

Variable selection for generalized linear mixed models and
non-Gaussian Genome-wide associated study data

Shuangshuang Xu

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

Marco A. R. Ferreira, Chair

Christopher T. Franck

Inyoung Kim

Allison N. Tegge

April 24, 2024

Blacksburg, Virginia

Keywords: GLMM, GWAS, Non Gaussian data, Bayesian variable selection

Copyright 2024, Shuangshuang Xu

Variable selection for generalized linear mixed models and non-Gaussian Genome-wide associated study data

Shuangshuang Xu

(ABSTRACT)

Genome-wide associated study (GWAS) aims to identify associated single nucleotide polymorphisms (SNP) for phenotypes. SNP has the characteristic that the number of SNPs is from hundred of thousands to millions. If p is the number of SNPs and n is the sample size, it is a $p \gg n$ variable selection problem. To solve this $p \gg n$ problem, the common method for GWAS is single marker analysis (SMA). However, since SNPs are highly correlated, SMA identifies true causal SNPs with high false discovery rate. In addition, SMA does not consider interaction between SNPs. In this dissertation, we propose novel Bayesian variable selection methods BG2 and IBG3 for non-Gaussian GWAS data. To solve ultra-high dimension problem and highly correlated SNPs problem, BG2 and IBG3 have two steps: screening step and fine-mapping step. In the screening step, BG2 and IBG3, like SMA method, only have one SNP in one model and screen to obtain a subset of most associated SNPs. In the fine-mapping step, BG2 and IBG3 consider all possible combinations of screened candidate SNPs to find the best model. Fine-mapping step helps to reduce false positives. In addition, IBG3 iterates these two steps to detect more SNPs with small effect size. In simulation studies, we compare our methods with SMA methods and fine-mapping methods. We also compare our methods with different priors for variables, including nonlocal prior, unit information prior, Zellner-g prior, and Zellner-Siow prior. Our methods are applied to substance use disorder (alcohol consumption and cocaine dependence), human health (breast cancer), and plant science (the number of root-like structure).

Variable selection for generalized linear mixed models and non-Gaussian Genome-wide associated study data

Shuangshuang Xu

(GENERAL AUDIENCE ABSTRACT)

Genome-wide associated study (GWAS) aims to identify genomics variants for targeted phenotype, such as disease and trait. The genomics variants which we are interested in are single nucleotide polymorphisms (SNP). SNP is a substitution mutation in the DNA sequence. GWAS solves the problem that which SNP is associated with the phenotype. However, the number of possible SNPs is from hundred of thousands to millions. The common method for GWAS is called single marker analysis (SMA). SMA only considers one SNP's association with the phenotype each time. In this way, SMA does not have the problem which comes from the large number of SNPs and small sample size. However, SMA does not consider the interaction between SNPs. In addition, SNPs that are close to each other in the DNA sequence may highly correlated SNPs causing SMA to have high false discovery rate. To solve these problems, this dissertation proposes two variable selection methods (BG2 and IBG3) for non-Gaussian GWAS data. Compared with SMA methods, BG2 and IBG3 methods detect true causal SNPs with low false discovery rate. In addition, IBG3 can detect SNPs with small effect sizes. Our methods are applied to substance use disorder (alcohol consumption and cocaine dependence), human health (breast cancer), and plant science (the number of root-like structure).

Dedication

This dissertation is dedicated to my mother Qing and my father Haifeng. Thank you so much for everything.

Contents

List of Figures	ix
List of Tables	xiii
1 Introduction	1
1.1 Bayesian Model Selection for Generalized Linear Mixed Models	2
1.2 BG2: Bayesian variable selection in generalized linear mixed models with nonlocal priors for non-Gaussian GWAS data	3
1.3 IBG3: Iterative Bayesian fine-mapping for GLMMs and non-Gaussian GWAS data	4
2 Bayesian Model Selection for Generalized Linear Mixed Models	6
2.1 Introduction	6
2.2 GLMMs	9
2.3 Pseudo likelihood Function for GLMMs	10
2.4 Model Selection	12
2.4.1 Priors for Model Parameters	13
2.4.2 Priors on the Model Space	15
2.4.3 Integrated Likelihood Methods	15

2.4.4	Fractional Bayes Factors	16
2.5	Simulation Study	18
2.6	Case Studies	23
2.6.1	Longitudinal Epilepsy Seizure Data	23
2.6.2	Spatial Lip Cancer Data	26
2.7	Discussion	27
2.8	Supplementary Material	29
2.8.1	The pseudo likelihood approach	29
2.8.2	Additional tables	34
2.8.3	Case Study - Binary Salamander Mating Data	35
2.8.4	Additional simulation studies	37
2.8.5	Additional figures	41
3	BG2: Bayesian variable selection in generalized linear mixed models with nonlocal priors for non-Gaussian GWAS data	59
3.1	Introduction	59
3.2	GLMMs for GWAS	63
3.3	BG2: Bayesian SNP selection in GLMMs for GWAS	64
3.3.1	Pseudo-likelihood model fitting	64
3.3.2	BG2 screening step	65
3.3.3	BG2 model selection step	68

3.4	Simulation Studies	70
3.4.1	Binary data	71
3.4.2	Count data	73
3.5	Case studies	75
3.5.1	Maximum number of alcoholic drinks	75
3.5.2	Cocaine dependence	76
3.5.3	Root-like structures in <i>A. Thaliana</i>	76
3.6	Discussion	77
3.7	Supplementary Material	78
3.7.1	The pseudo-likelihood approach	78
3.7.2	Results of simulation study of count data simulated with human genome	83
3.7.3	Boxplots of TP, FP, FDR, and F1 in the simulation studies	84
3.7.4	Robustness of BG2 when dealing with imbalanced binary data or highly skewed count data	87
3.7.5	Robustness of BG2 to genome spacing of SNPs	88
3.7.6	Sensitivity of BG2 to parameter values	90
3.7.7	Calibration of the pseudo-likelihood approach	92
3.7.8	Histograms of the response variables in the case studies	94
4	IBG3: Iterative Bayesian fine-mapping for GLMMs and non-Gaussian GWAS data	96

4.1	Introduction	96
4.2	GLMMs	99
4.3	Iterative model selection	100
4.3.1	Pseudo-likelihood method and Population Parameters Previously De- termined approach	100
4.3.2	Screening step	102
4.3.3	Fine-mapping step	104
4.4	Simulation Studies	105
4.5	Case studies	107
4.5.1	Maximum number of alcoholic drinks	108
4.5.2	Breast cancer	109
4.6	Discussion	114
4.7	Supplementary Material	115
4.7.1	Priors for IBG3 method	115
4.7.2	Plots for simulation study	119
5	Conclusions	121
	Bibliography	124

List of Figures

2.1	Simulation Results of HCM and ARM with $\tau_1 = 0.05$, $\tau_2 = 0.05$, $n=100, 400, 900$, $\beta_0 = 1, 4$	42
2.2	Simulation Results of HCM and ARM with $n = 400$, $\beta_0 = 1$, (a) $\tau_1 = 0.01$, $\tau_2 = 0$, and (b) $\tau_1 = 1$, $\tau_2 = 1$	43
2.3	Simulation Results of HCM and ARM with $\beta_0 = 2$, $\beta_1 = \beta_2 = 1$, $n=100, 400, 900$, and $\tau_2 = 0, 0.1$	44
2.4	Simulation Results of HCM and ARM with $\beta_0 = 2$, $\beta_1 = \beta_2 = 1$, $n=100, 400, 900$, and $\tau_1 = 0, 0.1$	45
2.5	Simulation Results of HCM and ARM with data generated from a Gaussian hierarchical model with mean zero and ICAR random effects.	46
2.6	Simulation Results of HCM and ARM selecting covariates with Bernoulli data.	47
2.7	Simulation Results of HCM and ARM selecting longitudinal random effects with Bernoulli data.	48
2.8	Simulation Results of HCM and ARM selecting covariates with Poisson data.	49
2.9	Simulation Results of HCM and ARM selecting covariates, comparing with DIC computed by INLA.	50
2.10	Simulation Results of HCM and ARM selecting covariates, comparing with WAIC computed by INLA.	51

2.11	Simulation Results of HCM and ARM selecting spatial random effects, comparing with DIC computed by INLA.	52
2.12	Simulation Results of HCM and ARM selecting spatial random effects, comparing with WAIC computed by INLA.	53
2.13	Simulation Results of HCM and ARM selecting overdispersion random effects, comparing with DIC computed by INLA.	54
2.14	Simulation Results of HCM and ARM selecting overdispersion random effects, comparing with WAIC computed by INLA.	55
2.15	Simulation Results of HCM and ARM selecting covariates, comparing with marginal likelihood computed by INLA.	56
2.16	Simulation Results of HCM and ARM selecting spatial random effects, comparing with marginal likelihood computed by INLA.	57
2.17	Simulation Results of HCM and ARM selecting overdispersion random effects, comparing with marginal likelihood computed by INLA.	58
3.1	Simulation Results of BG2 with binary data generated by A.Thaliana GWAS data.	72
3.2	Simulation Results of BG2 with count data generated by A.Thaliana GWAS data.	74
3.3	Simulation Results of BG2 with count data generated by human GWAS data.	83
3.4	Simulation Results of BG2 with binary data generated by human GWAS data.	84
3.5	Boxplots of BG2 and SMA with count data generated by human GWAS data.	85

3.6	Boxplots of BG2 and SMA with count data generated by A. Thaliana GWAS data.	86
3.7	Histograms for two simulated count datasets.	87
3.8	Simulation Results of BG2 with $\beta = 0.4, -0.4$	90
3.9	Simulation Results of BG2 with $\beta_0 = 1$	91
3.10	Simulation Results of BG2 with $\kappa = 0.3$	92
3.11	Q-Q plot of p-values based the pseudo-likelihood approach, human genome data.	93
3.12	Q-Q plot of p-values based the pseudo-likelihood approach, A. Thaliana genome data.	93
3.13	Histogram of the maximum number of alcoholic drinks.	94
3.14	Histogram of the cocaine dependence	95
3.15	Histogram of number of root-like structures in A. Thaliana.	95
4.1	The posterior probabilities of all SNPs in Iteration 1.	112
4.2	The posterior probabilities of all SNPs in Iteration 2.	112
4.3	The posterior probabilities of all SNPs in Iteration 3.	113
4.4	The posterior probabilities of all SNPs in Iteration 4.	113
4.5	The posterior probabilities of SNPs in two steps of Iteration 1 for simulation study.	119
4.6	The posterior probabilities of SNPs in two steps of Iteration 2 for simulation study.	120

4.7	The posterior probabilities of SNPs in two steps of Iteration 3 for simulation study.	120
-----	---	-----

List of Tables

2.1	Epilepsy data: posterior inclusion probabilities of fixed and random effects	24
2.2	Lip cancer data: posterior inclusion probabilities of fixed and random effects	27
2.3	Epilepsy data: summary of model selection results by competing methods	34
2.4	Estimates for epilepsy case study	34
2.5	Lip cancer data: DIC and WAIC for competing models	34
2.6	Lip cancer data: model selection summary	35
2.7	Estimates for lip cancer case study	35
2.8	Salamander mating data: posterior inclusion probabilities of fixed and random effects	36
2.9	Salamander mating data: model selection summary	37
2.10	Estimates for salamander mating case study	37
3.1	Performance of BG2 and SMA when data are skewed.	87
3.2	Performance of BG2 and SMA when data are balanced or imbalanced.	88
3.3	Robustness to genome spacing of SNPs.	89
4.1	True positives (TP) for IBG3, GEMMA, SuSiE-RSS, and BG2.	107
4.2	False positives (FP) for IBG3, GEMMA, SuSiE-RSS, and BG2.	107
4.3	False discovery rate (FDR) for IBG3, GEMMA, SuSiE-RSS, and BG2.	108

4.4	F1 score for IBG3, GEMMA, SuSiE-RSS, and BG2.	108
4.5	Time (min) for IBG3, GEMMA, SuSiE-RSS, and BG2.	109

List of Abbreviations

DIC Deviance information criterion

FBF Fractional Bayes factor

FDR False discovery rate

FP False positive

GA Genetic algorithm

GLMM Generalized linear mixed model

GWAS Genome-wide associated study

LMM Linear mixed model

MAF Minor allele frequency

P3D Population Parameters Previously Determined

PQL Penalized quasi likelihood method

SMA Single marker analysis

SNP Single nucleotide polymorphism

TP True positive

WAIC Watanabe-Akaike information criterion

Chapter 1

Introduction

Genome-wide associated study (GWAS) aims to identify associated single nucleotide polymorphisms (SNP) for phenotypes. Typically in GWAS the number of possible SNPs is from hundred of thousands (10^5) to millions (10^6) while the sample size n is from thousands (10^3) to tens of thousands (10^4). Thus, this is a variable selection problem with $p \gg n$. To solve this $p \gg n$ problem, the common method for GWAS is single marker analysis (SMA). However, since SNPs are highly correlated, SMA identifies true causal SNPs with high false discovery rate. In addition, SMA does not consider interaction between SNPs. In this dissertation, we propose novel Bayesian variable selection methods (ARM and HCM) for generalized linear mixed models (GLMMs), and two Bayesian variable selection methods BG2 and IBG3 for non-Gaussian GWAS data.

Bayesian variable selection for GLMMs has difficulty to obtain the marginal likelihood. Pseudo likelihood approach can approximate GLMMs for non-Gaussian data by computing adjusted observations that can be approximately modeled by Linear mixed models (LMMs). To solve ultra-high dimension problem and highly correlated SNPs problem, BG2 and IBG3 have two steps: screening step and fine-mapping step. In the screening step, similarly to SMAs, BG2 and IBG3 consider only one SNP per model and screen to obtain a subset of most associated SNPs. In the fine-mapping step, BG2 and IBG3 consider all possible combinations of screened candidate SNPs to find the best model. The fine-mapping step helps to reduce false positives. In addition, IBG3 iterates these two steps to detect more

SNPs with small effect size. In the simulation study, we compare our methods with SMA methods and fine-mapping methods. We also compare our methods with different priors for the regression coefficients, including nonlocal prior, unit information prior, Zellner-g prior, and Zellner-Siow prior. Our methods are applied to substance use disorder (alcohol consumption and cocaine dependence), human health (breast cancer), and plant science (the number of root-like structure).

Thus, this dissertation is about Bayesian variable selection for GLMMs and application to GWAS data. Chapter 2 presents ARM and HCM which are two novel Bayesian model selection methods for GLMMs. ARM and HCM can select fixed effects and random effects simultaneously. ARM and HCM solve the problem to obtain the inclusion posterior probability for each coefficient and random effect. Chapter 3 presents BG2, which is a Bayesian method based on nonlocal priors for ultrahigh-dimension variable selection in GLMMs for GWAS data. Chapter 4 presents IBG3, which is an iterative Bayesian fine-mapping method based on GLMMs for non-Gaussian GWAS data. Compared with BG2, IBG3 is an iterative approach that detects more SNPs. In addition, we develop unit information prior, Zellner-g prior, and Zellner-Siow prior for GLMMs and compare IBG3 methods with different priors.

1.1 Bayesian Model Selection for Generalized Linear Mixed Models

In Chapter 2, we propose a Bayesian model selection approach for generalized linear mixed models (GLMMs). We consider covariance structures for the random effects that are widely used in areas such as longitudinal studies, genome-wide association studies, and spatial statistics. Since the random effects cannot be integrated out of GLMMs analytically, we

approximate the integrated likelihood function using a pseudo likelihood approach. Our Bayesian approach assumes a flat prior for the fixed effects and includes both approximate reference prior and half-Cauchy prior choices for the variances of random effects. Since the flat prior on the fixed effects is improper, we develop a fractional Bayes factor approach to obtain posterior probabilities of the several competing models. Simulation studies with Poisson generalized linear mixed models with spatial random effects and overdispersion random effects show that our approach performs favorably when compared to widely used competing Bayesian methods including DIC and WAIC. We illustrate the usefulness and flexibility of our approach with three case studies including a Poisson longitudinal model, a Poisson spatial model, and a logistic mixed model. Our proposed approach is implemented in the R package `GLMMselect` [77] that is available on CRAN.

1.2 BG2: Bayesian variable selection in generalized linear mixed models with nonlocal priors for non-Gaussian GWAS data

In Chapter 3, we develop BG2 which is a Bayesian variable selection method in GLMMs for GWAS. Genome-wide association studies (GWASes) aim to identify single nucleotide polymorphisms (SNPs) associated with a given phenotype. A common approach for the analysis of GWAS is single marker analysis (SMA) based on linear mixed models (LMMs). However, LMM-based SMA usually yields a large number of false discoveries and cannot be directly applied to non-Gaussian phenotypes such as count data. To address these limitations, we present a novel Bayesian method to find SNPs associated with non-Gaussian phenotypes. To that end, we use generalized linear mixed models (GLMMs) and, thus, call our method

Bayesian GLMMs for GWAS (BG2). To deal with the high dimensionality of GWAS analysis, we propose novel nonlocal priors specifically tailored for GLMMs. In addition, we develop related fast approximate Bayesian computations. BG2 uses a two-step procedure: first, BG2 screens for candidate SNPs; second, BG2 performs model selection that considers all screened candidate SNPs as possible regressors. A simulation study shows favorable performance of BG2 when compared to GLMM-based SMA. We illustrate the usefulness and flexibility of BG2 with three case studies on cocaine dependence (binary data), alcohol consumption (count data), and number of root-like structures in a model plant (count data). BG2 is implemented in the R package BG2 [79] that is available on Bioconductor.

1.3 IBG3: Iterative Bayesian fine-mapping for GLMMs and non-Gaussian GWAS data

In Chapter 4, we propose a novel iterative Bayesian variable selection method for generalized linear mixed models (GLMMs) specifically designed for the analysis of non-Gaussian genome-wide association studies (GWAS) data. GWAS main goal is to identify single nucleotide polymorphisms (SNPs) associated with phenotypes of interest. Usually, GWAS data are analyzed with single marker analysis (SMA) methods. However, SMA methods usually suffer from high false discovery rate (FDR). Fine-mapping method can help reduce FDR. We compare the Zellner g prior with three other priors: nonlocal prior, unit information prior, and Zellner-Siow prior. The Zellner g prior has good performance in terms of false discovery rate and timing. We adapt to GLMMs the Zellner g prior and thus call our method iterative Bayesian GLMM for GWAS with Zellner g prior (IBG3). IBG3 is a genome-wide fine mapping method. IBG3 iterates two steps: a screening step that screens for candidate SNPs; and a fine-mapping step that considers all screened candidate SNPs as possible regressors. A

simulation study shows that IBG3 has favorable performance when compared to GEMMA, SuSiE-RSS, and BG2. We illustrate the usefulness and flexibility of IBG3 with two case studies on alcohol use disorder and breast cancer.

Chapter 2

Bayesian Model Selection for Generalized Linear Mixed Models

2.1 Introduction

This chapter is based on the following manuscript that has been published in *Biometrics* [76]: Shuangshuang Xu, Marco A.R. Ferreira, Erica M. Porter, and Christopher T. Franck. Bayesian model selection for generalized linear mixed models. *Biometrics*, 2023, 79(4): 3266-3278.

Generalized linear mixed models (GLMMs) are widely used to model non-Gaussian data with dependent observations. This type of data is often found in many areas of application such as epidemiology [43], meta-analysis of multiple clinical trials [59], survival analysis [65], and neuroimaging [41]. Even though Bayesian estimation procedures for GLMMs are well established, there are just a handful of papers that address Bayesian model selection for GLMMs. Currently, most applied papers use the deviance information criterion (DIC) [63] to perform Bayesian model selection for GLMMs [49, 68]. Even though the DIC is widely applicable, we show in a simulation study that the DIC has some undesirable behaviors when applied to GLMMs. To provide more reliable results, here we develop a novel Bayesian model selection approach for simultaneous selection of covariates and random effects for GLMMs.

Specifically, we focus on GLMMs where each random effect has a covariance matrix that is the product of an unknown variance component parameter and a known positive semi-definite symmetric matrix. The class of GLMMs we consider can be used for the analysis of spatial areal data [2, 14], genome-wide association studies (GWAS) [71], and longitudinal data [9, 75]. However, inference for GLMMs is difficult because the integrated likelihood function is not available in closed form. To deal with the issue of integration of random effects, we approximate the integrated likelihood function using a pseudo likelihood approach [72] that leads to a Gaussian likelihood approximation. We then assign a flat prior for the vector of regression coefficients and an approximate reference prior [20] for the variance components of the GLMMs, which is inspired by the reference prior proposed by [34] for Gaussian data. In addition, we also consider a half-Cauchy prior for the square root of variance components [21, 52]. Because the prior on the vector of regression coefficients is improper, we develop a fractional Bayes factor (FBF) approach [50]. We note that [53] have proposed FBF for Gaussian mixed models for the particular case of spatial areal data. In contrast, here we consider generalized linear mixed models. In addition, we consider not only spatial random effects but also many other types of random effects such as overdispersion random effects and longitudinal random effects. Because we use default priors combined with FBF, our proposed model selection approach is fully automatic, which obviates the need for subjective specification of hyperparameters and makes the method more accessible for practitioners. We call our two proposed model selection approaches the approximate reference method (ARM) and the half-Cauchy method (HCM).

To compare the performance of our methods ARM and HCM to the performance of the DIC, the Watanabe-Akaike information Criterion (WAIC) [70], and marginal likelihood computed by INLA under different parameter settings, we present a simulation study based on Poisson generalized linear mixed models with a spatial random effect and an overdispersion random

effect. In this simulation study, we vary the sample size, coefficient of non-null covariates, level of spatial dependence, and overdispersion level. The simulation study shows that DIC and WAIC cannot reliably distinguish the random effect when there is another random effect. In contrast, our methods ARM and HCM perform well at detecting covariates and correct dependence structure. In particular, ARM and HCM always correctly detect the case of no random effects. Finally, while the performances of the DIC and WAIC do not improve much with large sample sizes, our proposed ARM and HCM have large improvement with increasing sample size. In addition, the simulation study shows that marginal likelihood computed by INLA has similar performance to our methods ARM and HCM when selecting covariates. However, marginal likelihood computed by INLA does not perform well when selecting random effects.

Apart from the DIC, WAIC, and marginal likelihood, there are not many other Bayesian model selection approaches for GLMMs. One such approach proposed by [10] for simultaneously selecting fixed and random effects in GLMMs assumes that the subject-specific random effects have a covariance matrix with all its elements being free parameters to be estimated. As a consequence, the method proposed by [10] is only applicable to problems with replications and can not be readily applied to problems where the vector of observations is a realization from a structured multivariate distribution such as GWAS data and spatial areal data. In contrast, because we assume that each random effect has a covariance matrix that is the product of an unknown variance component parameter with a known positive semi-definite covariance matrix, our methods ARM and HCM can be applied to longitudinal data, GWAS data, and spatial areal data.

The remainder of this paper is organized as follows. Section 2.2 describes the GLMMs that we consider. Section 2.3 outlines how the pseudo likelihood approach approximates GLMMs for non-Gaussian data by computing adjusted observations that are modeled using Gaussian

LMMs. Section 2.4 introduces priors for model selection, the FBF approach for dealing with improper priors, and posterior computation. Section 2.5 presents the results of a simulation study. Section 2.6 illustrates our method with applications to two case studies. Section 2.7 concludes with a discussion and future directions.

The Appendix contains details about the pseudo likelihood method (Appendix 2.8.1), additional tables for the case studies (Appendix 2.8.2), one additional case study (Appendix 2.8.3), several additional simulation studies (Appendix 2.8.4), and additional figures (Appendix 2.8.5).

2.2 GLMMs

Consider a response vector $\mathbf{y} = (y_1, y_2, \dots, y_n)^\top$ of n observations. Let \mathbf{X} be an n by p matrix of explanatory variables and $\boldsymbol{\beta}$ be the corresponding p -dimensional vector of fixed effects. Let \mathbf{Z}_j be an n by q_j design matrix and $\boldsymbol{\alpha}_j$ be the corresponding q_j -dimensional vector of random effects, $j = 1, \dots, Q$. Let vectors \mathbf{x}_i and \mathbf{z}_{ij} be the i th rows of \mathbf{X} and \mathbf{Z}_j , respectively. Conditional on linear predictors η_1, \dots, η_n , the observations y_1, \dots, y_n are independent with probability density function belonging to the exponential family, that is $f(y_i|\eta_i) = \exp[y_i\eta_i - B_i(\eta_i) + C_i(y_i)]$, $i = 1, \dots, n$, where the canonical parameter η_i is modeled as a linear function of fixed and random effects as $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \sum_j \mathbf{z}_{ij}^\top \boldsymbol{\alpha}_j$. Each observation y_i has mean $\mu_i = B_i'(\eta_i)$ and variance $v_i = B_i''(\eta_i)$. In addition, we assume that each vector of random effects $\boldsymbol{\alpha}_j$ has a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\tau_j \boldsymbol{\Sigma}_j$, where the variance component parameter τ_j is unknown and $\boldsymbol{\Sigma}_j$ is a known symmetric positive semi-definite matrix. For example, if $\boldsymbol{\alpha}$ is a vector of overdispersion random effects then the corresponding matrix $\boldsymbol{\Sigma}$ is an identity matrix.

As another example, in the case of spatial areal data, we assume that $\boldsymbol{\alpha}$ is a vector of spatial random effects that follows a sum-zero constrained Gaussian Intrinsic Conditional

Autoregressive Model [33, 34], that is,

$$\boldsymbol{\alpha}|\boldsymbol{\tau} \sim N(\mathbf{0}, \boldsymbol{\tau}\boldsymbol{\Sigma}), \quad (2.1)$$

where $\boldsymbol{\Sigma}$ is a known positive semi-definite covariance matrix that depends on the neighborhood structure of the spatial subregions. Specifically, an adjacency matrix \mathbf{W} is defined such that if subregion i and subregion j are adjacent, the entries in cells (i, j) and (j, i) are 1, otherwise 0. Let \mathbf{D}_w be a diagonal matrix with each diagonal element equal to the summation of the corresponding row of \mathbf{W} . Then, the covariance matrix $\boldsymbol{\Sigma}$ is the Moore-Penrose inverse of $\mathbf{D}_w - \mathbf{W}$ [33, 34]. We note that computations for this model may be performed using the precision matrix. In addition, we note that the knowledge about the covariance matrix $\boldsymbol{\Sigma}$ has allowed, for the case of Gaussian hierarchical models with ICAR random effects, the derivation of a reference prior for the parameters [34], and formal Bayesian model selection [53].

2.3 Pseudo likelihood Function for GLMMs

A key step in Bayesian model selection is to integrate out random effects from the likelihood function. However, while for LMMs the random effects can be integrated out analytically, for GLMMs that is not possible. To overcome this difficulty, here we use a pseudo likelihood approach that approximates a GLMM for non-Gaussian data by computing adjusted observations that are modeled using an approximate Gaussian LMM.

Let $\boldsymbol{\alpha}$ represent all random effects and $\boldsymbol{\tau}$ represent all variance components. Then, the

likelihood function with the relevant but intractable integral over random effects $\boldsymbol{\alpha}$ is

$$\begin{aligned}
L(\boldsymbol{\beta}, \boldsymbol{\tau} | \mathbf{y}) &= \int p(\mathbf{y} | \boldsymbol{\alpha}, \boldsymbol{\beta}) p(\boldsymbol{\alpha} | \boldsymbol{\tau}) d\boldsymbol{\alpha} \\
&= \int \prod_{i=1}^N \left[\exp \left\{ y_i \left(\mathbf{x}_i^\top \boldsymbol{\beta} + \sum_j \mathbf{z}_{ij}^\top \boldsymbol{\alpha}_j \right) - B_i \left(\mathbf{x}_i^\top \boldsymbol{\beta} + \sum_j \mathbf{z}_{ij}^\top \boldsymbol{\alpha}_j \right) + C_i(y_i) \right\} \right] \\
&\quad \prod_j \left[(2\pi)^{-\frac{q_j}{2}} |\tau_j \boldsymbol{\Sigma}_j|^{-\frac{1}{2}} \exp \left\{ -\frac{\boldsymbol{\alpha}_j^\top \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\alpha}_j}{2\tau_j} \right\} \right] d\boldsymbol{\alpha}. \tag{2.2}
\end{aligned}$$

In Equation (2.2), the random effects $\boldsymbol{\alpha}$ cannot be integrated out analytically. Our method approximates the integral in Equation (2.2) with a Gaussian LMM via a pseudo likelihood approach. For a Gaussian LMM, the corresponding integral can be solved analytically, and then the likelihood function of parameters has an analytic expression.

The pseudo likelihood approach was first proposed by [72]. The pseudo likelihood approach is an iterative procedure that starts by writing the model as $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)'$ and $\boldsymbol{\epsilon}$ is a vector of errors with $cov(\boldsymbol{\epsilon}) = \mathbf{V} = diag(v_1, \dots, v_n)$. Let $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{V}}$ be the current estimates of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\mu}$ and \mathbf{V} . Here, $\hat{\boldsymbol{\beta}}$ is initialized at the estimate from a GLM fit. Now, approximate μ_i with a first-order Taylor expansion around $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$. Rearrange all the terms in $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$ such that the terms that depend on \mathbf{y} , $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\mu}}$ appear on the left side of the equation and the remaining terms appear on the right side of the equation. Multiply both sides by $\hat{\mathbf{V}}^{-1}$. As a result, the left side of the equation will have $\mathbf{y}^* = \hat{\mathbf{V}}^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}) + \mathbf{X}\hat{\boldsymbol{\beta}} + \sum_j \mathbf{Z}_j \hat{\boldsymbol{\alpha}}_j$. The vector \mathbf{y}^* is known as the vector of pseudo-observations or the vector of adjusted observations. Equating \mathbf{y}^* to the right side of the

equation, we obtain the following model for the adjusted observations.

$$\begin{aligned}\mathbf{y}^* &\approx \mathbf{X}\boldsymbol{\beta} + \sum_j \mathbf{Z}_j \boldsymbol{\alpha}_j + \widehat{\mathbf{V}}^{-1} \boldsymbol{\epsilon}, \\ \boldsymbol{\alpha}_j &\sim N(\mathbf{0}, \tau_j \boldsymbol{\Sigma}_j), \\ \boldsymbol{\epsilon} &\sim N(\mathbf{0}, \mathbf{V}).\end{aligned}\tag{2.3}$$

Thus, the pseudo likelihood approach assumes that $\boldsymbol{\epsilon}$ follows a Gaussian distribution with mean vector $\mathbf{0}$ and covariance matrix \mathbf{V} . Substituting \mathbf{V} with $\widehat{\mathbf{V}}$ in Equation (2.3), \mathbf{y}^* can be approximately modeled with the LMM $\mathbf{y}^* \sim N(\mathbf{X}\boldsymbol{\beta}, \sum_j \tau_j \mathbf{Z}_j \boldsymbol{\Sigma}_j \mathbf{Z}_j^\top + \widehat{\mathbf{V}}^{-1})$. Therefore, we have the closed-form pseudo likelihood function

$$\begin{aligned}p(\mathbf{y}^* | \boldsymbol{\beta}, \tau) &= (2\pi)^{-\frac{n}{2}} \left| \sum_j \tau_j \mathbf{Z}_j \boldsymbol{\Sigma}_j \mathbf{Z}_j^\top + \widehat{\mathbf{V}}^{-1} \right|^{-\frac{1}{2}} \\ &\quad \exp \left\{ -\frac{1}{2} (\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta})^\top \left(\sum_j \tau_j \mathbf{Z}_j \boldsymbol{\Sigma}_j \mathbf{Z}_j^\top + \widehat{\mathbf{V}}^{-1} \right)^{-1} (\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}) \right\}.\end{aligned}\tag{2.4}$$

Further details about the pseudo likelihood approach appear in Section 2.8.1. To perform model selection, we first use the pseudo likelihood function in Equation (2.4) in an iterative manner to estimate the parameters and to obtain adjusted observations \mathbf{y}^* . We then use these adjusted observations \mathbf{y}^* rather than the original observations \mathbf{y} to perform model selection.

2.4 Model Selection

We perform model selection based on the pseudo likelihood function given in Equation (2.4). Similarly to [66], we use the same vector of adjusted observations to compare all candidate

models' posterior probabilities. Specifically, we compute the vector of adjusted observations using the full model with all candidate regressors and all candidate random effects. In addition, consider the model space $\mathcal{M} = \{M_c, c = 1 \dots C\}$, with C possible models. We assume model M_c has K_c regressors, where \mathbf{X}_c is the corresponding matrix of explanatory variables and $\boldsymbol{\beta}_c$ is the corresponding vector of coefficients. Further, model M_c has Q_c types of random effects. Let $\boldsymbol{\tau}_c = (\tau_{c,1}, \dots, \tau_{c,Q_c})$ be the vector of variance components of the Q_c types of random effects in the model M_c . The integrated likelihood based on the vector of adjusted observations \mathbf{y}^* is

$$p(\mathbf{y}^*|M_c) = \int \int p(\mathbf{y}^*|\boldsymbol{\beta}_c, \boldsymbol{\tau}_c) \pi(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c|M_c) d\boldsymbol{\beta}_c d\boldsymbol{\tau}_c, \quad (2.5)$$

where $\pi(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c|M_c)$ is the prior distribution of parameters conditional on model M_c . Application of Bayes' Theorem then yields posterior model probabilities:

$$P(M_c|\mathbf{y}^*) = \frac{p(\mathbf{y}^*|M_c)\pi(M_c)}{\sum_{r=1}^C p(\mathbf{y}^*|M_r)\pi(M_r)} \propto p(\mathbf{y}^*|M_c)\pi(M_c),$$

where $\pi(M_c)$ is the prior probability of model M_c .

In Section 2.4.1, we specify the priors for model parameters. In Section 2.4.2, we specify the priors on the model space. Section 2.4.3 discusses approximation of the integral in Equation (2.5). In Section 2.4.4, we propose an FBF approach [53] to perform model selection with improper priors.

2.4.1 Priors for Model Parameters

We consider the approximate reference prior proposed by [20] in the context of LMMs for $\boldsymbol{\beta}$ and the reciprocal of τ , which is based on the reference prior proposed by [34]. In what

follows, we consider the implied reference prior for τ obtained by transformation of variables. For simple notation, let M without subscript represent a general model, $\boldsymbol{\beta}$ represent the corresponding vector of regressor coefficients, and τ represent the variance component. In the reference prior [34], all the parameters are independent. The vector of regression coefficients $\boldsymbol{\beta}$ is assigned a uniform prior on \mathcal{R}^p . In addition, as τ goes to infinity the reference prior $\pi(\tau)$ is proportional to τ^{-2} . Further, as τ goes to 0 the reference prior is proportional to a constant. Based on the tail behavior of the reference prior for τ , [20] proposed the approximate reference prior $\pi(\tau) \propto (1 + \frac{\tau}{a_\tau})^{-2}$, where a_τ is a hyperparameter. We set a_τ equal to 2. The choice of $a_\tau = 2$ is equivalent to the choice made by [20] for Gaussian data. In addition, our simulation study shows that this choice also works well for GLMMs. Hence, for $\boldsymbol{\beta}$ we use the flat prior $\pi(\boldsymbol{\beta}|M) \propto 1$, and for τ we use the approximate reference prior

$$\pi_1(\tau|M) = \frac{1}{2(\tau/2 + 1)^2}, \quad \tau \geq 0. \quad (2.6)$$

This approximate reference prior is related to the half-Cauchy prior $\pi(\tau) \propto \frac{1}{\tau^2+1}$, which has the same tail behavior. [21] proposed a half-Cauchy prior, however, for the standard deviation of random effects in a two-level Gaussian model. Assuming a half-Cauchy prior for the square root of the variance component parameter τ implies for τ the prior density $\pi_2(\tau) \propto \tau^{-\frac{1}{2}}(\tau + 1)^{-1}$ [52]. Thus, $\pi(\tau) = O(\tau^{-\frac{1}{2}})$ for $\tau \rightarrow 0$ and $\pi(\tau) = O(\tau^{-\frac{3}{2}})$ for $\tau \rightarrow \infty$. Hence, the half-Cauchy prior for $\sqrt{\tau}$ has more mass near zero and more mass for large values of τ than the approximate reference prior for τ given in Equation (2.6). Here, we consider two variants of our pseudo-likelihood-based method: ARM, which uses the approximate reference prior given in Equation (2.6); and HCM, which uses the half-Cauchy prior for $\sqrt{\tau}$. We compare our methods ARM and HCM to the DIC and WAIC in the simulation studies presented in Section 2.5.

2.4.2 Priors on the Model Space

Let K denote the number of candidate covariates and Q denote the number of candidate random effects types. For example, in an application where we may have spatial random effects and/or overdispersion random effects, $Q = 2$. In addition, let K_c denote the number of covariates in Model M_c . For fixed effects, we use priors from [61], which automatically correct for multiplicity. Specifically, the prior probability for model M_c with K_c covariates is $P(M_c \text{ with } K_c \text{ covariates}) = 1 / \left[(K + 1) \binom{K}{K_c} \right]$. With respect to random effects, there are 2^Q possibilities for inclusion and exclusion of random effects. Assuming that each random effect has 0.5 prior inclusion probability, the prior probability for Model M_c with Q_c types of random effects is $P(M_c \text{ with } Q_c \text{ types of random effects}) = 1/2^{Q_c}$. Because usually in practice the number of candidate random effects types Q is small, a discrete uniform prior for the inclusion of random effects is reasonable. Assuming *a priori* independence of inclusion of fixed effects and random effects, the prior probability for model M_c is $P(M_c) = 1 / \left[2^{Q_c} (K + 1) \binom{K}{K_c} \right]$.

2.4.3 Integrated Likelihood Methods

After the priors for parameters have been defined, the integrated likelihood given in Equation (2.5) based on the adjusted observations \mathbf{y}^* becomes

$$\begin{aligned}
 p(\mathbf{y}^* | M_c) &= \int \int p(\mathbf{y}^* | \boldsymbol{\beta}_c, \boldsymbol{\tau}_c) \pi(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c | M_c) d\boldsymbol{\beta}_c d\boldsymbol{\tau}_c \\
 &\propto \int \int \exp \left\{ -\frac{1}{2} (\mathbf{y}^* - \mathbf{X}_c \boldsymbol{\beta}_c)^\top \left(\sum_j^{Q_c} (\tau_{cj} \mathbf{Z}_{cj} \boldsymbol{\Sigma}_{cj} \mathbf{Z}_{cj}^\top) + \widehat{\mathbf{V}}^{-1} \right)^{-1} (\mathbf{y}^* - \mathbf{X}_c \boldsymbol{\beta}_c) \right\} \\
 &\quad \left| \sum_j^{Q_c} (\tau_{cj} \mathbf{Z}_{cj} \boldsymbol{\Sigma}_{cj} \mathbf{Z}_{cj}^\top) + \widehat{\mathbf{V}}^{-1} \right|^{-\frac{1}{2}} \pi(\boldsymbol{\tau}_c) d\boldsymbol{\beta}_c d\boldsymbol{\tau}_c.
 \end{aligned}$$

The vector of regression coefficients $\boldsymbol{\beta}_c$ can be integrated out analytically. After integrating out $\boldsymbol{\beta}_c$, we can write the integrated likelihood as

$$\begin{aligned} p(\mathbf{y}^*|M_c) &= \int p(\mathbf{y}^*, \boldsymbol{\tau}_c|M_c) d\boldsymbol{\tau}_c \\ &\propto \int \exp \left[\frac{1}{2} \mathbf{y}^{*\top} \{ \mathbf{H}_c^{-1} \mathbf{X}_c (\mathbf{X}_c^\top \mathbf{H}_c^{-1} \mathbf{X}_c)^{-1} \mathbf{X}_c^\top \mathbf{H}_c^{-1} - \mathbf{H}_c^{-1} \} \mathbf{y}^* \right] \\ &\quad \left| \mathbf{H}_c^{-1} (\mathbf{X}_c^\top \mathbf{H}_c^{-1} \mathbf{X}_c)^{-1} \right|^{\frac{1}{2}} \pi(\boldsymbol{\tau}_c) d\boldsymbol{\tau}_c, \end{aligned} \quad (2.7)$$

where $\mathbf{H}_c = \sum_j^{Q_c} (\tau_{cj} \mathbf{Z}_{cj} \boldsymbol{\Sigma}_{cj} \mathbf{Z}_{cj}^\top) + \widehat{\mathbf{V}}^{-1}$. Note that the vector of variance components $\boldsymbol{\tau}_c$ cannot be integrated out analytically. To compute the integral in Equation (2.7), we first perform a logarithm transformation on $\boldsymbol{\tau}_c$. Let $\boldsymbol{\delta}_c = \log(\boldsymbol{\tau}_c)$ be the vector obtained by applying the logarithm transformation to each element of $\boldsymbol{\tau}_c$. Then, we integrate out $\boldsymbol{\delta}_c$ using a Laplace approximation to obtain

$$\begin{aligned} \int p(\mathbf{y}^*, \boldsymbol{\tau}_c|M_c) d\boldsymbol{\tau}_c &= \int p(\mathbf{y}^*, \exp(\boldsymbol{\delta}_c)|M_c) \exp(\boldsymbol{\delta}_c) d\boldsymbol{\delta}_c \\ &\approx (2\pi)^{\frac{Q_c}{2}} \left| q''(\widehat{\boldsymbol{\delta}}_c) \right|^{-\frac{1}{2}} \exp \left\{ -q(\widehat{\boldsymbol{\delta}}_c) \right\}, \end{aligned} \quad (2.8)$$

where $q(\boldsymbol{\delta}_c) = -\frac{1}{2} \mathbf{y}^{*\top} [\mathbf{H}_c^{-1} \mathbf{X}_c (\mathbf{X}_c^\top \mathbf{H}_c^{-1} \mathbf{X}_c)^{-1} \mathbf{X}_c^\top \mathbf{H}_c^{-1} - \mathbf{H}_c^{-1}] \mathbf{y}^* - \log \pi(\exp(\boldsymbol{\delta}_c)) + \boldsymbol{\delta}_c - \frac{1}{2} \log |\mathbf{H}_c^{-1} (\mathbf{X}_c^\top \mathbf{H}_c^{-1} \mathbf{X}_c)^{-1}|$, $\widehat{\boldsymbol{\delta}}_c$ is the point that minimizes $q(\boldsymbol{\delta}_c)$, and $q''(\boldsymbol{\delta}_c)$ is the Hessian matrix.

2.4.4 Fractional Bayes Factors

In order to obtain the posterior model probabilities of interest, we use a fractional Bayes factor (FBF) approach. The FBF is a modification of the Bayes factor that allows for improper priors on parameters [50].

To define the usual Bayes factor, let the baseline model M_l be the model with the largest integrated likelihood in the model space. Then, the Bayes factor BF_{cl} of model M_c versus the baseline model M_l is defined as the ratio of their integrated likelihoods $BF_{cl} = \frac{p(\mathbf{y}^*|M_c)}{p(\mathbf{y}^*|M_l)}$. Hence, we can compute the posterior probability of model M_c as proportional to its prior probability times its Bayes factor versus the baseline model, that is $P(M_c|\mathbf{y}^*) \propto P(M_c)p(\mathbf{y}^*|M_c)/p(\mathbf{y}^*|M_l) \propto BF_{cl}P(M_c)$.

Note that the prior on the regression coefficients $\pi(\boldsymbol{\beta}_c|M_c) \propto d$ is improper, where d is an arbitrary constant. Thus, the Bayes factor computed with the integrated likelihood from Equations (2.7) and (2.8) is only defined up to an unspecified constant of proportionality and cannot be used to compare models directly.

To solve this problem, we use the fractional Bayes factor (FBF, [50]) to approximate the Bayes factor. [53] developed the fractional Bayes factor method for Gaussian hierarchical models with ICAR random effects. We use the FBF approach to train the improper prior so that we can compute a meaningful Bayes factor. By training the improper prior, we mean using Bayes Theorem to combine the improper prior with a portion of the data or a fraction of the likelihood to obtain a proper distribution [6, 50, 53]. We can then use this latter distribution as a trained prior to compute a meaningful Bayes factor. Specifically, here we train the prior with a fraction b of the likelihood function. The trained prior density for model M_c is obtained by Bayes Theorem as

$$\pi^b(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c) = \frac{p^b(\mathbf{y}^*|\boldsymbol{\beta}_c, \boldsymbol{\tau}_c)\pi(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c|M_c)}{\int p^b(\mathbf{y}^*|\boldsymbol{\beta}_c, \boldsymbol{\tau}_c)\pi(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c|M_c) d\boldsymbol{\beta}_c d\boldsymbol{\tau}_c}.$$

The integrated likelihood is then computed as an integral of the product of the likelihood function raised to $1 - b$ and the trained prior. Following [50], the resulting integrated

likelihood of model M_c , called the fractional integrated likelihood, is equal to

$$\begin{aligned}
q_c(b, \mathbf{y}^*) &= \int p^{1-b}(\mathbf{y}^* | \boldsymbol{\beta}_c, \boldsymbol{\tau}_c) \pi^b(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c) d\boldsymbol{\beta}_c d\boldsymbol{\tau}_c \\
&= \int p^{1-b}(\mathbf{y}^* | \boldsymbol{\beta}_c, \boldsymbol{\tau}_c) \frac{p^b(\mathbf{y}^* | \boldsymbol{\beta}_c, \boldsymbol{\tau}_c) \pi(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c | M_c)}{\int p^b(\mathbf{y}^* | \boldsymbol{\beta}_c, \boldsymbol{\tau}_c) \pi(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c | M_c) d\boldsymbol{\beta}_c d\boldsymbol{\tau}_c} d\boldsymbol{\beta}_c d\boldsymbol{\tau}_c \\
&= \frac{\int p(\mathbf{y}^* | \boldsymbol{\beta}_c, \boldsymbol{\tau}_c) \pi(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c | M_c) d\boldsymbol{\beta}_c d\boldsymbol{\tau}_c}{\int p^b(\mathbf{y}^* | \boldsymbol{\beta}_c, \boldsymbol{\tau}_c) \pi(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c | M_c) d\boldsymbol{\beta}_c d\boldsymbol{\tau}_c}. \tag{2.9}
\end{aligned}$$

The size of the training fraction b should be chosen carefully. When b is too small, the denominator in Equation (2.9) may diverge. If b is too large, a substantial part of the integrated likelihood is used to train the prior on the parameters, and then the remaining information in the integrated likelihood used to update the prior model probabilities will lead to less distinctive posterior model probabilities. Here, we consider a training fraction size equal to $b = m/n$, where m is the equivalent training size. To guide the choice of m in our considered GLMM context, we use the fact that for LMMs with the reference prior proposed by [34] the minimal value of m that guarantees propriety of the fractional integrated likelihood is $p + 1$ [53]. In particular, in all the GLMM applications we present in Section 2.6, the training fraction $b = (p + 1)/n$ yields well-defined Bayes factors.

Then, the FBF of model M_c versus model M_l is defined as $BF_{cl}^b = \frac{q_c(b, \mathbf{y}^*)}{q_l(b, \mathbf{y}^*)}$. Next, we compute the posterior probability of model M_c as proportional to the FBF, BF_{cl}^b , times the prior probability of model M_c , that is $P^b(M_c | \mathbf{y}) = BF_{cl}^b \times P(M_c) / \left[\sum_{k=1}^C BF_{kl}^b \times P(M_k) \right]$.

2.5 Simulation Study

To investigate the performance of our proposed model selection methods ARM and HCM when compared to the widely used DIC, WAIC and marginal likelihood computed by INLA, we perform a simulation study for different combinations of parameter settings. Here we

present results for Poisson GLMMs. In the Section 2.8.4 we present results for Bernoulli GLMMs. For each combination of parameter settings, we generate 100 datasets. We simulate samples on regular square grids and consider three sample sizes, $n = 100, 400,$ and 900 . Each sample may have spatial dependence based on a first-order neighborhood structure modeled with a vector of spatial random effects $\boldsymbol{\alpha}_1$ following the ICAR distribution given in Equation (2.1). For the variance component τ_1 of the spatial random effects, we consider values $0, 0.03, 0.05, 0.1,$ or 1 , where $\tau_1 = 0$ implies no spatial dependence. We also consider the possibility of overdispersion random effect $\boldsymbol{\alpha}_2$ in the model. We set the variance component τ_2 of the overdispersion random effect to $0, 0.05, 0.1, 0.5,$ or 1 , where $\tau_2 = 0$ implies no overdispersion. We consider 4 candidate covariates x_{1i}, x_{2i}, x_{3i} and x_{4i} sampled from a standard normal distribution. We assume that $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, 0, 0)^\top$, thus the last two covariates are not in the true model. Here β_0 is the intercept, with values equal to $1, 2,$ or 4 . We let $\beta_1 = \beta_2$ with values $0, 0.1, 0.2, 0.3, 0.5,$ or 1 . When β_1 and β_2 are both equal to 0 , there is no covariate in the true model. Conditionally independent Poisson observations y_i are generated with the GLMM $y_i | \lambda_i \stackrel{ind}{\sim} \text{Poisson}(\lambda_i), i = 1 \dots n$, with $\log \lambda_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \alpha_{1i} + \alpha_{2i}$, spatial random effects $\boldsymbol{\alpha}_1 \sim N(\mathbf{0}, \tau_1 \boldsymbol{\Sigma})$, and overdispersion random effects $\boldsymbol{\alpha}_2 \sim N(\mathbf{0}, \tau_2 \mathbf{I})$.

For each parameter setting, there are $C = 64$ candidate models in total. Specifically, there are 2^4 possible combinations of fixed effects. In addition, there are 4 possible combinations of random effects types, one with both spatial random effects and overdispersion random effects, one with only spatial random effects, one with only overdispersion random effects, and one without any random effects. We calculate posterior model probabilities for all 64 models, and we compute posterior inclusion probabilities for each of the 4 covariates, for the spatial random effect, and for the overdispersion random effect.

We compare our model selection methods ARM and HCM to the DIC, the WAIC and

marginal likelihood computed by the R package INLA [56]. For the ARM and HCM, we decide to include a component in the selected model if the posterior inclusion probability of that component is larger or equal to 0.5, that is, if that component is in the median probability model [4]. For the criteria computed by INLA, we select the model with the lowest DIC and WAIC values, and the highest marginal likelihood, respectively. For the three criteria computed by INLA, we consider the INLA default prior specification as well as our proposed AR prior and HC prior.

Because currently the most widely used criteria for Bayesian selection of GLMMs are the DIC and WAIC computed with INLA default priors, here we compare these criteria with our ARM and HCM. We present a comparison of our methods ARM and HCM to DIC and WAIC computed using our AR and HC priors in Section 2.8.4. The conclusions are similar to those for DIC and WAIC computed with default INLA priors presented here. Figure 2.1 presents the probability of each competing method selecting the correct covariate structure as a function of the value of their regression coefficients $\beta_1 = \beta_2$. Here, there are spatial random effects with $\tau_1 = 0.05$ and overdispersion random effects with $\tau_2 = 0.05$. Three sample sizes are considered: $n = 100, 400, 900$. Two values for the intercept are considered: $\beta_0 = 1$ and 4. Figure 2.1 shows that the ARM and HCM perform much better than the DIC and the WAIC computed with INLA's default priors. For example, in the most challenging case considered with $n = 100$ and $\beta_0 = 1$, the ARM and HCM have a higher probability than the DIC and WAIC of selecting the correct covariate structure when their regression coefficients β_1 and β_2 are zero. In addition, as the value of $\beta_1 = \beta_2$ increases, the probability of the ARM and HCM to correctly select the true non-null covariates x_1 and x_2 increases more quickly than that of the DIC and the WAIC. Finally, the probability of ARM and HCM to correctly select the two non-null regressors increases much closer to one than those of the DIC and WAIC as the sample size increases and as the intercept value increases. As

the sample size increases, the probability of ARM and HCM detecting covariates with small coefficients increases substantially. For example, the left panels of Figure 2.1 show that when the intercept is equal to 1, the probabilities of our proposed methods choosing the correct covariates structure when the coefficient is equal to 0.1 are about 10%, 60%, and 90% for sample sizes 100, 400, and 900, respectively.

Figure 2.2 investigates the impact of different magnitudes of the variance components on the probability of selecting the correct covariate structure as a function of the value of the regression coefficients $\beta_1 = \beta_2$. Panels (a) and (b) of Figure 2.2 present settings with small ($\tau_1 = 0.01$ and $\tau_2 = 0$) and large ($\tau_1 = 1$ and $\tau_2 = 1$) variance components, respectively. In both panels, the sample size is $n = 400$ and the intercept is $\beta_0 = 1$. In the small variance components setting, ARM and HCM perform comparably to the DIC and WAIC for small values of $\beta_1 = \beta_2$, but our methods ARM and HCM greatly outperform the DIC and WAIC for moderate to large values of $\beta_1 = \beta_2$. Meanwhile, in the more challenging large variance components setting, when $\beta_1 = \beta_2 = 0$, our ARM and HCM correctly select the model with no regressor in the model for 100% of the simulated datasets samples. In contrast, when $\beta_1 = \beta_2 = 0$, the DIC and WAIC select the wrong covariate structure for 20% of the simulated datasets, respectively. Finally, as the magnitude of $\beta_1 = \beta_2$ increases, in comparison to the DIC and WAIC, ARM and HCM achieve much higher probabilities of selecting the correct model.

Figure 2.3 presents the probability of selecting correct spatial random effects structure as a function of the value of the variance component for the spatial random effects. Results are shown for sample sizes $n = 100, 400$ and 900 , and variance of overdispersion random effects $\tau_2 = 0$ and 0.1 . Figure 2.3 shows that the DIC and WAIC have low probability of selecting the model with no spatial random effects when the true model does not have spatial random effects; In addition, this performance does not improve much as the sample size increases

from 400 to 900. In contrast, our methods ARM and HCM have large probabilities of selecting the correct spatial random effects structure when the true model does not have spatial random effects, and have large probabilities of selecting spatial random effects when the variance component for the spatial random effects is large. Finally, the performance of ARM and HCM at correctly detecting spatial dependence greatly improves as the sample size increases.

ARM, HCM, DIC and WAIC's performance when selecting overdispersion random effects is similar to selecting spatial random effects. Figure 2.4 in the supporting information presents the probability of selecting correct overdispersion structure as a function of the value of the variance for overdispersion. Figure 2.4 shows that the DIC and WAIC have low probability of selecting a model with no overdispersion random effects even when overdispersion is not present in the true model, and this undesirable performance does not improve much when the sample size increases. In contrast, our methods ARM and HCM have large probabilities of selecting correct overdispersion structure when overdispersion is not present in the true model, and have large probabilities of selecting overdispersion when the proportion of variance due to overdispersion is large. Finally, the performance of ARM and HCM at correctly detecting overdispersion greatly improves as the sample size increases.

Figures 2.15, 2.16, and 2.17 present a comparison of the performance of INLA marginal likelihood with our ARM and HCM methods. Figure 2.15 shows that INLA marginal likelihood with INLA's default priors is worse than our methods at selecting covariates when coefficients of covariates are small. INLA marginal likelihood with INLA's default prior or INLA marginal likelihood with our proposed priors are better than our methods ARM and HCM when the regression coefficient is large. For spatial random effects inclusion, Figure 2.16 shows that INLA marginal likelihood with any of the considered priors has difficulties to detect spatial random effects. For overdispersion random effects, Figure 2.17 shows that when

there is no spatial random effects in the model, INLA marginal likelihood can correctly select overdispersion random effects. However, when there are spatial random effects in the model, marginal likelihood computed by INLA cannot correctly select overdispersion random effects. In summary, INLA marginal likelihood with our proposed priors works well for selection of regressors but does not work well for the selection of random effects. Meanwhile, our ARM and HCM methods work well in both cases.

2.6 Case Studies

2.6.1 Longitudinal Epilepsy Seizure Data

We analyze a dataset on epilepsy seizures previously analyzed by [67], [9], and others. The data were collected in four biweekly visits of 59 epileptics during a clinical trial to evaluate the effectiveness of a drug to control seizures [39]. The response variable is the number of seizures y_{ij} for patient i on visit j . The most general model we consider is $y_{ij}|\mu_{ij} \stackrel{ind}{\sim} \text{Poisson}(\mu_{ij})$, with $\log(\mu_{ij}) = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \alpha_{1i} + z_j \alpha_{2i} + \alpha_{3ij}$, $\boldsymbol{\alpha}_1 \sim N(\mathbf{0}, \tau_1 I_{59})$, $\boldsymbol{\alpha}_2 \sim N(\mathbf{0}, \tau_2 I_{59})$, and $\boldsymbol{\alpha}_3 \sim N(\mathbf{0}, \tau_3 I_{236})$, $i = 1 \dots 59$ and $j = 1 \dots 4$, where \mathbf{x}_{ij} denotes a 6-dimensional vector with a one for intercept and 5 covariates. The 59 subjects were randomly assigned to a new drug or a placebo. The first covariate is the treatment indicator (Trt), where Trt=1 indicates that the patient received the treatment and Trt=0 indicates that the patient received the placebo. The second covariate is the baseline level of seizures (Base), equal to the logarithm of the average number of epileptic seizures per two weeks recorded in the 8-week period before treatment. The third covariate is the interaction term of Base and Trt. The fourth covariate is the logarithm of age (Age). And the fifth covariate is the visit number (Visit), with the 4 visits coded as -3, -1, 1 and 3. [9] mentioned that preliminary analysis indicated that the

counts were substantially lower during the fourth visit. Thus, they also define a binary variable $V4$, such that $V4=1$ indicates the fourth visit and $V4=0$ indicates the other visits. In the model above, $\boldsymbol{\beta}$ is the vector of regression coefficients, $\boldsymbol{\alpha}_1 = (\alpha_{11}, \dots, \alpha_{1\ 59})$ is the vector of patient-specific random effects, $\boldsymbol{\alpha}_2 = (\alpha_{21}, \dots, \alpha_{2\ 59})$ is the vector of patient-specific random effects for the slope of the variable Visit with $\mathbf{z} = (-0.3, -0.1, 0.1, 0.3)$, and $\boldsymbol{\alpha}_3 = (\alpha_{311}, \dots, \alpha_{3\ 59\ 1}, \alpha_{312}, \dots, \alpha_{3\ 59\ 2}, \dots, \alpha_{314}, \dots, \alpha_{3\ 59\ 4})$ is the vector of overdispersion random effects.

The covariates Trt, Base, Age and Visit can be included in the model independently. However, the interaction term between Trt and Base is only included when both Trt and Base are in the model. Thus, there are 20 possible combinations of covariates. For the dependence structure, we follow the four cases considered by [9]: no random effects in the model; only patient-specific random effects $\boldsymbol{\alpha}_1$; $\boldsymbol{\alpha}_1$ and overdispersion random effects $\boldsymbol{\alpha}_3$; $\boldsymbol{\alpha}_1$ and patient-specific random effects for the slope of the variable Visit $\boldsymbol{\alpha}_2$. Finally, we assume that the vectors of random effects $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_3$ are independent. Therefore, the model space has 80 models, composed by 20 combinations of covariates and 4 possible settings of random effects.

Table 2.1: Epilepsy data: posterior inclusion probabilities of fixed and random effects

	variable	ARM	HCM
fixed effect	Base	1	1
	Trt	0.14	0.04
	Trt \times Base	0	0
	Age	0.03	0.01
	V4	0.12	0.11
random effect	$\boldsymbol{\alpha}_1$	1	1
	$\boldsymbol{\alpha}_2$	0	0
	$\boldsymbol{\alpha}_3$	1	1

Table 2.1 presents the posterior inclusion probabilities of the fixed and random effects. Both the ARM and the HCM indicate that the baseline level of seizures (Base) should be included

in the model. However, the posterior inclusion probabilities do not provide support for any of the other covariates. Further, both ARM and HCM strongly indicate that α_2 , the patient-specific random effects for the slope of the variable Visit should not be included in the model. Finally, both ARM and HCM strongly indicate the need to include the patient-specific random effect α_1 and overdispersion random effect α_3 .

Table 2.3 presents a summary of the model selection results for the epilepsy data by comparing methods ARM, HCM, DIC and WAIC. A check mark appears next to the effects (rows) selected by each method (column). In addition, Table 2.3 presents the selection of fixed effects and variance components based on estimates and standard errors reported by [9] for two models fitted with PQL, which we denote by PQL1 and PQL2. Table 2.4 presents estimates and standard errors for the parameters based on the full model. Model PQL1 includes random effects α_1 and α_2 while Model PQL2 includes random effects α_1 and α_3 . Interestingly, while the original PQL method cannot choose between Model PQL1 or Model PQL2, our ARM and HCM clearly show that the data support exclusion of random effect α_2 and inclusion of random effects α_1 and α_3 . Further, the DIC and WAIC agree with the ARM and HCM and also choose random effects α_1 and α_3 . Furthermore, in terms of fixed effects the DIC and WAIC are the least parsimonious, choosing Base, Trt and Trt \times Base, while PQL chooses Base and Trt. Finally, the ARM and HCM are the most parsimonious and choose only the Base covariate.

In addition to selecting more parsimonious models, our ARM and HCM provide more definitive support for the inclusion or exclusion of each effect in the form of Bayesian posterior probabilities. For example, the posterior inclusion probabilities of the patient-level random effects α_1 , overdispersion random effects α_3 , and the covariate Base are all equal to one. Further, there is a lot less support for the covariate V4, which has posterior inclusion probability of 0.12 by the ARM and 0.11 by the HCM. Furthermore, both ARM and HCM provide pos-

terior inclusion probability equal to zero for the interaction between Trt and Base. Finally, the simulation study presented in Section 2.5 shows that we can rely on the uncertainty quantification provided by the ARM and HCM.

2.6.2 Spatial Lip Cancer Data

In this section, we present an analysis of the Scottish lip cancer dataset previously analyzed by [14], [9], [19], among many others. This dataset provides the number of male lip cancer cases in the 56 counties of Scotland during the period 1975-1980, as well as the percentage of the work force employed in agriculture, fishing, or forestry (AFF) as a covariate. The most general model we consider is $y_i|\mu_i \stackrel{ind}{\sim} \text{Poisson}(\mu_i)$, $\log(\mu_i) = \log(n_i) + \mathbf{x}_i^\top \boldsymbol{\beta} + \alpha_{1i} + \alpha_{2i}$, $\boldsymbol{\alpha}_1 \sim N(\mathbf{0}, \tau_1 \boldsymbol{\Sigma})$, and $\boldsymbol{\alpha}_2 \sim N(\mathbf{0}, \tau_2 \mathbf{I}_{56})$, $i = 1 \dots 56$, where n_i is the expected number of lip cancer cases in the i^{th} county, calculated based on the age distributions by counties. In this analysis, the n_i 's are assumed to be known constants. In addition, the vector \mathbf{x}_i is a two-dimensional vector with one as the first element and AFF for the i^{th} county as the second element. Further, $\boldsymbol{\alpha}_1$ is a vector of spatial random effects following a sum-zero constrained Gaussian Intrinsic Conditional Autoregressive model [33] and modeled by Equation (2.1). Finally, $\boldsymbol{\alpha}_2$ is a vector of overdispersion random effects.

There are two possible combinations for the fixed effects: with or without the covariate AFF. For the random effects, we follow the options considered by [9]. When $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are in the model at the same time, the PQL estimate of the overdispersion variance τ_2 is 0. Thus, we consider models with only three random effects combinations: spatial random effects $\boldsymbol{\alpha}_1$; overdispersion random effects $\boldsymbol{\alpha}_2$; and no random effects.

Table 2.2 presents the posterior inclusion probabilities for the fixed and random effects. Both the ARM and HCM select the model with the covariate AFF and spatial random effect $\boldsymbol{\alpha}_1$.

Table 2.2: Lip cancer data: posterior inclusion probabilities of fixed and random effects

	variable	ARM	HCM
fixed effect	AFF	0.93	0.92
random effect	α_1	1	1
	α_2	0	0

Table 2.5 presents the DIC and WAIC for the 6 models considered. In contrast to the results of the ARM and HCM, DIC and WAIC select the model without the covariate AFF but with spatial random effect α_1 . Table 2.6 summarizes model selection results for the ARM, HCM, DIC, WAIC, as well as the selection of model components based on PQL methods reported by [9] for two models: PQL1 includes α_1 and PQL2 includes α_2 . Results from PQL for the AFF regressor agree with the results from the HCM and ARM. An advantage of the HCM and ARM over PQL is that they clearly indicate that the model should include a spatial random effect and not include overdispersion.

2.7 Discussion

We have proposed a novel Bayesian method for model selection for GLMMs. Our approach is based on a pseudo likelihood approximation of GLMMs by LMMs leading to a closed form solution for integrating out the random effects from the likelihood. We consider two priors for the model parameters. First, we use an approximate reference prior that is uniform for the fixed effects and has the behavior of the half-Cauchy prior for the variance parameters. Second, while keeping the improper uniform prior for the fixed effects, we consider the half-Cauchy prior for the square root of the variance parameters [21, 52]. Finally, to deal with the prior impropriety we have developed a fractional Bayes factor approach.

The simulation study has shown that our proposed methods ARM and HCM perform well for correctly selecting both covariates and dependence structure. ARM and HCM assign

high posterior inclusion probability to covariates with large coefficients and also high posterior inclusion probability to random effects with large variance components. In particular, ARM and HCM are better than DIC and WAIC at correctly selecting covariates. In cases where random effects have large variances, the ability of DIC and WAIC to correctly select covariates is tremendously reduced. In contrast, ARM and HCM do not suffer as badly when selecting covariates in the presence of random effects with large variances. In addition, DIC and WAIC have high false positive rates and often select null fixed and random effects. In contrast, ARM and HCM correctly assign low posterior inclusion probability to null covariates and to null random effects. We also compared our methods with marginal likelihood computed by INLA. Our results show that when we use INLA with our priors instead of the default INLA priors, the marginal likelihood computed by INLA and the marginal likelihood computed by our pseudo likelihood approach work similarly for the selection of regression coefficients. However, the marginal likelihood computed by INLA does not work well for the selection of spatial random effects and overdispersion random effects. Therefore, it seems that our pseudo likelihood approximation works better than the INLA approximation to the marginal likelihood for the selection of random effects.

We illustrate the application of our proposed methods ARM and HCM with three case studies. The data for these three case studies are available in the supporting information. In the first case study, we consider epilepsy seizures as a type of longitudinal count data. ARM and HCM are more parsimonious, selecting baseline covariate, patient-level random effects and overdispersion random effects. DIC and WAIC select two more covariates: treatment and interaction term between baseline and treatment. In the second case study, we study Scottish lip cancer data as a type of spatial count data. Our methods ARM and HCM select spatial dependence and covariate AFF, whereas DIC and WAIC select the model without covariate AFF but include spatial random effects. In the third case study, presented in

Section 2.8.3, we look at binary salamander mating data. For fixed effects, our methods ARM and HCM select WSF and WSF×WSM, whereas DIC and WAIC select all three covariates. For random effects, our two methods ARM and HCM have totally different results than DIC and WAIC: while DIC and WAIC select male random effect, our methods ARM and HCM select female random effect. Given the results from the simulation study, we recommend the models selected by ARM and HCM.

There are many potential avenues for future research. One possible future research topic is the use of Bayesian model averaging for computing credible intervals for regression coefficients. This can be facilitated by the fact that our methods provide posterior probabilities for different models. Another promising research direction is the use of nonlocal priors [26, 27, 73] for the fixed effects. Finally, another possible research topic is model selection for GLMMs when the number of possible regressors is much larger than the sample size. We are currently working on the latter two research topics and will report the results in a future manuscript.

2.8 Supplementary Material

2.8.1 The pseudo likelihood approach

In the pseudo likelihood approach, parameters are set to their current estimates, and the estimates are updated iteratively. Let $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ be current estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, respectively. Write the observation y_i as the sum of its mean and an error term ϵ_i , that is $y_i = \mu_i + \epsilon_i$, with mean $\mu_i = B'_i(\eta_i)$ and variance $Var(\epsilon_i) = v_i = B''_i(\eta_i)$. Now expand μ_i in a first-order

Taylor expansion around $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ to obtain

$$\begin{aligned}
y_i &= \mu_i + \epsilon_i \\
&= B'_i \left(\mathbf{x}_i^T \boldsymbol{\beta} + \sum_j \mathbf{z}_{ij}^T \boldsymbol{\alpha}_j \right) + \epsilon_i \\
&\approx B'_i \left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}} + \sum_j \mathbf{z}_{ij}^T \hat{\boldsymbol{\alpha}}_j \right) + B''_i \left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}} + \sum_j \mathbf{z}_{ij}^T \hat{\boldsymbol{\alpha}}_j \right) \left[\mathbf{x}_i^T (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \sum_j \mathbf{z}_{ij}^T (\boldsymbol{\alpha}_j - \hat{\boldsymbol{\alpha}}_j) \right] + \epsilon_i.
\end{aligned} \tag{S.1}$$

Denote the current estimate of the mean vector $\boldsymbol{\mu}$ evaluated at $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ by $\hat{\boldsymbol{\mu}}$. That is,

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} B'_1 \left(\mathbf{x}_1^T \hat{\boldsymbol{\beta}} + \sum_j \mathbf{z}_{1j}^T \hat{\boldsymbol{\alpha}}_j \right) \\ \vdots \\ B'_n \left(\mathbf{x}_n^T \hat{\boldsymbol{\beta}} + \sum_j \mathbf{z}_{nj}^T \hat{\boldsymbol{\alpha}}_j \right) \end{bmatrix}.$$

In addition, denote the current estimate of the covariance matrix of $\boldsymbol{\epsilon}$ evaluated at $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ by $\hat{\mathbf{V}}$, that is

$$\hat{\mathbf{V}} = \begin{bmatrix} B''_1 \left(\mathbf{x}_1^T \hat{\boldsymbol{\beta}} + \sum_j \mathbf{z}_{1j}^T \hat{\boldsymbol{\alpha}}_j \right) & & \\ & \ddots & \\ & & B''_n \left(\mathbf{x}_n^T \hat{\boldsymbol{\beta}} + \sum_j \mathbf{z}_{nj}^T \hat{\boldsymbol{\alpha}}_j \right) \end{bmatrix}.$$

Reorganize Equation (S.1) by moving $\hat{\boldsymbol{\mu}}$ to the left side, pre-multiplying $\hat{\mathbf{V}}^{-1}$ and then moving $\mathbf{X}\hat{\boldsymbol{\beta}}$ and $\mathbf{Z}_j\hat{\boldsymbol{\alpha}}_j$ to the left side. Let \mathbf{y}^* be equal to the left side of the resulting equation, that is

$$\mathbf{y}^* = \hat{\mathbf{V}}^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}) + \mathbf{X}\hat{\boldsymbol{\beta}} + \sum_j \mathbf{Z}_j\hat{\boldsymbol{\alpha}}_j.$$

The vector \mathbf{y}^* is known as the vector of adjusted observations, which is computed as a function of the current estimates of fixed effects $\hat{\boldsymbol{\beta}}$ and random effects $\hat{\boldsymbol{\alpha}}_j$. The right side of the resulting equation is $\mathbf{X}\boldsymbol{\beta} + \sum_j \mathbf{Z}_j \boldsymbol{\alpha}_j + \hat{\mathbf{V}}^{-1} \boldsymbol{\epsilon}$. Hence, we obtain the following approximate model for the adjusted observations

$$\mathbf{y}^* \approx \mathbf{X}\boldsymbol{\beta} + \sum_j \mathbf{Z}_j \boldsymbol{\alpha}_j + \hat{\mathbf{V}}^{-1} \boldsymbol{\epsilon}.$$

Further, assuming that $\hat{\mathbf{V}}^{-1} \mathbf{V} \hat{\mathbf{V}}^{-1} \approx \hat{\mathbf{V}}^{-1}$ and applying properties of expectation and variance, we get

$$\begin{aligned} E(\mathbf{y}^*) &= \mathbf{X}\boldsymbol{\beta}, \\ \text{Var}(\mathbf{y}^*) &\approx \sum_j \tau_j \mathbf{Z}_j \boldsymbol{\Sigma}_j \mathbf{Z}_j^T + \hat{\mathbf{V}}^{-1}. \end{aligned}$$

If we further assume that $\boldsymbol{\epsilon}$ has an approximate normal distribution, then

$$\mathbf{y}^* \sim N \left(\mathbf{X}\boldsymbol{\beta}, \sum_j \tau_j \mathbf{Z}_j \boldsymbol{\Sigma}_j \mathbf{Z}_j^T + \hat{\mathbf{V}}^{-1} \right).$$

As we explain below, estimates of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}_j$, and τ_j are updated iteratively. After convergence, we have the final pseudo response data \mathbf{y}^* and approximate LMM

$$\begin{aligned} \mathbf{y}^* &\approx \mathbf{X}\boldsymbol{\beta} + \sum_j \mathbf{Z}_j \boldsymbol{\alpha}_j + \hat{\mathbf{V}}^{-1} \boldsymbol{\epsilon}, \\ \boldsymbol{\alpha}_j &\sim N(\mathbf{0}, \tau_j \boldsymbol{\Sigma}_j), \\ \boldsymbol{\epsilon} &\sim N(\mathbf{0}, \mathbf{V}). \end{aligned}$$

The closed form of the likelihood function with respect to the unknown parameters is

$$L(\boldsymbol{\beta}, \boldsymbol{\tau} | \mathbf{y}^*) = (2\pi)^{-\frac{n}{2}} \left| \sum_j \tau_j \mathbf{Z}_j \boldsymbol{\Sigma}_j \mathbf{Z}_j^T + \widehat{\mathbf{V}}^{-1} \right|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta})^T \left(\sum_j \tau_j \mathbf{Z}_j \boldsymbol{\Sigma}_j \mathbf{Z}_j^T + \widehat{\mathbf{V}}^{-1} \right)^{-1} (\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}) \right\}.$$

Let $\widehat{\tau}_j$ be current estimates of the variance components. We update $\widehat{\boldsymbol{\beta}}$ with the conditional posterior mean of $\boldsymbol{\beta}$, given by $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T \widehat{\mathbf{H}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \widehat{\mathbf{H}}^{-1} \mathbf{y}^*$, which is also the best linear unbiased estimator of $\boldsymbol{\beta}$, where $\widehat{\mathbf{H}} = \sum_j \widehat{\tau}_j \mathbf{Z}_j \boldsymbol{\Sigma}_j \mathbf{Z}_j^T + \widehat{\mathbf{V}}^{-1}$. Further, we use the conditional posterior mean as the estimator of $\boldsymbol{\alpha}$. Note that

$$\begin{pmatrix} \mathbf{y}^* \\ \boldsymbol{\alpha} \end{pmatrix} \sim N \left(\begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \widehat{\mathbf{H}} & \widehat{\boldsymbol{\tau}} \mathbf{Z} \boldsymbol{\Sigma} \\ \widehat{\boldsymbol{\tau}} \boldsymbol{\Sigma} \mathbf{Z}^T & \widehat{\boldsymbol{\tau}} \boldsymbol{\Sigma} \end{pmatrix} \right),$$

$$\widehat{\boldsymbol{\alpha}} = E(\boldsymbol{\alpha} | \mathbf{y}^*) = \widehat{\boldsymbol{\tau}} \boldsymbol{\Sigma} \mathbf{Z}^T \widehat{\mathbf{H}}^{-1} (\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}). \quad (\text{S.2})$$

Substitute $\boldsymbol{\beta}$ by its current estimate $\widehat{\boldsymbol{\beta}}$ in Equation (S.2). Then, update the current estimates of $\boldsymbol{\alpha}$ to

$$\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\tau}} \boldsymbol{\Sigma} \mathbf{Z}^T \widehat{\mathbf{H}}^{-1} (\mathbf{y}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}).$$

And then, we use maximum likelihood (ML) to estimate $\boldsymbol{\tau}$. The profile log likelihood of $\boldsymbol{\tau}$, follows as:

$$\begin{aligned} \log L(\boldsymbol{\tau}|\mathbf{y}^*) &\propto -\frac{1}{2} \log \left| \sum_j \tau_j \mathbf{Z}_j \boldsymbol{\Sigma}_j \mathbf{Z}_j^T + \widehat{\mathbf{V}}^{-1} \right| \\ &\quad - \frac{1}{2} (\mathbf{y}^* - \mathbf{X}\widehat{\boldsymbol{\beta}})^T \left(\sum_j \tau_j \mathbf{Z}_j \boldsymbol{\Sigma}_j \mathbf{Z}_j^T + \widehat{\mathbf{V}}^{-1} \right)^{-1} (\mathbf{y}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}), \\ \widehat{\boldsymbol{\tau}} &= \operatorname{argmax} \log L(\boldsymbol{\tau}|\mathbf{y}^*). \end{aligned}$$

We use the full model to estimate parameters and generate the pseudo response data \mathbf{y}^* . The algorithm that we use sets initial values of $\boldsymbol{\beta}$ to GLM estimates and the initial value of $\boldsymbol{\alpha}$ equal to $\mathbf{0}$. And then, we update these estimators until convergence. Lastly, we obtain the vector of adjusted observations \mathbf{y}^* from the LMM. The algorithm is as follows:

Algorithm 1 Pseudo likelihood approach

procedure PSEUDO LIKELIHOOD($\mathbf{y}, \mathbf{X}, \mathbf{Z}_j$)

Initial values: $\boldsymbol{\beta}^{(0)}$ = estimate from GLM, $\boldsymbol{\alpha}_j^{(0)} = \mathbf{0}$, $\tau_j^{(0)} = 0$.

Calculate $\boldsymbol{\mu}^{(0)}$, $\mathbf{V}^{(0)}$, $\mathbf{H}^{(0)}$ and $\mathbf{y}^{*(0)}$.

while $\boldsymbol{\beta}$, $\boldsymbol{\tau}$ not converge **do**

$$\boldsymbol{\beta}^{(t)} = (\mathbf{X}^T \mathbf{H}^{(t-1)} - \mathbf{X})^{-1} \mathbf{X}^T \mathbf{H}^{(t-1)} \mathbf{y}^{*(t-1)}$$

$$\boldsymbol{\alpha}_j^{(t)} = \tau_j^{(t-1)} \boldsymbol{\Sigma}_j \mathbf{Z}_j^T \mathbf{H}^{(t-1)} (\mathbf{y}^{*(t-1)} - \mathbf{X}\boldsymbol{\beta}^{(t)})$$

$$\boldsymbol{\tau}^{(t)} = \operatorname{argmax} \log L(\boldsymbol{\tau}|\mathbf{y}^{*(t-1)})$$

Update $\boldsymbol{\mu}^{(t)}$, $\mathbf{V}^{(t)}$, $\mathbf{H}^{(t)}$ and $\mathbf{y}^{*(t)}$

end while

end procedure

2.8.2 Additional tables

Table 2.3: Epilepsy data: summary of model selection results by competing methods

	variable	ARM	HCM	DIC	WAIC	PQL1	PQL2
fixed effect	Base	✓	✓	✓	✓	✓	✓
	Trt			✓	✓	✓	✓
	Trt × Base			✓	✓		
	Age						
	V4						
random effect	α_1	✓	✓	✓	✓	✓	✓
	α_2					✓	
	α_3	✓	✓	✓	✓		✓

Table 2.4: Estimates for epilepsy case study

Parameter	Estimate	Standard deviation
β_0	-1.29	1.14
β_1	0.86	0.13
β_2	-0.92	0.39
β_3	0.34	0.20
β_4	0.48	0.34
β_5	-0.10	0.08
τ_1	0.20	0.05
τ_2	0	0.27
τ_3	0.12	0.03

Table 2.5: Lip cancer data: DIC and WAIC for competing models

	α_1	α_1 +AFF	α_2	α_2 +AFF	-	AFF
DIC	298	299	313	311	590	451
WAIC	294	297	307	307	605	461

Table 2.6: Lip cancer data: model selection summary

		ARM	HCM	DIC	WAIC	PQL1	PQL2
fixed effect	AFF	✓	✓			✓	✓
random effect	α_1	✓	✓	✓	✓	✓	
	α_2						✓

Table 2.7: Estimates for lip cancer case study

Parameter	Estimate	Standard deviation
β_0	-0.28	0.12
β_1	0.43	0.13
τ_1	0.47	0.15
τ_2	0	0.05

2.8.3 Case Study - Binary Salamander Mating Data

Here we apply our model selection approaches to the well known salamander mating data set that has been analyzed by many authors such as [42], [9], and [64]. The response consists of mating occurrences between two salamander populations, White Side (WS) and Rough Butt (RB). There were three experiments. In each experiment, 10 RB males (RBM) and 10 WS males (WSM) were paired with 10 RB females (RBF) and 10 WS females (WSF). The first experiment took place in the Summer. In the second experiment, performed in the Fall, the salamanders were identical to those in the first experiment. The third experiment, performed in the Fall, used different salamanders than the first two studies. The binary observation was recorded as 1 when mating occurred and 0 otherwise. We consider two random effects, α_1 and α_2 , to account for variation among females and males, respectively. To satisfy GLMMs' conditional independence assumption, we follow the setup of [64] that considers only data from the first experiment performed in the Summer and the third experiment performed in the Fall with different salamanders. For fixed effects, we consider an intercept, an indicator for WSF, an indicator for WSM, and their interaction term. The most general model we consider is $y_{ij} | \alpha_1, \alpha_2 \stackrel{ind}{\sim} \text{Bernoulli}(p_{ij})$, $i = 1 \dots 40$ and $j = 1 \dots 40$,

with $\text{logit}(p_{ij}) = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \alpha_{1i} + \alpha_{2j}$, where p_{ij} is the probability of successful mating. Female random effects $\boldsymbol{\alpha}_1 \sim N(\mathbf{0}, \tau_1 I_{40})$, and male random effects $\boldsymbol{\alpha}_2 \sim N(\mathbf{0}, \tau_2 I_{40})$. Then, there are eight possible combinations of fixed effect inclusions. In addition, there are four possibilities for the random effects, including inclusion or exclusion of female random effects and inclusion or exclusion of male random effects. Therefore, there are 32 possible models.

Table 2.8: Salamander mating data: posterior inclusion probabilities of fixed and random effects

	variable	ARM	HCM
fixed effect	WSF	1	1
	WSM	0.46	0.45
	WSF \times WSM	1	1
random effect	$\boldsymbol{\alpha}_1$	0.91	0.81
	$\boldsymbol{\alpha}_2$	0.22	0.11

Table 2.8 presents posterior inclusion probabilities for the fixed and random effects. Both ARM and HCM strongly suggest to include covariate WSF, covariate WSF \times WSM and female random effects in the model. Table 2.8 provides a summary of the model selection results for all considered methods. The DIC and WAIC select the model with all potential covariates and the male random effects. [64] used a quasi-likelihood (QL) method to calculate estimates and standard errors of fixed effect coefficients and variances of random effects. According to their methods, covariates WSF and WSF \times WSM as well as both random effects are significant. While the random effect selection results from the QL method differ from ours, [64] found a much larger estimate of variance for the female random effects than that for the male random effects, with similar standard errors. This indicates that the female random effects may be more important, which is consistent with the results from ARM and HCM.

Table 2.9: Salamander mating data: model selection summary

	variable	ARM	HCM	DIC	WAIC	QL
fixed effect	WSF	✓	✓	✓	✓	✓
	WSM			✓	✓	
	WSF×WSM	✓	✓	✓	✓	✓
random effect	α_1	✓	✓			✓
	α_2			✓	✓	✓

Table 2.10: Estimates for salamander mating case study

Parameter	Estimate	Standard deviation
β_0	0.79	0.31
β_1	-2.34	0.48
β_2	-0.47	0.40
β_3	2.70	0.60
τ_1	1.16	0.67
τ_2	0	0.33

2.8.4 Additional simulation studies

Spatially correlated covariates

In this simulation study, observations y_i are generated with the Poisson GLMM $y_i|\lambda_i \stackrel{ind}{\sim} \text{Poisson}(\lambda_i)$, $i = 1 \dots n$, where mean $\log \lambda_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \alpha_{1i} + \alpha_{2i}$, spatial random effects $\alpha_1 \sim N(\mathbf{0}, \tau_1 \Sigma)$, and overdispersion random effects $\alpha_2 \sim N(\mathbf{0}, \tau_2 I)$. While in this simulation study x_1 and x_3 are still generated from standard normal distribution, x_2 and x_4 are generated from a Gaussian hierarchical model with mean zero and ICAR random effects that have the same covariance matrix $\tau_1 \Sigma$ as the vector of spatial random effects α_1 .

Figure 2.5 shows that our methods ARM and HCM have high probability of correctly selecting the correct covariate structure when coefficients are zero or coefficients are large. In addition, ARM and HCM have substantial improvement as the sample size increases.

Bernoulli data

In this simulation study, we follow the same setup of the salamander case study. We perform a simulation study with Bernoulli data. Specifically, we consider cases with n individuals, $n=40, 80,$ and 160 , where individual has 6 measurements. The intercept β_0 is equal to 0 or 0.75. In addition, there are 4 candidate covariates with corresponding coefficients equal to $-2.5, 2.5, \beta_3,$ and 0, where $\beta_3 = 0, 0.5, 1, 1.5, 2, 2.5$. Further, there is one candidate vector of longitudinal random effects with variance component equal to 0, 0.5, 1, 1.5, or 2. Thus, the model we consider in this new simulation study is

$$\begin{aligned} y_{ij}|p_{ij} &\stackrel{ind}{\sim} \text{Bernoulli}(p_{ij}), \quad i = 1, \dots, n, \quad j = 1, \dots, 6, \\ \log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \beta_0 - 2.5\mathbf{x}_1 + 2.5\mathbf{x}_2 + \beta_3\mathbf{x}_3 + 0\mathbf{x}_4 + Z\boldsymbol{\alpha}_1, \\ \boldsymbol{\alpha}_1 &\sim N(\mathbf{0}, \tau_1 I_{n \times n}). \end{aligned}$$

Figure 2.6 shows that our methods ARM and HCM can correctly select covariates $x_{1,2,3}$ in the case of Bernoulli data. When covariate x_3 is not in the model, DIC and WAIC have high probability of selecting the wrong covariate structure. In contrast, our methods ARM and HCM have higher probability of selecting the right covariate structure. When sample size is small and the value of β_3 is large, DIC and WAIC have higher probability than ARM and HCM to detect the correct covariates structure. When sample size becomes large, our methods have large improvement to correctly select covariates and are better than DIC and WAIC.

Figure 2.7 shows that for Bernoulli data, our methods ARM and HCM can correctly select longitudinal random effects with higher probability when sample size is larger or the value of variance component is larger. ARM and HCM struggle to correctly detect longitudinal

random effects when variance component is small. In addition, ARM and HCM correctly exclude longitudinal random effects when they are not in the true model. In contrast, DIC and WAIC have high probability of incorrectly selecting longitudinal random effects structure when there are no longitudinal random effects in the model. Therefore, selection of correct random effect structure seems to be a much more difficult problem for Bernoulli data than for Poisson data.

Comparison with BAS package

Here, we compare our methods ARM and HCM with the BAS package [15] for the case when the true model is a Poisson GLM, that is, when there are no random effects in the true model. Note that here we still assume a GLMM when using ARM and HCM. Meanwhile, the BAS package assumes a GLM. For ARM, HCM and BAS methods, we select the covariates with posterior inclusion probability larger than 0.5. Figure 2.8 presents the probability of selecting the correct covariate structure as a function of β_1 and β_2 . The BAS package performs better than ARM and HCM when the regression coefficients are larger. But surprisingly, our methods ARM and HCM perform much better than the BAS package when there is no covariate or covariates have small coefficients.

Comparison with DIC and WAIC computed by INLA with AR prior or HC prior

Here, we compare our methods ARM and HCM with DIC and WAIC computed by INLA but with the priors that we propose, which are flat priors for the regression coefficients, and AR prior or HC prior for the variance component of the random effects.

Figure 2.9 shows a comparison of our methods ARM and HCM with the DIC computed by INLA with our AR prior and HC prior. Figure 2.10 shows a comparison of our methods ARM

and HCM with the WAIC computed by INLA with our AR prior and HC prior. Figures 2.9 and Figure 2.10 show that even with our priors, the DIC and WAIC are worse than our methods ARM and HCM when selecting covariates. Figure 2.11 shows a comparison of our methods ARM and HCM with DIC computed by INLA with our AR prior and HC prior. Figure 2.12 shows a comparison of our methods ARM and HCM with WAIC computed by INLA with our AR prior and HC prior. Figure 2.11 and Figure 2.12 show that DIC and WAIC with our priors can correctly select spatial random effects. When there is no spatial random effects in the model, DIC and WAIC have a little trouble to detect the correct spatial random effects structure. Compared with DIC and WAIC with INLA's default priors, DIC and WAIC with our priors have higher probability to select correct spatial random effects structure when there are no spatial random effects in the model, which is large improvement. Figure 2.13 shows a comparison of our methods ARM and HCM with DIC computed by INLA with our AR prior and HC prior. Figure 2.14 shows a comparison of our methods ARM and HCM with WAIC computed by INLA package with our AR prior and HC prior. However, as shown in Figure 2.13 and Figure 2.14, DIC and WAIC with our priors still have trouble selecting the correct overdispersion random effects structure when there is no overdispersion random effects in the model.

Comparison with marginal likelihood computed by INLA

Here, we compare our methods ARM and HCM with marginal likelihood computed by INLA. Figure 2.15 shows that model selection using the marginal likelihood computed by INLA using INLA's default priors is worse than our methods at selecting covariates when the regression coefficients are small. Marginal likelihood methods with INLA's default prior or our proposed priors are better than our methods ARM and HCM when the regression coefficients are large. For spatial random effects inclusion, Figure 2.16 shows that model selection

using the marginal likelihood computed by INLA with any of the considered priors cannot detect spatial random effects. For overdispersion random effects, Figure 2.17 shows that when there is no other type of random effects in the model, marginal likelihood computed by INLA with our priors can correctly select overdispersion random effects. However, when there are spatial random effects in the model (right panel of Figure 2.17), marginal likelihood computed by INLA with any of the considered priors has a very high probability of erroneously detecting overdispersion when the true model has no overdispersion.

2.8.5 Additional figures

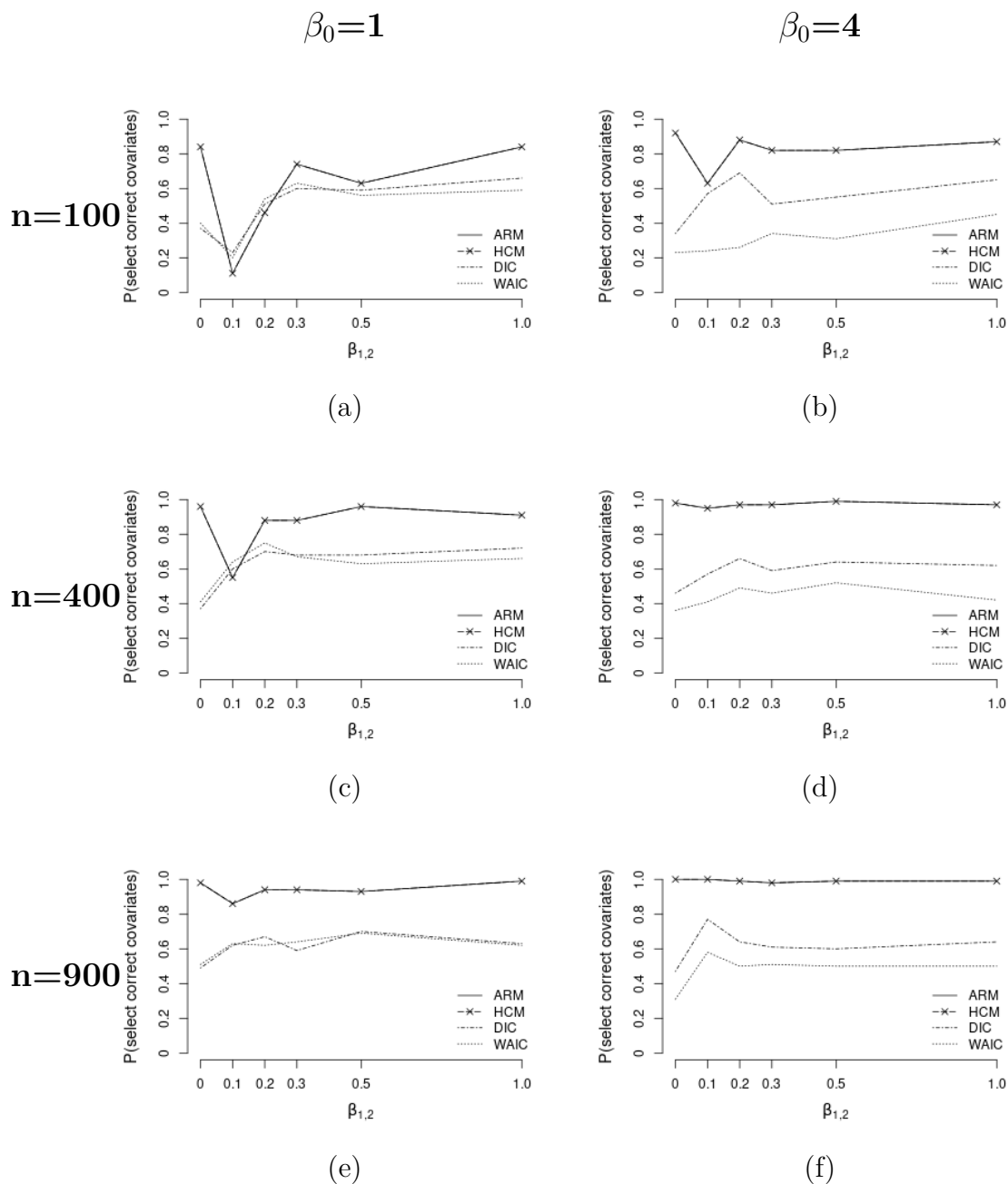


Figure 2.1: Probability of selecting the correct covariate structure as a function of the value of the regression coefficient, settings: $\tau_1 = 0.05$, $\tau_2 = 0.05$, $n=100$ (top row), $n=400$ (middle row), $n=900$ (bottom row), and $\beta_0 = 1$ (left column), $\beta_0 = 4$ (right column).

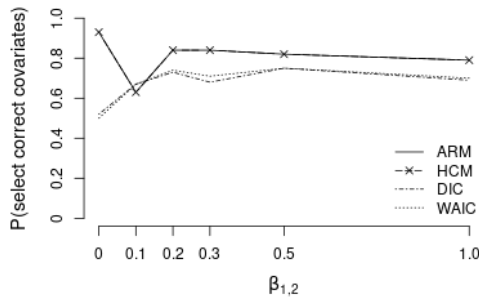
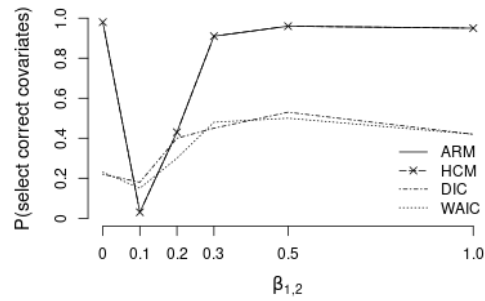
(a) $\tau_1 = 0.01, \tau_2 = 0$ (b) $\tau_1 = 1, \tau_2 = 1$

Figure 2.2: Probability of selecting the correct covariate structure as a function of the value of the regression coefficient, settings: (a) $\tau_1 = 0.01$ and $\tau_2 = 0$, and (b) $\tau_1 = 1$ and $\tau_2 = 1$, both with sample size $n = 400$ and intercept value $\beta_0 = 1$. (a) has weak dependence structure. (b) has strong dependence structure. Dependence structure can affect our method's performance for detecting covariates with small coefficients. However, the DIC and WAIC have difficulty detecting covariates even with large coefficients when spatial dependence is strong.

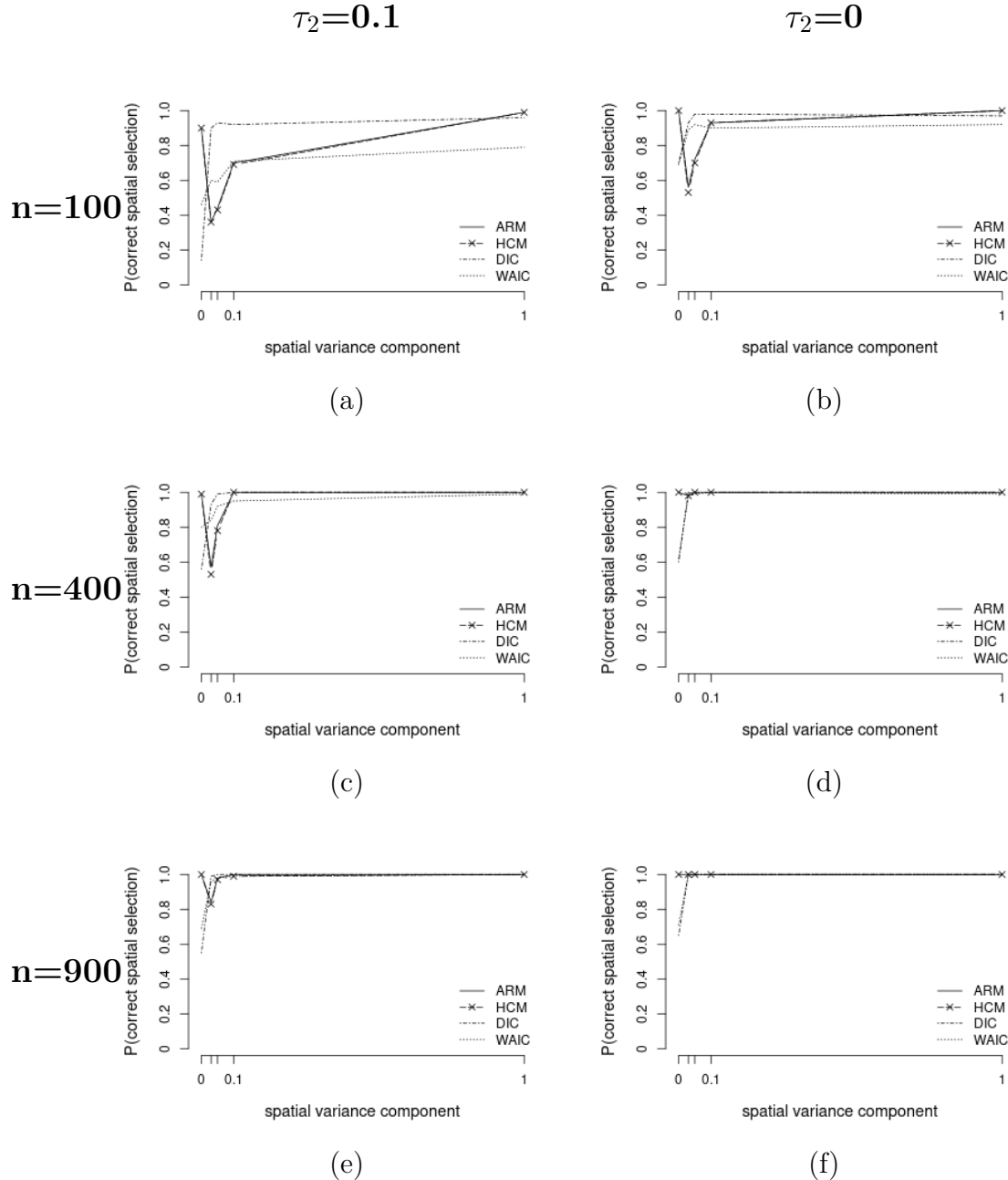


Figure 2.3: Probability of selecting the correct spatial random effects structure as a function of the value of variance component for spatial random effects. Settings: $\beta_0 = 2$, $\beta_1 = \beta_2 = 1$, $n=100$ (top row), $n=400$ (middle row), $n=900$ (bottom row), and $\tau_2 = 0.1$ (left column), $\tau_2 = 0$ (right column). If the spatial variance proportion is zero then there is no vector of spatial random effects in the model, and the correct decision is to not select the vector of spatial random effects.

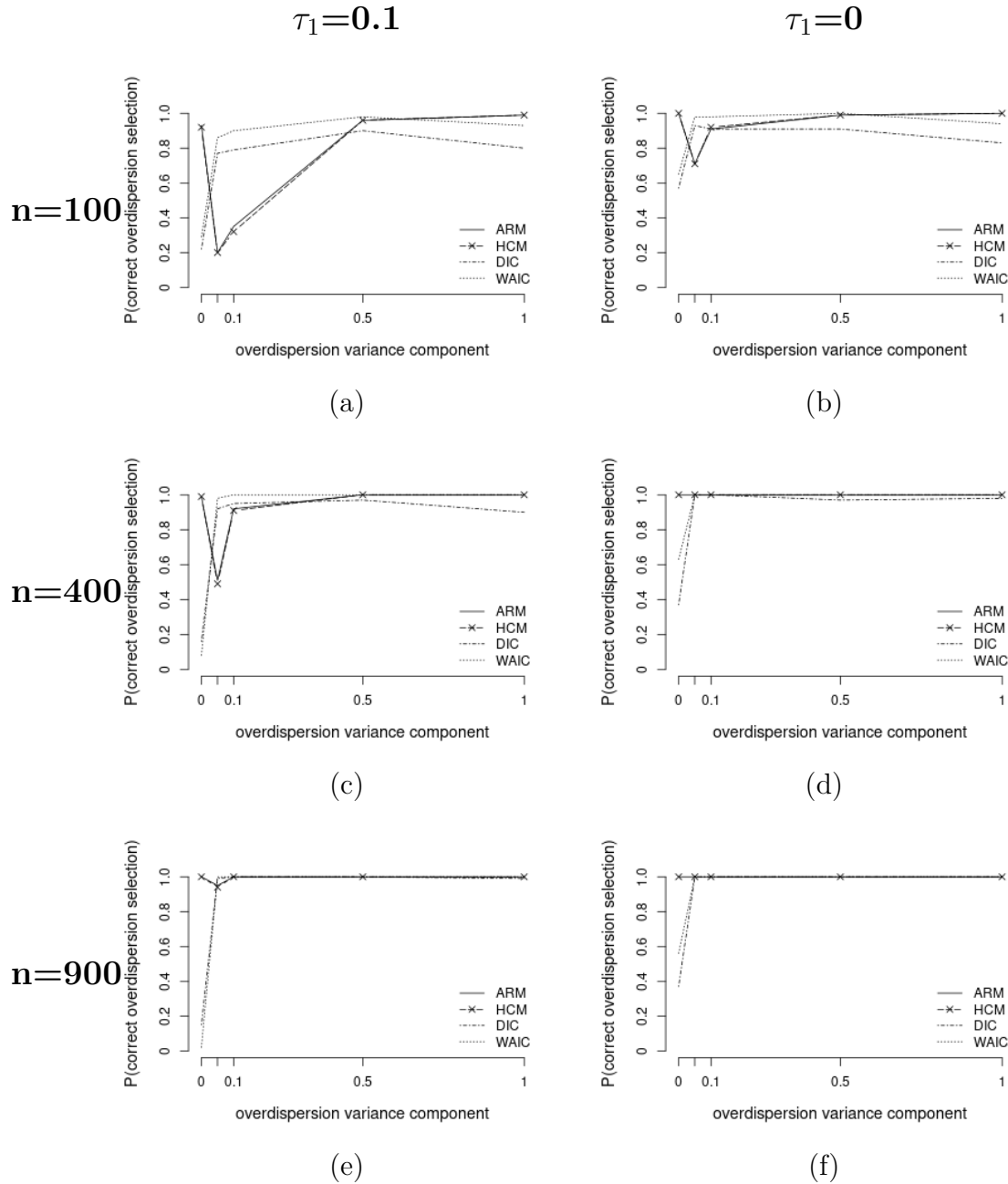


Figure 2.4: Probability of selecting the correct overdispersion random effects structure as a function of the value of variance component for overdispersion random effects. Settings: $\beta_0 = 2$, $\beta_1 = \beta_2 = 1$, $n=100$ (top row), $n=400$ (middle row), $n=900$ (bottom row), and $\tau_1 = 0.1$ (left column), $\tau_1 = 0$ (right column). If the overdispersion variance proportion is zero then there is no vector of overdispersion random effects in the model, and the correct decision is to not select the vector of overdispersion random effects.

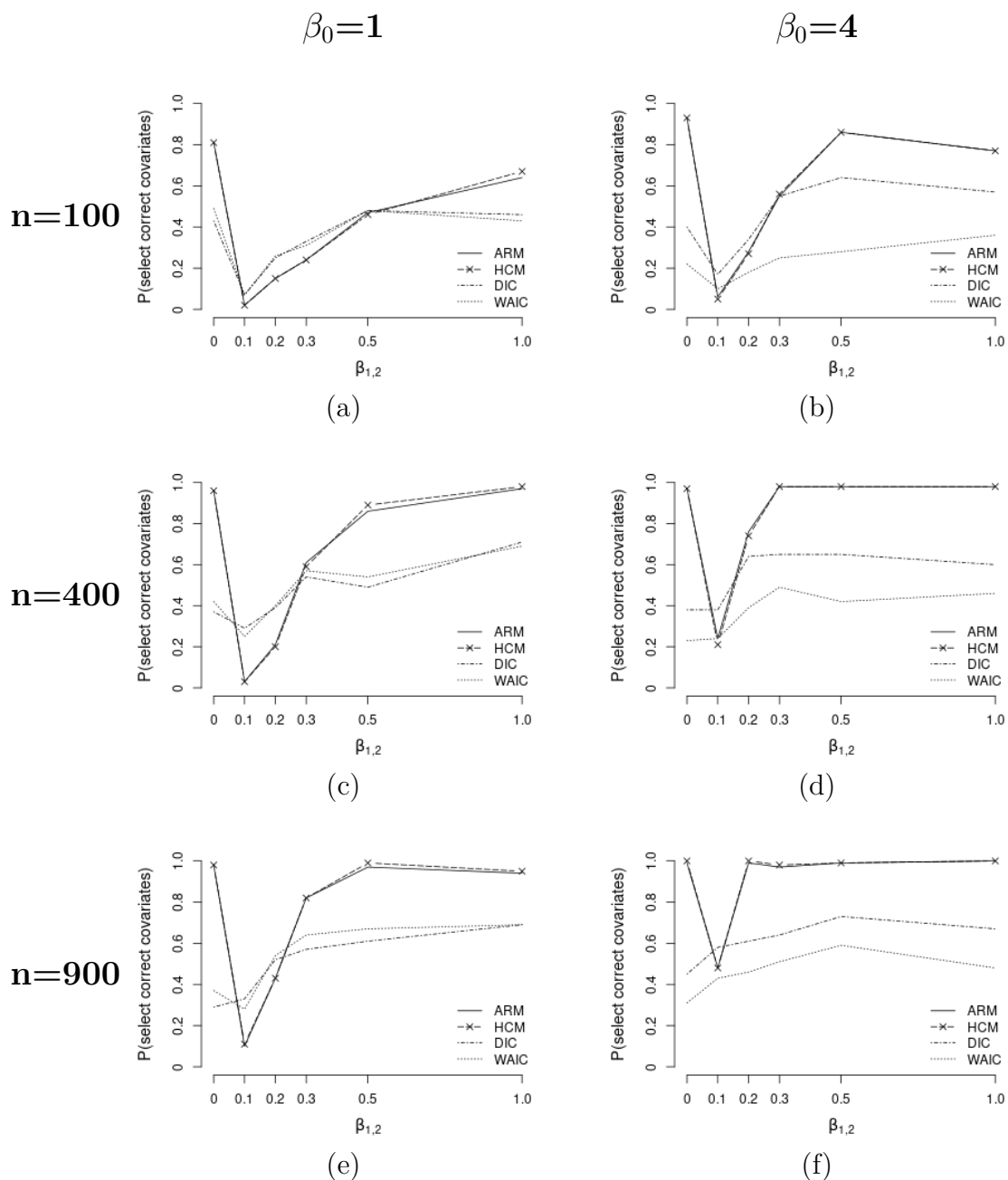


Figure 2.5: Probability of selecting the correct covariate structure as a function of the value of the regression coefficient. Settings: $\tau_1 = 0.05$, $\tau_2 = 0.05$, $n=100$ (top row), $n=400$ (middle row), $n=900$ (bottom row), and $\beta_0 = 1$ (left column), $\beta_0 = 4$ (right column). Covariates x_1 and x_3 are generated from $N(\mathbf{0}, I)$. Covariates x_2 and x_4 are generated from a Gaussian hierarchical model with mean zero and ICAR random effects that have the same covariance matrix $\tau_1 \Sigma_1$ as the vector of spatial random effects α_1 .

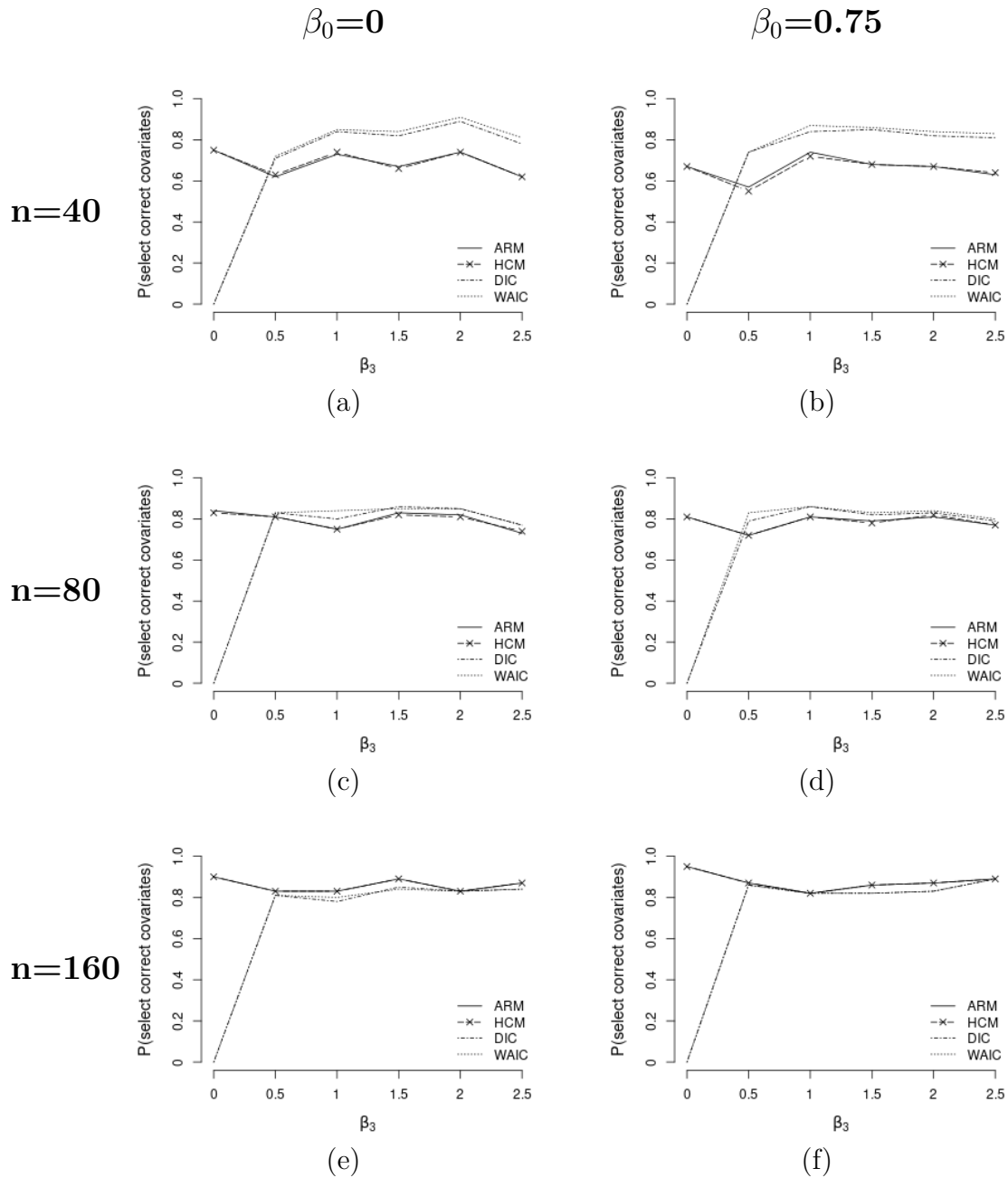
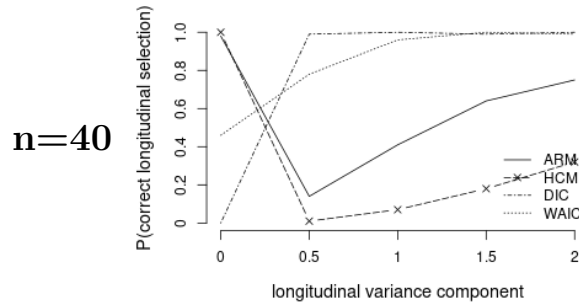
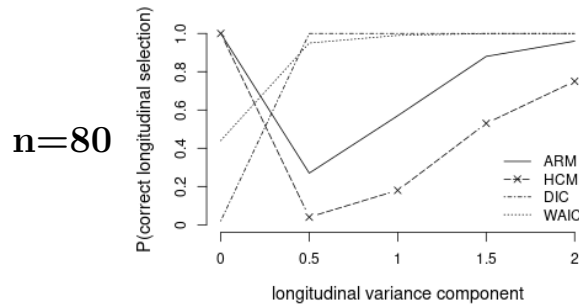


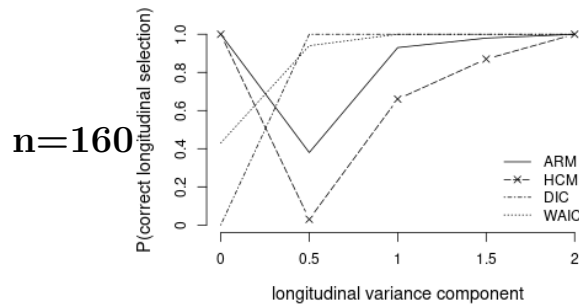
Figure 2.6: Bernoulli data. Probability of selecting correct covariate structure as a function of the value of the regression coefficient. Settings: $\tau_1 = 0.05$, $n=40$ (top row), $n=80$ (middle row), $n=160$ (bottom row), and $\beta_0 = 0$ (left column), $\beta_0 = 0.75$ (right column).



(a)



(b)



(c)

Figure 2.7: Bernoulli data. Probability of selecting correct longitudinal random effects structure as a function of the value of the variance component. Settings: $\beta_0 = 0$, $\beta_3 = 0.5$, $n=40$ (top row), $n=80$ (middle row), $n=160$ (bottom row).

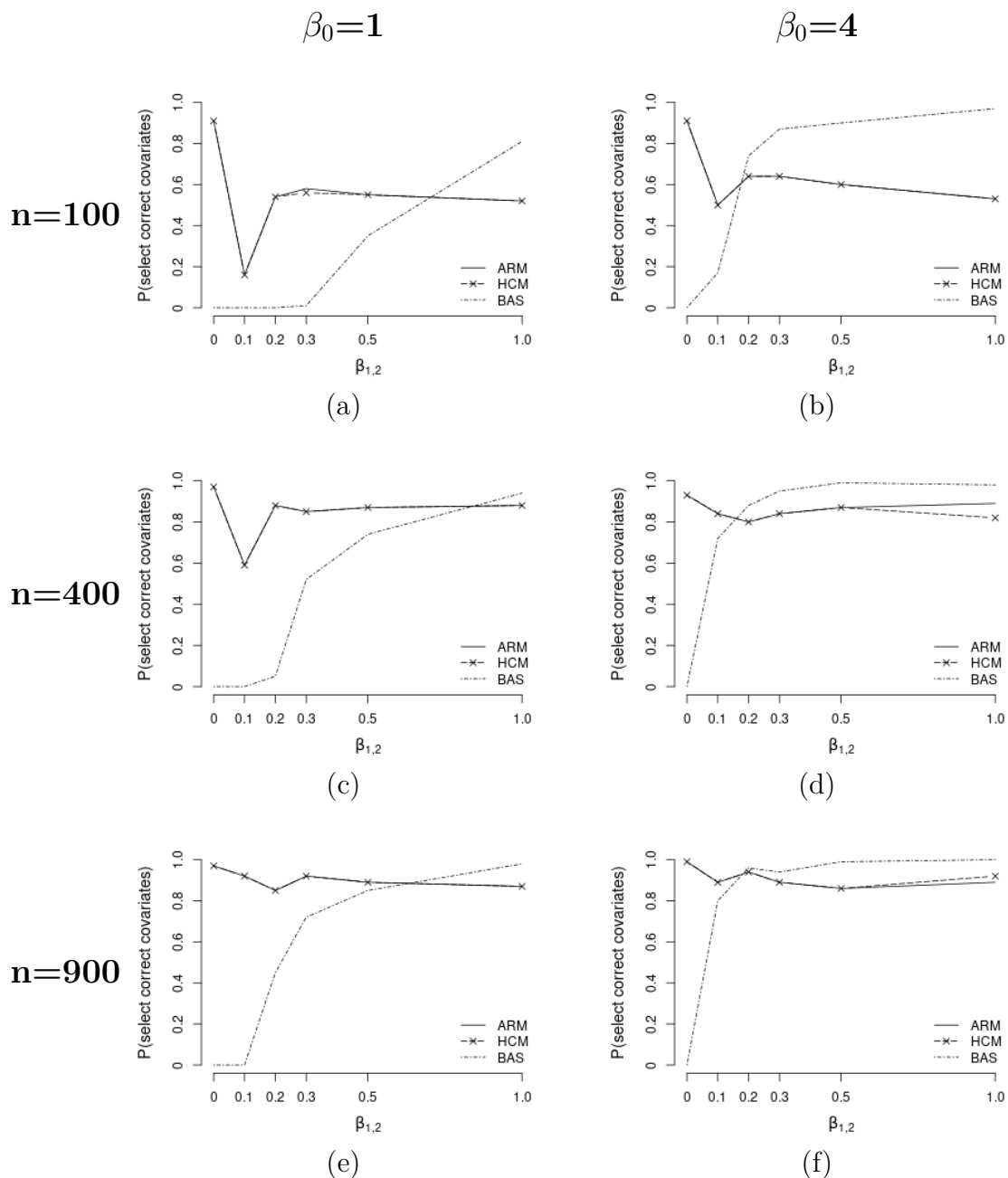


Figure 2.8: Poisson data. Probability of selecting the correct covariate structure as a function of the value of the regression coefficient by ARM, HCM and BAS package. Settings: $\tau_1 = 0$, $\tau_2 = 0$, $n=100$ (top row), $n=400$ (middle row), $n=900$ (bottom row), and $\beta_0 = 1$ (left column), $\beta_0 = 4$ (right column).

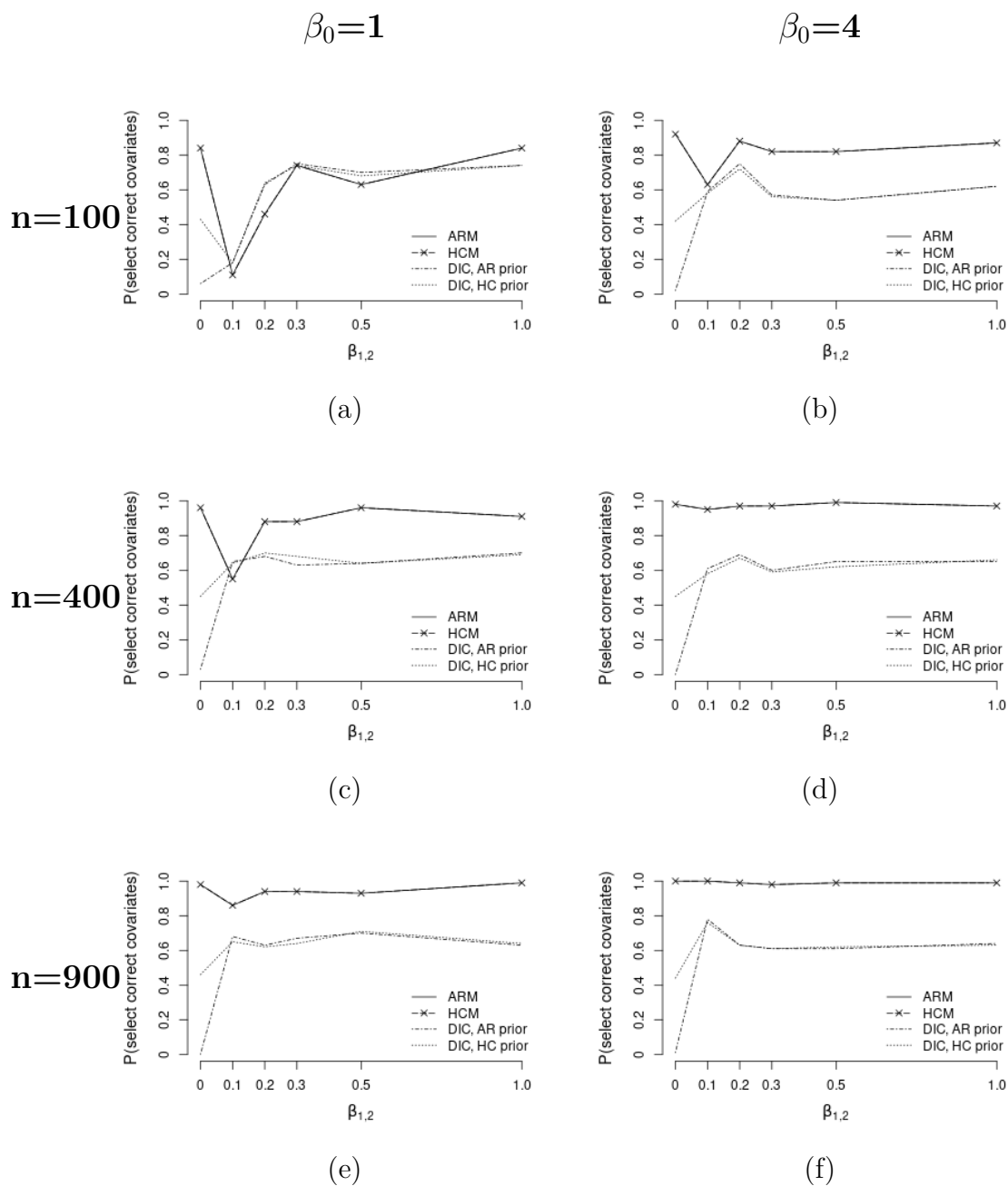


Figure 2.9: Poisson data. Comparison of our methods ARM and HCM with the DIC computed by INLA with our AR prior and HC prior. Probability of selecting correct covariate structure as a function of the value of the regression coefficient. Settings: $\tau_1 = 0.05$, $\tau_2 = 0.05$, $n=100$ (top row), $n=400$ (middle row), $n=900$ (bottom row), and $\beta_0 = 1$ (left column), $\beta_0 = 4$ (right column).

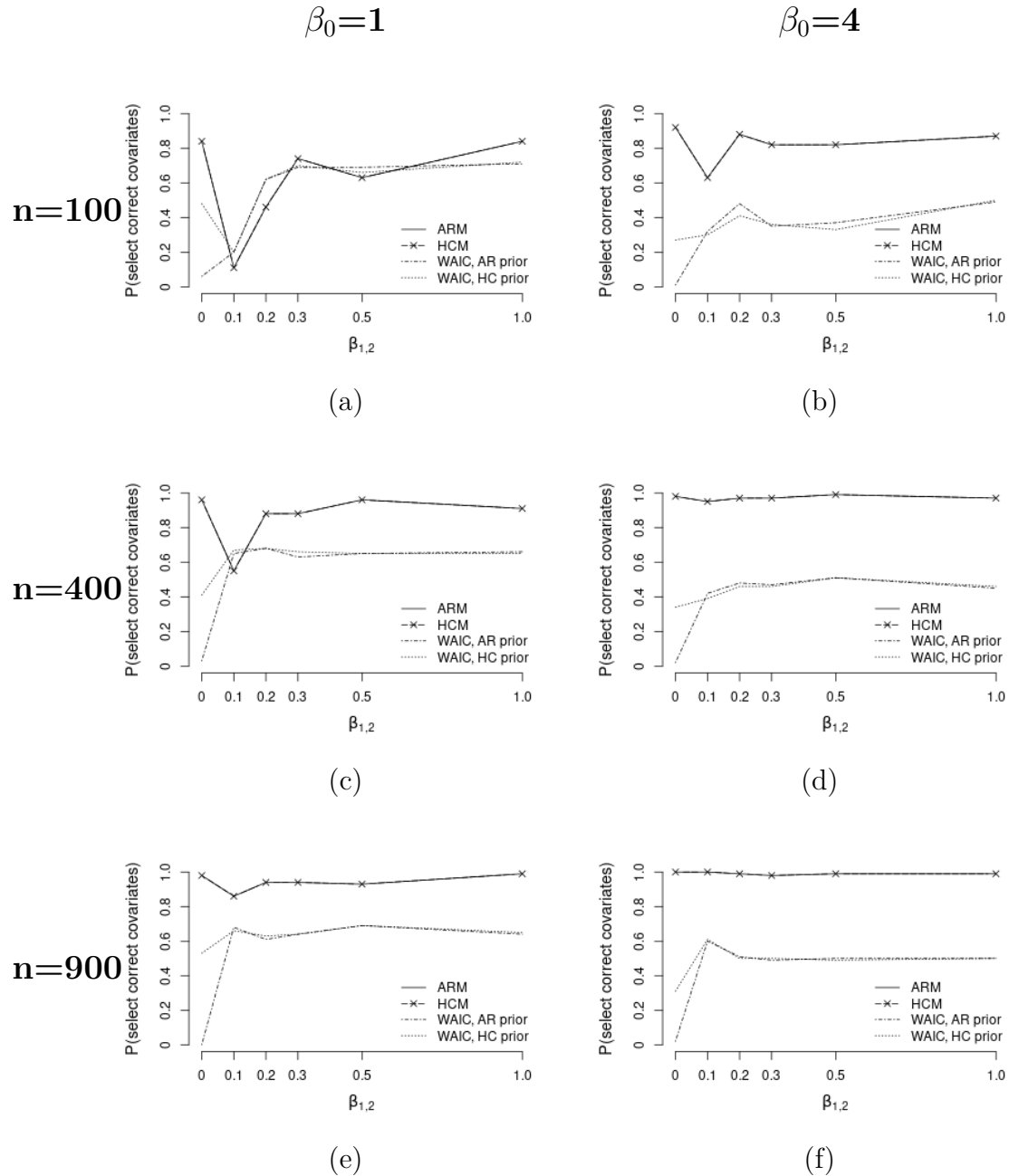


Figure 2.10: Poisson data. Comparison of our methods ARM and HCM with the WAIC computed by INLA with our AR prior and HC prior. Probability of selecting correct covariate structure as a function of the value of the regression coefficient. Settings: $\tau_1 = 0.05$, $\tau_2 = 0.05$, $n=100$ (top row), $n=400$ (middle row), $n=900$ (bottom row), and $\beta_0 = 1$ (left column), $\beta_0 = 4$ (right column).

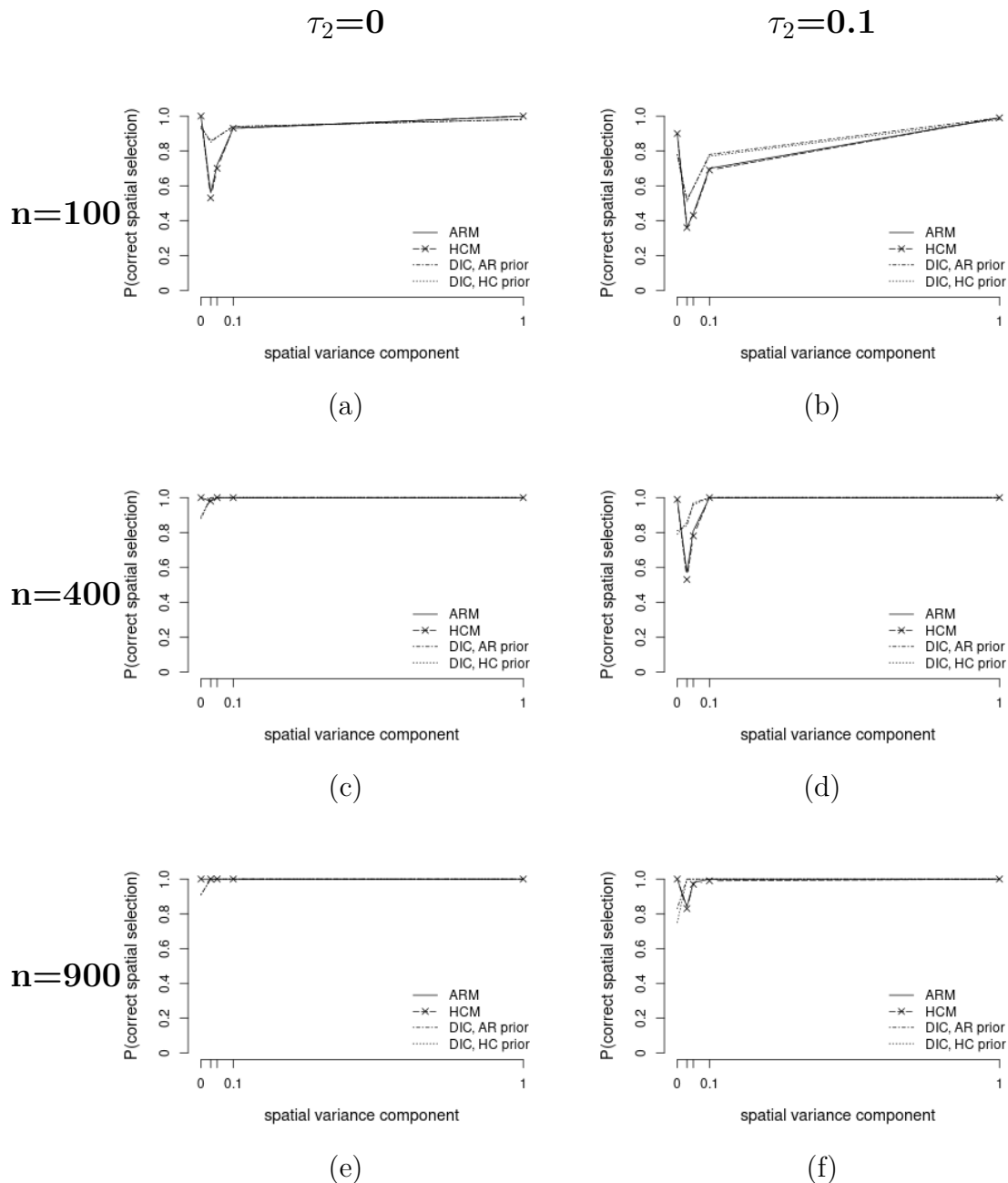


Figure 2.11: Poisson data. Comparison of our methods ARM and HCM with DIC computed by INLA with our AR prior and HC prior. Probability of selecting correct spatial random effects structure as a function of the value of the variance component. Settings: $\beta_0 = 2$, $\beta_{1,2} = 1$, $n=100$ (top row), $n=400$ (middle row), $n=900$ (bottom row), and $\tau_2 = 0$ (left column), $\tau_2 = 0.1$ (right column).

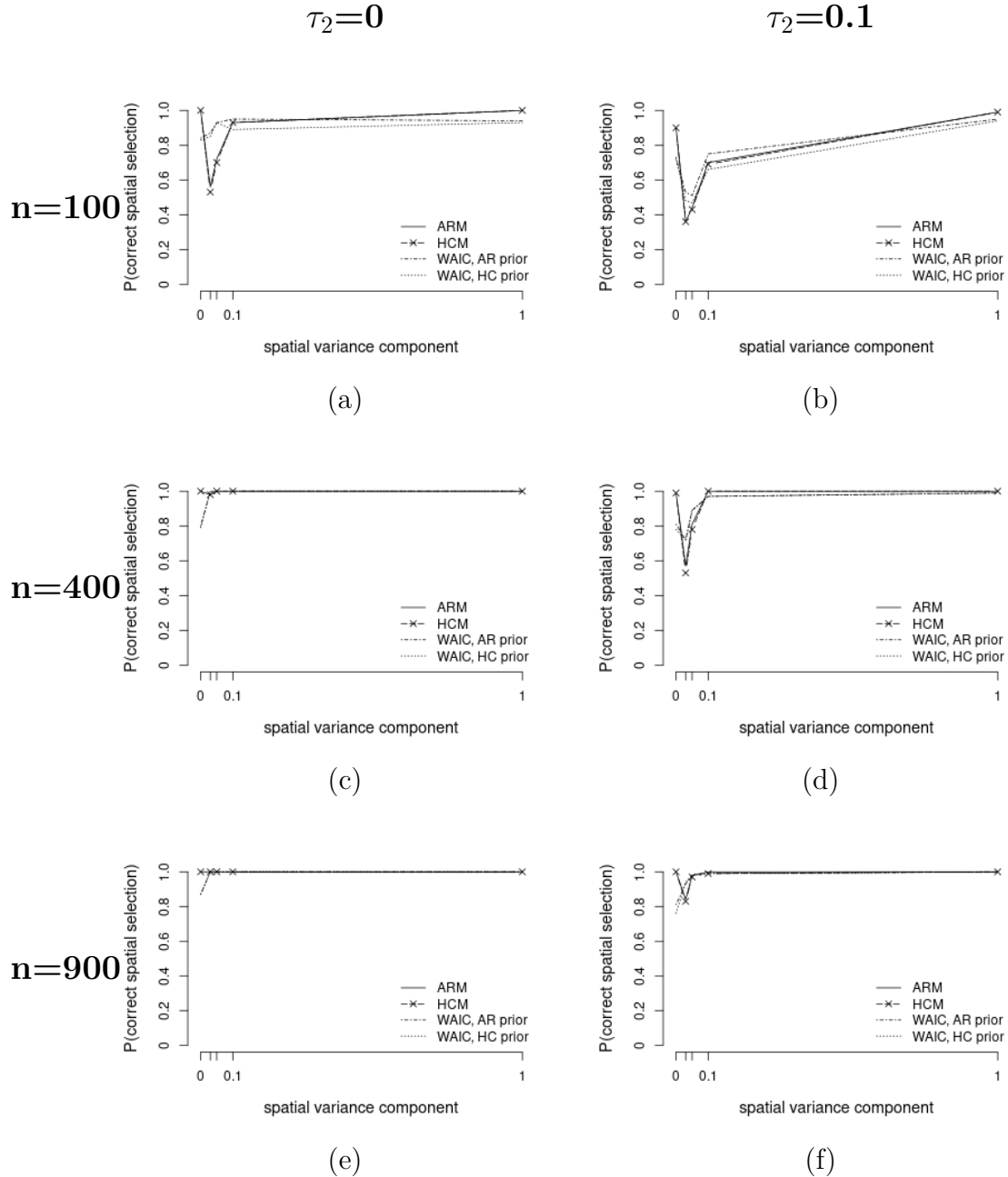


Figure 2.12: Poisson data. Comparison of our methods ARM and HCM with WAIC computed by INLA with our AR prior and HC prior. Probability of selecting correct spatial random effects structure as a function of the value of the variance component. Settings: $\beta_0 = 2$, $\beta_{1,2} = 1$, $n=100$ (top row), $n=400$ (middle row), $n=900$ (bottom row), and $\tau_2 = 0$ (left column), $\tau_2 = 0.1$ (right column).

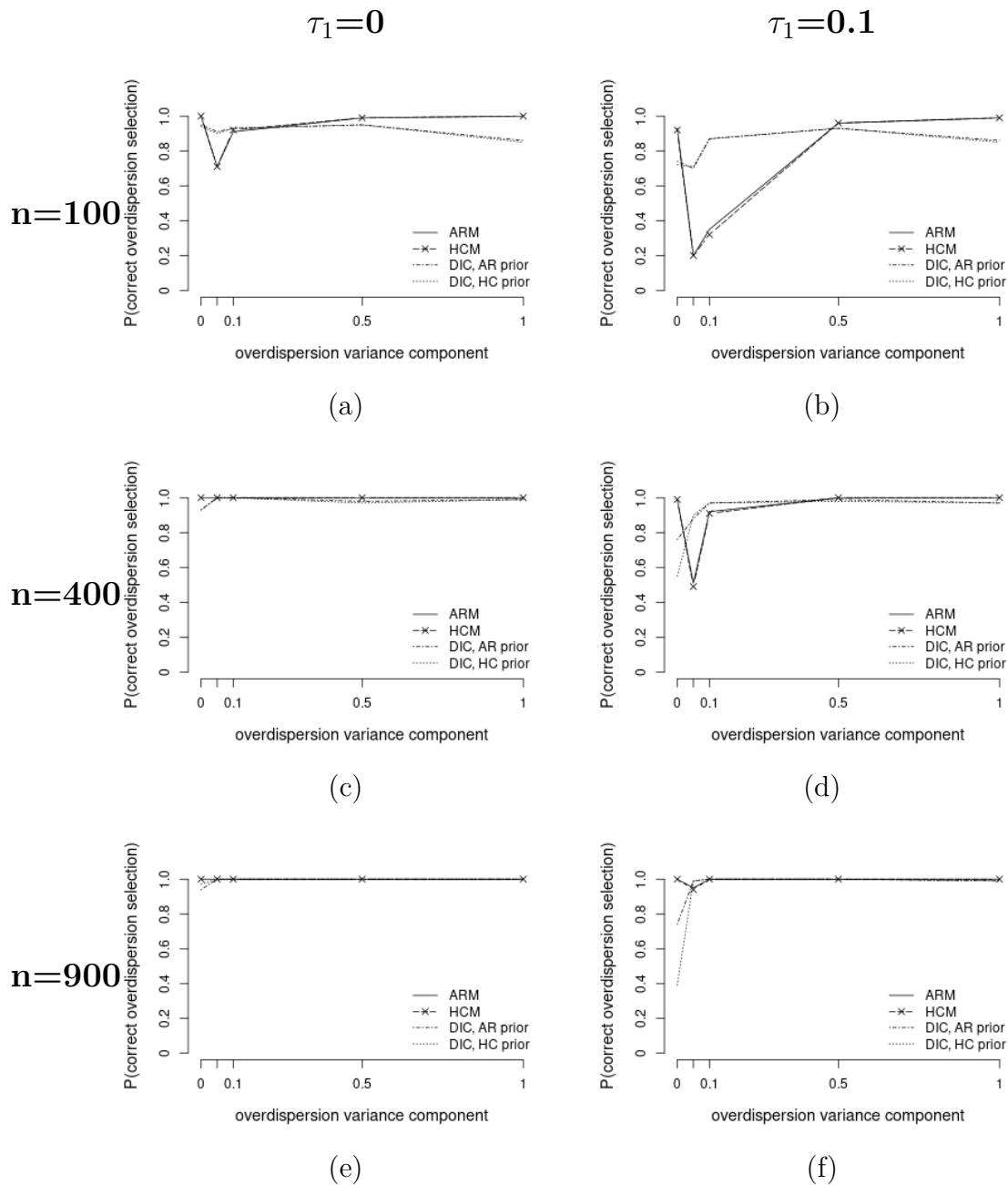


Figure 2.13: Poisson data. Comparison of our methods ARM and HCM with DIC computed by INLA with our AR prior and HC prior. Probability of selecting correct overdispersion random effects structure as a function of the value of the variance component. Settings: $\beta_0 = 2$, $\beta_{1,2} = 1$, $n=100$ (top row), $n=400$ (middle row), $n=900$ (bottom row), and $\tau_1 = 0$ (left column), $\tau_1 = 0.1$ (right column).

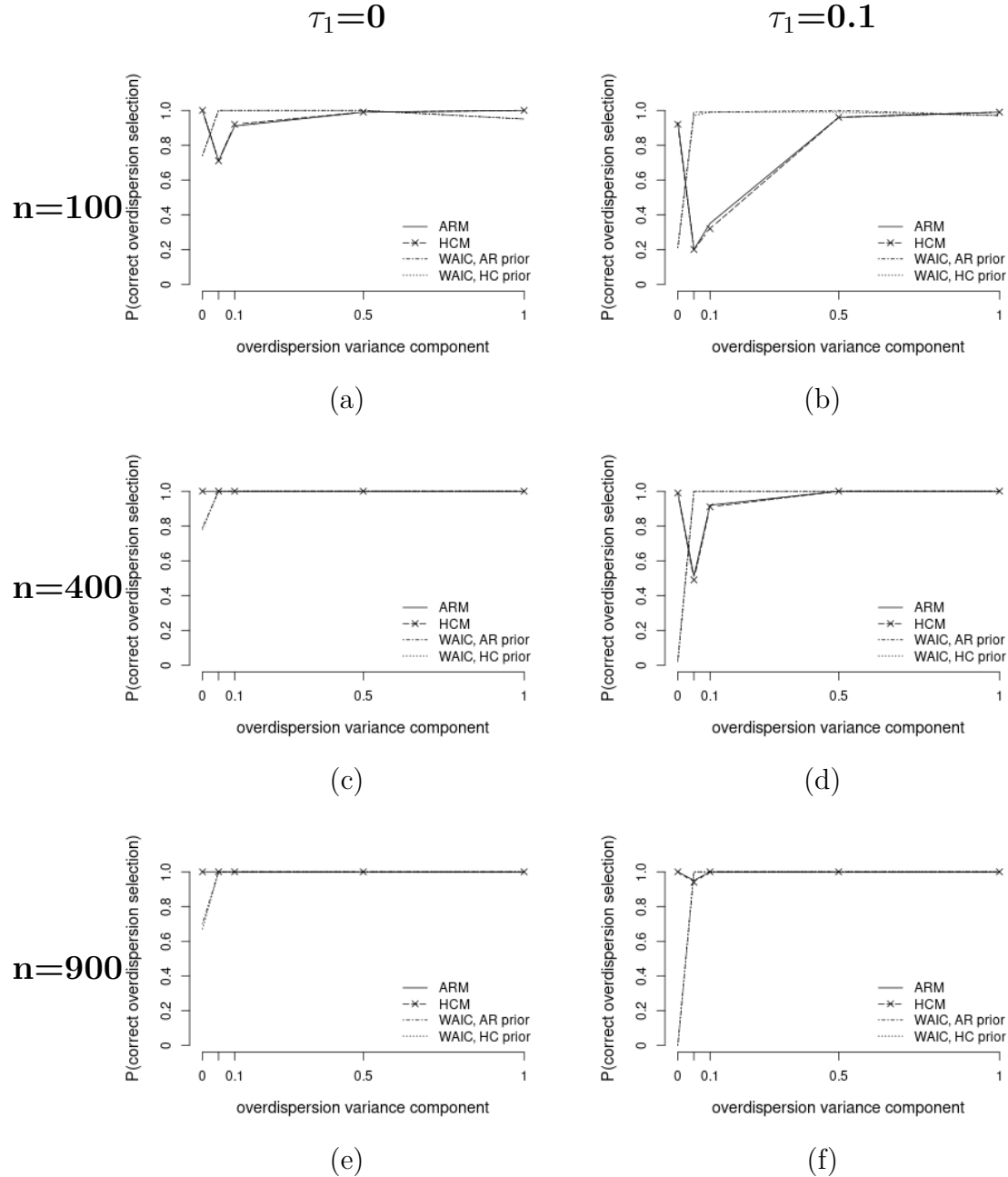


Figure 2.14: Poisson data. Comparison of our methods ARM and HCM with WAIC computed by INLA package with our AR prior and HC prior. Probability of selecting correct overdispersion random effects structure as a function of the value of the variance component. Settings: $\beta_0 = 2$, $\beta_{1,2} = 1$, $n=100$ (top row), $n=400$ (middle row), $n=900$ (bottom row), and $\tau_1 = 0$ (left column), $\tau_1 = 0.1$ (right column).

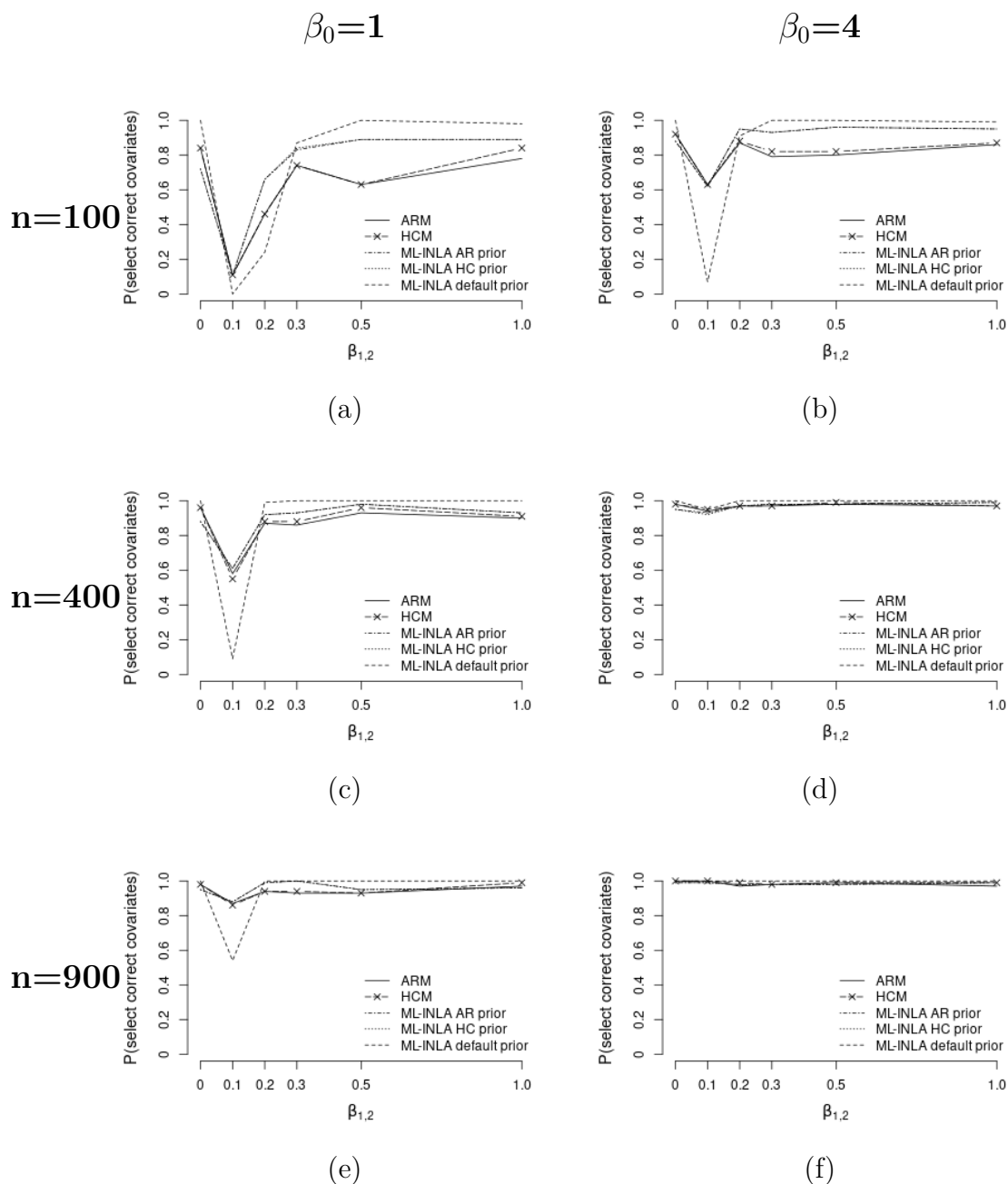


Figure 2.15: Poisson data. Comparison of our methods ARM and HCM with marginal likelihood (ML) computed by INLA. Probability of selecting the correct covariate structure as a function of the value of the regression coefficient. Settings: $\tau_1 = 0.05$, $\tau_2 = 0.05$, $n=100$ (top row), $n=400$ (middle row), $n=900$ (bottom row), and $\beta_0 = 1$ (left column), $\beta_0 = 4$ (right column).

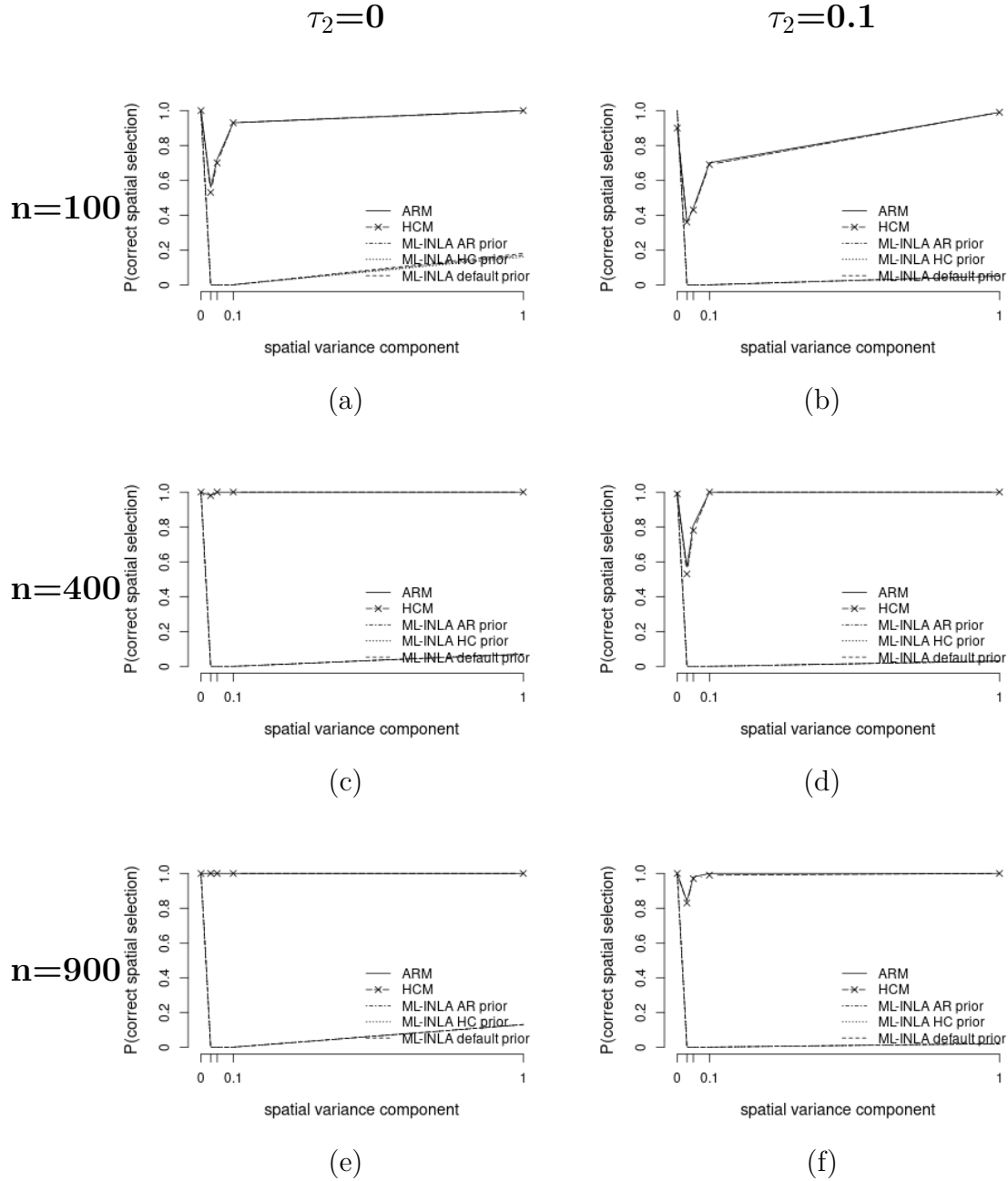


Figure 2.16: Poisson data. Comparison of our methods ARM and HCM with marginal likelihood (ML) computed by INLA. Probability of selecting correct spatial random effects structure as a function of the value of the variance component. Settings: $\beta_0 = 2$, $\beta_{1,2} = 1$, $n=100$ (top row), $n=400$ (middle row), $n=900$ (bottom row), and $\tau_2 = 0$ (left column), $\tau_2 = 0.1$ (right column).

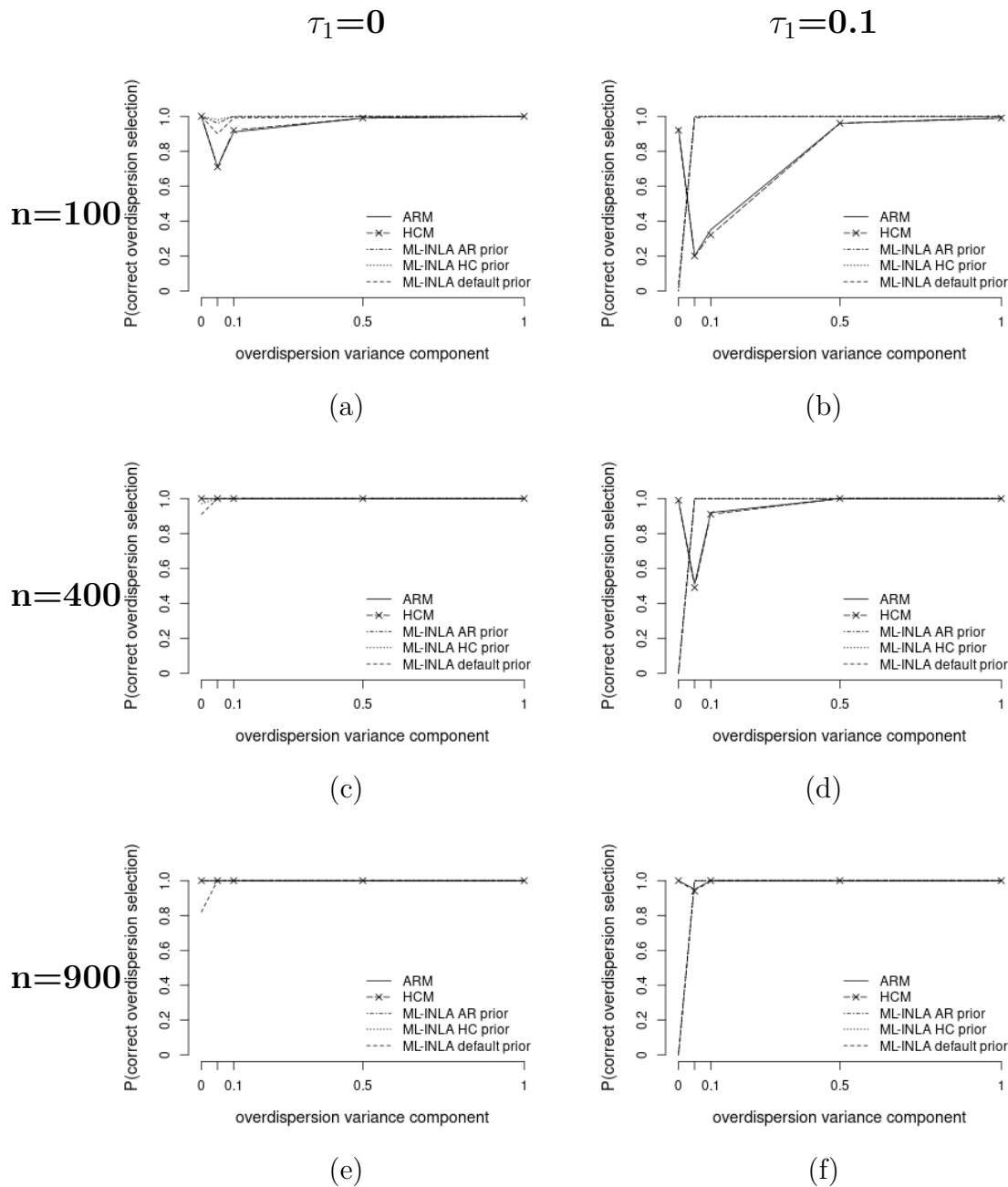


Figure 2.17: Poisson data. Comparison of our methods ARM and HCM with marginal likelihood (ML) computed by INLA. Probability of selecting correct overdispersion random effects structure as a function of the value of the variance component. Settings: $\beta_0 = 2$, $\beta_{1,2} = 1$, $n=100$ (top row), $n=400$ (middle row), $n=900$ (bottom row), and $\tau_2 = 0$ (left column), $\tau_2 = 0.1$ (right column).

Chapter 3

BG2: Bayesian variable selection in generalized linear mixed models with nonlocal priors for non-Gaussian GWAS data

3.1 Introduction

This chapter is based on the following manuscript that has been published in BMC Bioinformatics [78]: Shuangshuang Xu, Jacob Williams, and Marco A.R. Ferreira. BG2: Bayesian variable selection in generalized linear mixed models with nonlocal priors for non-Gaussian GWAS data. BMC Bioinformatics, 2023, 24(1): 343.

Genome-wide association studies (GWAS) have uncovered many single nucleotide polymorphisms (SNP) associated to important phenotypes such as plant productivity [85], plant response to salt stress [29], and human diseases [84]. To take into account the correlation among GWAS observations, the most widely used methods for the analysis of GWAS Gaussian data are single marker analysis (SMA) methods based on linear mixed models (LMMs) [30, 31, 81]. Recently, SMA based on logistic regression with random effects has

been proposed for the analysis of GWAS binary data [25]. However, to the best of our knowledge, there are no published methods for the analysis of other types of *correlated GWAS non-Gaussian* data such as count data. One of our contributions is to propose the use of generalized linear mixed models for the analysis of GWAS non-Gaussian data. To that end, we use generalized linear mixed models (GLMMs) and, thus, call our method Bayesian GLMMs for GWAS (BG2).

BG2 has two steps: a screening step and a model selection step. The screening step, similarly to SMA methods, fits p GLMMs where each model has just one SNP, and uses Bayesian FDR control [46, 47] to provide a set of candidate SNPs. After that, the model selection step performs a model search through the space of GLMMs that may include any number of screened candidate SNPs as possible regressors. BG2 implements both steps using a pseudo-likelihood approach. We note that a similar pseudo-likelihood approach can be used to implement SMA methods for non-Gaussian GWAS data, and a particular case of such an approach has been proposed for GWAS binary data [25]. However, simulation studies presented in Section 3.4 show that, when compared to such SMA methods for non-Gaussian data, BG2 leads to much lower FDR.

The GLMMs for GWAS data considered by BG2 may have two types of random effects: kinship random effects and overdispersion random effects. The kinship random effects account for correlation among GWAS observations due to population stratification and hidden relatedness. Similarly to existing literature for Gaussian GWAS data, we assume that the vector of kinship random effects follows a multivariate Gaussian distribution with a mean vector of zeros and a covariance matrix that is the product of a one-dimensional unknown variance parameter and a positive semi-definite kinship matrix [45, 51]. The overdispersion random effects allow for extra variability not accounted for by the model for observations; for example, when assuming a conditional Poisson model for the observations, the overdispersion

random effects account for extra-Poisson variability.

Both screening and model selection steps in BG2 are based on nonlocal priors. To the best of our knowledge, this is the first time that nonlocal priors are proposed for regression coefficients in GLMMs. Previous literature in Bayesian model selection for GLMMs has assigned for regression coefficients local priors [3]. While local priors have positive density at null parameter values, nonlocal priors have density equal to zero at null parameter values. Nonlocal priors were first proposed by [26, 27] for Gaussian linear models. Nonlocal priors have been successfully developed for many different problems such as model selection in Gaussian directed acyclic graphical models [1], classification with Bayesian probit models [55], variable selection in logistic models [48], Bayesian wavelet analysis [57], and variable selection in generalized linear models [73]. Nonlocal priors lead to faster accumulation of evidence in favor of a true null hypothesis [26], and have been shown to be advantageous for high-dimensional problems [27, 54, 55]. Therefore, BG2 uses nonlocal priors for SNP search in GWAS analysis.

Due to the large number of GLMMs that need to be fitted, BG2 relies on two approximations to speed up computations: a pseudo-likelihood approximation; and a Population Parameters Previously Determined (P3D) approximation. For GLMMs, the integrated likelihood function obtained by integrating out the random effects is not available in closed form. Repeated numerical integration of the random effects for each GLMM fitted for a GWAS analysis is computationally too expensive. Thus, BG2 uses a pseudo-likelihood approach [72] to facilitate integrating out the random effects. Such pseudo-likelihood approach leads to a Gaussian approximation for adjusted observations that allows analytically integrating out the random effects. To further speed up computations, we propose a P3D approximation for GLMMs. A P3D approximation was first proposed by [87] for Gaussian linear mixed models (LMMs) and variation of this approximation is used in the celebrated and widely

used method EMMAX for the analysis of GWAS Gaussian data [31].

In our P3D approach, for each BG2 step (screening and model selection) we fit a baseline GLMM to obtain adjusted observations and estimates of the variance parameters. We then keep the adjusted observations and the variance parameters fixed at the values computed with the baseline GLMM when fitting all other models in that BG2 step. In our P3D approach, the baseline model is different for the screening step and for the model selection step. For the screening step, the baseline model is a GLMM without any SNPs. For the model selection step, the baseline model is a GLMM with all candidate SNPs obtained from the screening step. This choice of baseline GLMM for the model selection step is based on [66], who have suggested for GLMMs the use of adjusted observations based on the full model – the model with all the regressors – when computing BICs for all possible models. When compared to a usual pseudo-likelihood approach to GLMMs, our P3D approximation greatly reduces the computational time and allows the analysis of non-Gaussian GWAS data within a reasonable time frame.

The remainder of this paper is organized as follows. Section 3.2 describes the GLMMs that we consider for non-Gaussian GWAS data. Section 3.3 describes our BG2 method for the identification of causal SNPs. Section 3.4 presents the results of two simulation studies for binary data and for count data. Section 3.5 illustrates our method with applications to three case studies: human cocaine dependence, alcohol consumption, and the number of root-like structures in the plant *A. Thaliana*. Section 3.6 concludes with a discussion and future directions.

3.2 GLMMs for GWAS

Consider observations y_1, \dots, y_n that, given random effects, are conditionally independent and have a distribution from the exponential family of distributions. This flexible family of distributions includes the Bernoulli, binomial, Poisson, and gamma distributions. Thus, this family may be used to model observed GWAS phenotypes such as an indicator of disease presence/absence, number of lateral roots in plants, or survival time. Then, the density function of y_i is

$$f(y_i|\eta_i) = \exp[y_i\eta_i - B(\eta_i) + C(y_i)], \quad (3.1)$$

for $i = 1, \dots, n$, where $B(\cdot)$ and $C(\cdot)$ are known functions. Further, each observation y_i has mean $\mu_i = B'(\eta_i)$ and variance $v_i = B''(\eta_i)$. Let X_s be a matrix of SNPs and $\boldsymbol{\beta}_s$ be the corresponding vector of regression coefficients. In addition, let X_c be a matrix that contains a column of ones for the intercept and other columns for control covariates (e.g., age, sex, and environmental factors) and $\boldsymbol{\beta}_c$ be the corresponding vector of regression coefficients. Further, let $\boldsymbol{\alpha}_1$ be a vector of random effects that accounts for kinship correlation. Specifically, $\boldsymbol{\alpha}_1$ has a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\kappa_1 \Sigma$, where κ_1 is an unknown scalar and Σ is a kinship matrix. Furthermore, let $\boldsymbol{\alpha}_2$ be a vector of overdispersion random effects following $N(\mathbf{0}, \kappa_2 I)$. Let $\mathbf{y} = (y_1, \dots, y_n)$ be the vector of observed phenotypes. Then, the conditional expectation $E(\mathbf{y}|\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ is linked to the linear predictor $X_s\boldsymbol{\beta}_s + X_c\boldsymbol{\beta}_c + \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2$ by the link function g :

$$g(E(\mathbf{y}|\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)) = X_s\boldsymbol{\beta}_s + X_c\boldsymbol{\beta}_c + \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2. \quad (3.2)$$

The class of GLMMs given by Equations (3.1) and (3.2) can be expanded to deal with other cases. For example, to account for the experimental design used for data collection, we may add another random effect $\boldsymbol{\alpha}_3$ following a multivariate normal distribution with

mean vector $\mathbf{0}$ and covariance matrix $\kappa_3 \Sigma_3$, where κ_3 is a unknown parameter and Σ_3 is a symmetric positive semi-definite matrix that describes the dependence structure among the observations due to the experimental design.

3.3 BG2: Bayesian SNP selection in GLMMs for GWAS

Our method BG2 consists of two steps: screening and model selection. The BG2 screening step uses a novel Bayesian single marker analysis for non-Gaussian data and applies Bayesian false discovery rate control to yield a set of candidate SNPs. After that, the BG2 model selection step performs a search through the model space of all GLMMs that may include any number of SNPs from the set of candidate SNPs. In both steps, BG2 uses a pseudo-likelihood approach to fit models. In what follows, Section 3.3.1 presents the pseudo-likelihood approach, Section 3.3.2 introduces the BG2 screening step, and Section 3.3.3 presents the BG2 model selection step.

3.3.1 Pseudo-likelihood model fitting

In both the screening and the model selection steps, BG2 uses a pseudo-likelihood approach. This is an iterative approach that writes the model for the observations as $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is a vector of errors and $V = Var(\boldsymbol{\epsilon}) = Var(\mathbf{y})$. In addition, the pseudo-likelihood approach expands $\boldsymbol{\mu} = E(\mathbf{y}|\boldsymbol{\beta}_s, \boldsymbol{\beta}_c, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ in a first-order Taylor expansion around current estimates of $\boldsymbol{\beta}_s$, $\boldsymbol{\beta}_c$, $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, κ_1 , and κ_2 . The resulting equation is rearranged such that the left-hand side depends only on known quantities (observations, current estimates of parameters, regression matrices). Then, this equation is pre-multiplied by V^{-1} . The left-hand side of the resulting equation, known as the vector of adjusted observations, is $\mathbf{y}^* =$

$\widehat{V}^{-1}(\mathbf{y} - \widehat{\boldsymbol{\mu}}) + X_s \widehat{\boldsymbol{\beta}}_s + X_c \widehat{\boldsymbol{\beta}}_c + \widehat{\boldsymbol{\alpha}}_1 + \widehat{\boldsymbol{\alpha}}_2$. Equating \mathbf{y}^* to the right-hand side of the resulting equation yields

$$\mathbf{y}^* = X_s \boldsymbol{\beta}_s + X_c \boldsymbol{\beta}_c + \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 + \widehat{V}^{-1} \boldsymbol{\epsilon}. \quad (3.3)$$

Then, the pseudo-likelihood approach approximates the GLMM by an LMM given by Equation (3.3) with vectors of random effects $\boldsymbol{\alpha}_1 \sim N(\mathbf{0}, \kappa_1 \Sigma)$ and $\boldsymbol{\alpha}_2 \sim N(\mathbf{0}, \kappa_2 I)$. Based on this LMM, new estimates are computed for $\boldsymbol{\beta}_s$, $\boldsymbol{\beta}_c$, $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, κ_1 , κ_2 , and V . The pseudo-likelihood algorithm then iterates until convergence of these estimates.

3.3.2 BG2 screening step

The BG2 screening step uses a P3D approach based on a baseline model that assumes a linear predictor given in Equation (3.2) specialized to contain no SNPs, that is, $g(E(\mathbf{y} | \boldsymbol{\beta}_c, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)) = X_c \boldsymbol{\beta}_c + \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2$.

Our P3D approach keeps $\boldsymbol{\beta}_c$, κ_1 , κ_2 , and V fixed at their pseudo-likelihood estimates when performing the Bayesian SMA in the BG2 screening step. Let us denote these estimates by $\widehat{\boldsymbol{\beta}}_c$, $\widehat{\kappa}_1$, $\widehat{\kappa}_2$, and \widehat{V} . In addition, our P3D approach keeps the vector of adjusted observations fixed equal to \mathbf{y}^* obtained at the last iteration of the pseudo-likelihood algorithm for the baseline model. Let \mathbf{x}_s be the vector of covariate values for SNP s . Then, the BG2 screening step assumes for each SNP s , $s = 1, \dots, p$, that the adjusted observations \mathbf{y}^* can be modeled by the LMM

$$\mathbf{y}^* = X_c \widehat{\boldsymbol{\beta}}_c + \mathbf{x}_s \beta_s + \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 + \widehat{V}^{-1} \boldsymbol{\epsilon}. \quad (3.4)$$

Let $\widehat{H} = \widehat{\kappa}_1 \Sigma + \widehat{\kappa}_2 I + \widehat{V}^{-1}$ be the covariance matrix of the adjusted observations \mathbf{y}^* . Then, the adjusted observations \mathbf{y}^* have an approximate multivariate Gaussian distribution $N(X_c \widehat{\boldsymbol{\beta}}_c +$

$\mathbf{x}_s \beta_s, \widehat{H}$). Consider the spectral decomposition $\widehat{H} = PDP^T$. Let $\tilde{\mathbf{y}} = P^T(\mathbf{y}^* - X_c \widehat{\beta}_c)$ and $\tilde{\mathbf{x}}_s = P^T \mathbf{x}_s$. Then, an estimator of β_s is $\widehat{\beta}_s = (\tilde{\mathbf{x}}_s^T D^{-1} \tilde{\mathbf{x}}_s)^{-1} \tilde{\mathbf{x}}_s^T D^{-1} \tilde{\mathbf{y}}$. In addition, the estimator $\widehat{\beta}_s$ has approximate distribution $N(\beta_s, \sigma_s^2)$, where $\sigma_s^2 = \text{var}(\widehat{\beta}_s) = (\tilde{\mathbf{x}}_s^T D^{-1} \tilde{\mathbf{x}}_s)^{-1}$.

We assign for β_s a prior that is a mixture of a Dirac delta function and a nonlocal prior, that is,

$$p(\beta_s | \tau, \pi_0) = \pi_0 \delta(\beta_s = 0) + (1 - \pi_0) \frac{\beta_s^2}{n\tau\sigma_s^2} N(\beta_s | 0, n\tau\sigma_s^2),$$

where π_0 is the probability of the null hypothesis that β_s is equal to zero and $\tau > 0$ is a scale parameter. Larger values of τ cause the prior to be more spread out and lead BG2 to focus on identifying SNPs with relatively large regression coefficients. Then, the predictive density of $\widehat{\beta}_s$ is

$$\begin{aligned} p(\widehat{\beta}_s | \tau, \pi_0) &= \int p(\widehat{\beta}_s | \beta_s) p(\beta_s | \tau, \pi_0) d\beta_s \\ &= \pi_0 N(\widehat{\beta}_s | 0, \sigma_s^2) + (1 - \pi_0) (2\pi\sigma_s^2)^{-\frac{1}{2}} (n\tau + 1)^{-\frac{3}{2}} \\ &\quad \exp \left\{ -\frac{\widehat{\beta}_s^2}{2\sigma_s^2(n\tau + 1)} \right\} \left[1 + \frac{n\tau\widehat{\beta}_s^2}{(n\tau + 1)\sigma_s^2} \right]. \end{aligned} \quad (3.5)$$

Based on this predictive density and assuming that $\widehat{\beta}_1, \dots, \widehat{\beta}_p$ are approximately conditionally independent given π_0 and τ , we obtain the approximate likelihood function of τ and π_0

$$L(\widehat{\beta}_1, \dots, \widehat{\beta}_p | \tau, \pi_0) = \prod_{s=1}^p p(\widehat{\beta}_s | \tau, \pi_0). \quad (3.6)$$

Let $\pi(\tau)$ and $\pi(\pi_0)$ be the prior densities of τ and π_0 , respectively. Then, by Bayes Theorem

an approximate posterior density for (τ, π_0) is

$$\pi(\tau, \pi_0 | \hat{\beta}_1, \dots, \hat{\beta}_p) \propto \pi(\tau)\pi(\pi_0) \prod_{s=1}^p p(\hat{\beta}_s | \tau, \pi_0). \quad (3.7)$$

BG2 estimates τ and π_0 by maximizing (3.7) to obtain posterior modes $\hat{\tau}$ and $\hat{\pi}_0$.

We assign a noninformative uniform prior on $(0, 1)$ for π_0 and consider two prior specifications for τ . The first prior specification is a uniform prior for τ on $(0, \infty)$. The second prior specification for τ is an inverse gamma distribution with shape parameter $0.55/0.022 + 1$ and rate parameter 0.55 , that is $\tau \sim IG(0.55/0.022 + 1, 0.55)$. This prior specification implies a prior mean for τ equal to 0.022 , which was the value for a fixed τ recommended by [58] for GWAS studies. In addition, we note that values of τ that are too small lead to numerical instability. Therefore, our prior specification implies that *a priori* $P(\tau > 0.01) = 0.999$, stochastically keeping τ away from 0.

Alternatively, we may fix τ at pre-specified values [27, 58]. Specifically, in the context of GWAS analysis, [58] suggested fixing $\tau = 0.022$ because GWAS effect sizes are generally very small. When $\tau = 0.022$, the nonlocal MOM prior assigns a probability of 0.01 to the event that a standardized effect size falls in the interval $(-0.05, 0.05)$. Thus, in the simulation studies presented in Section 3.4, we also consider fixing τ at 0.022.

After estimating τ and π_0 , BG2 takes an Empirical Bayes approach and keep them at their estimates $\hat{\tau}$ and $\hat{\pi}_0$ while using Bayes Theorem to compute the posterior probability that the regression coefficient of SNP s ($s = 1, \dots, p$) in the screening step is different than zero, that is

$$P(\beta_s \neq 0 | \hat{\beta}_s, \hat{\tau}, \hat{\pi}_0) = 1 - \frac{\pi_0 N(\hat{\beta}_s | 0, \sigma_s^2)}{p(\hat{\beta}_s | \hat{\tau}, \hat{\pi}_0)}, \quad (3.8)$$

where $p(\widehat{\beta}_s|\widehat{\tau}, \widehat{\pi}_0)$ is the predictive density given in Equation (3.5) with $\tau = \widehat{\tau}$ and $\pi_0 = \widehat{\pi}_0$.

Finally, based on the posterior probabilities computed with Equation (3.8), the BG2 screening step uses Bayesian FDR control [16, 46, 47, 71, 74] to obtain a list of candidate SNPs while keeping the nominal FDR at 5%. Let k be the number of candidate SNPs.

3.3.3 BG2 model selection step

The BG2 model selection step considers GLMMs with any number of SNPs from the list of k candidate SNPs obtained from the BG2 screening step. Thus, the model selection step considers $S = 2^k$ possible models. Let M_m be the m -th model, $m = 1, \dots, S$. Let X_m be the matrix of SNPs in model M_m , β_m be the corresponding vector of regression coefficients, and p_m be the number of SNPs in model M_m . Let X_S be the model with all k candidate SNPs.

We assume that the k candidate SNPs may or may not be in a model according to a sequence of exchangeable Bernoulli trials. Specifically, the prior probability of model M_m is $P(M_m) = \widehat{\pi}_0^{k-p_m}(1 - \widehat{\pi}_0)^{p_m}$ where $\widehat{\pi}_0$ is the estimate of the probability of null hypothesis obtained in the screening step. We do this to ensure that the Bayesian control of false discoveries in the BG2 model selection step is as strict as the control of false discoveries in the BG2 screening step.

The BG2 model selection step uses a P3D approach where the baseline model is the full model M_S with linear predictor $g(E(\mathbf{y}|\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)) = X_c\beta_c + X_S\beta_S + \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2$. The pseudo-likelihood approach then yields estimates $\widehat{\beta}_c$, $\widehat{\kappa}_1$ and $\widehat{\kappa}_2$, \widehat{V} , and adjusted observations \mathbf{y}^* . We then consider all models $M_m, m = 1, \dots, S$, where we keep β_c , κ_1 , κ_2 , and V fixed at these estimates. In addition, following the recommendation of [66], we keep the adjusted observations for all the S considered models fixed at the adjusted observations \mathbf{y}^* obtained while fitting the full model.

Therefore, under model M_m and with the P3D approach, the adjusted observations \mathbf{y}^* follow the approximate distribution $N\left(X_c\widehat{\boldsymbol{\beta}}_c + X_m\boldsymbol{\beta}_m, H\right)$, where $H = \widehat{\kappa}_1\Sigma + \widehat{\kappa}_2I + \widehat{V}^{-1}$. Consider the spectral decomposition of the matrix H given by $H = PDP^T$. In addition, let $\widetilde{\mathbf{y}} = P^T(\mathbf{y}^* - X_c\widehat{\boldsymbol{\beta}}_c)$ and $\widetilde{X}_m = P^T X_m$. Then, we can rewrite the LMM as $\widetilde{\mathbf{y}}|\boldsymbol{\beta}_m \stackrel{a}{\sim} N(\widetilde{X}_m\boldsymbol{\beta}_m, D)$. Because D is a diagonal matrix, computations for this latter model are very fast.

We propose a novel nonlocal prior for GLMMs. Specifically, we propose a prior density that is the product of a multivariate Gaussian density and the product of the square of each element of the vector of regression coefficients $\boldsymbol{\beta}_m$. In this multivariate Gaussian density, the covariance matrix is $\tau n(X_m^T H^{-1} X_m)^{-1}$. Using the spectral decomposition of the matrix H , the prior we propose for $\boldsymbol{\beta}_m$ is

$$\begin{aligned} p(\boldsymbol{\beta}_m|M_m) &= d_m(2\pi)^{-p_m/2}(\widehat{\tau}n)^{-3p_m/2}|\widetilde{X}_m^T D^{-1} \widetilde{X}_m|^{\frac{3}{2}} \\ &\quad \exp\left[-\frac{1}{2\widehat{\tau}n}\boldsymbol{\beta}_m^T \widetilde{X}_m^T D^{-1} \widetilde{X}_m \boldsymbol{\beta}_m\right] \prod_{i=1}^{p_m} \beta_{mi}^2, \end{aligned} \quad (3.9)$$

where d_m is a normalizing constant.

Let $C_m = \widetilde{X}_m^T D^{-1} \widetilde{X}_m(1+(\widehat{\tau}n)^{-1})$, $\widetilde{\boldsymbol{\beta}}_m = C_m^{-1} \widetilde{X}_m^T D^{-1} \widetilde{\mathbf{y}}$, and $R_m = \widetilde{\mathbf{y}}^T D^{-1}(D - \widetilde{X}_m C_m^{-1} \widetilde{X}_m^T) D^{-1} \widetilde{\mathbf{y}} = \widetilde{\mathbf{y}}^T D^{-1} \widetilde{\mathbf{y}} - \widetilde{\mathbf{y}}^T D^{-1} \widetilde{X}_m \widetilde{\boldsymbol{\beta}}_m$. Then, the marginal density of the adjusted observations $\widetilde{\mathbf{y}}$ conditional on model M_m is

$$\begin{aligned} m(\widetilde{\mathbf{y}}|M_m) &= \int N(\widetilde{\mathbf{y}}|\widetilde{X}_m\boldsymbol{\beta}_m, D)\pi(\boldsymbol{\beta}_m|M_m) d\boldsymbol{\beta}_m \\ &= (2\pi)^{-\frac{n}{2}}|D|^{-\frac{1}{2}}(1+\widehat{\tau}n)^{-p_m/2} \\ &\quad \exp\left(-\frac{R_m}{2}\right) \frac{E_2\left(\prod_{i=1}^{p_m} \beta_{mi}^2\right)}{E_1\left(\prod_{i=1}^{p_m} \beta_{mi}^2\right)}, \end{aligned} \quad (3.10)$$

where $E_1\left(\prod_{i=1}^{p_m} \beta_{mi}^2\right)$ is the expected value with respect to $N(\mathbf{0}, (1+\widehat{\tau}n)C_m^{-1})$ and $E_2\left(\prod_{i=1}^{p_m} \beta_{mi}^2\right)$ is the expected value with respect to $N(\widetilde{\boldsymbol{\beta}}_m, C_m^{-1})$. To approximate $E_1\left(\prod_{i=1}^{p_m} \beta_{mi}^2\right)$ and

$E_2(\prod_{i=1}^{p_m} \beta_{mi}^2)$, we simulate 2000 samples from $N(\tilde{\boldsymbol{\beta}}_m, C_m^{-1})$, denoted as $\boldsymbol{\beta}_{2m}^{(j)}$, $j = 1, \dots, 2000$. We compute $\sum_{j=1}^{2000} (\prod_{i=1}^{p_m} \beta_{2mi}^{2(j)})/2000$ as an approximation to $E_2(\prod_{i=1}^{p_m} \beta_{mi}^2)$. Let $\boldsymbol{\beta}_{1m}^{(j)} = (1 + \hat{\tau}n)^{\frac{1}{2}}(\boldsymbol{\beta}_{2m}^{(j)} - \tilde{\boldsymbol{\beta}}_m)$, $j = 1, \dots, 2000$. Finally, we compute $\sum_{j=1}^{2000} (\prod_{i=1}^{p_m} \beta_{1mi}^{2(j)})/2000$ as an approximation to $E_1(\prod_{i=1}^{p_m} \beta_{mi}^2)$.

Then, the posterior probability of model M_m is

$$P(M_m|\tilde{\mathbf{y}}) \propto P(M_m)m(\tilde{\mathbf{y}}|M_m). \quad (3.11)$$

If the number of candidate covariates k is small ($k < 16$), we compute the posterior probabilities for all 2^k candidate models and select the highest posterior probability model as the best model. If the number of candidate covariates is large, we use a genetic algorithm from the R package GA [62] to search for the highest posterior probability model.

3.4 Simulation Studies

We have performed simulation studies to compare our BG2 method with SMA for binary data and count data. Specifically, we consider single marker analysis with Bonferroni correction with nominal FDR set to 0.05. To assess the performance of our methods, in these simulation studies we use genotype SNP data from humans and from *A. Thaliana*. These are the same genotype data used in the case studies we present in Section 3.5. We use four criteria to compare the competing methods: true positives (TP), false positives (FP), false discovery rate (FDR) and F1 score. Within each simulation study, for each method we compute the average TP, FP, FDR and F1 over 100 simulated datasets. We use a buffer to define what is a true positive and a false positive. Specifically, if one or more detected SNPs are adjacent (within 5000 base pairs) to a same causal SNP, that is counted as a true positive. In addition,

each detected SNP not adjacent to a causal SNP is counted as a false positive.

3.4.1 Binary data

We simulate binary GWAS data using genotype information from the Study of Addiction: Genetics and Environment (SAGE) which is part of the National Human Genome Research Institute’s Gene Environment Association Study Initiative [Database for Genotypes and Phenotypes (dbGaP) study accession phs000092.v1.p1]. Specifically, we use genotype information from 2,772 European Americans in a total of 800,000 SNPs with minor allele frequency (MAF) larger than 0.01.

From these 800,000 SNPs, we selected 20 evenly spaced SNPs to be the causal SNPs. We set the regression coefficients for 5 of these causal SNPs to 0.2, and for 5 other causal SNPs to -0.2. In addition, the regression coefficients for the other 10 causal SNPs have the same value β , but that value varies in six settings: 0.2, 0.3, 0.4, 0.5, 0.7 and 1. Further, we set the intercept at $\beta_0 = -0.5$. Furthermore, the variance component κ of the kinship random effects $\boldsymbol{\alpha}$ is set to 0.15. Thus, the binary phenotype data are simulated from a Bernoulli GLMM with logistic link function and linear predictor $\beta_0 + \sum_{i=1}^{10} \beta x_{ij} + \sum_{i=11}^{15} 0.2x_{ij} + \sum_{i=16}^{20} (-0.2)x_{ij} + \alpha_i$, with $\boldsymbol{\alpha} \sim N(\mathbf{0}, \kappa\Sigma)$ where Σ is the kinship matrix.

Figure 3.1 shows for binary data the performance of our BG2 method with three different ways to choose the parameter τ , as well as the performance of the SMA method. These performances in terms of TP, FP, FDR, and F1 averaged over 100 datasets for each setting are plotted as functions of the varying regression coefficient β . In addition, Figure 3.1 shows the computational time. Our BG2 methods take twice as long as SMA, which is to be expected since SMA has only a screening step whereas BG2 has a screening step and a model selection step. Among the four ways considered to choose τ for BG2, estimating τ

based on a uniform prior provides higher F1 scores for smaller values of β , and provides comparable F1 scores for larger values of β . In addition, when compared to SMA, BG2 with uniform prior provides larger average number of true positives TP than when β is small, and a smaller TP when β is large. However, BG2 with uniform prior leads to a much smaller average number of false positives than SMA. As a result, when compared to SMA, for all considered values of the regression coefficient β , BG2 with uniform prior has much larger F1. Therefore, for binary GWAS data we recommend BG2 with a uniform prior for τ .

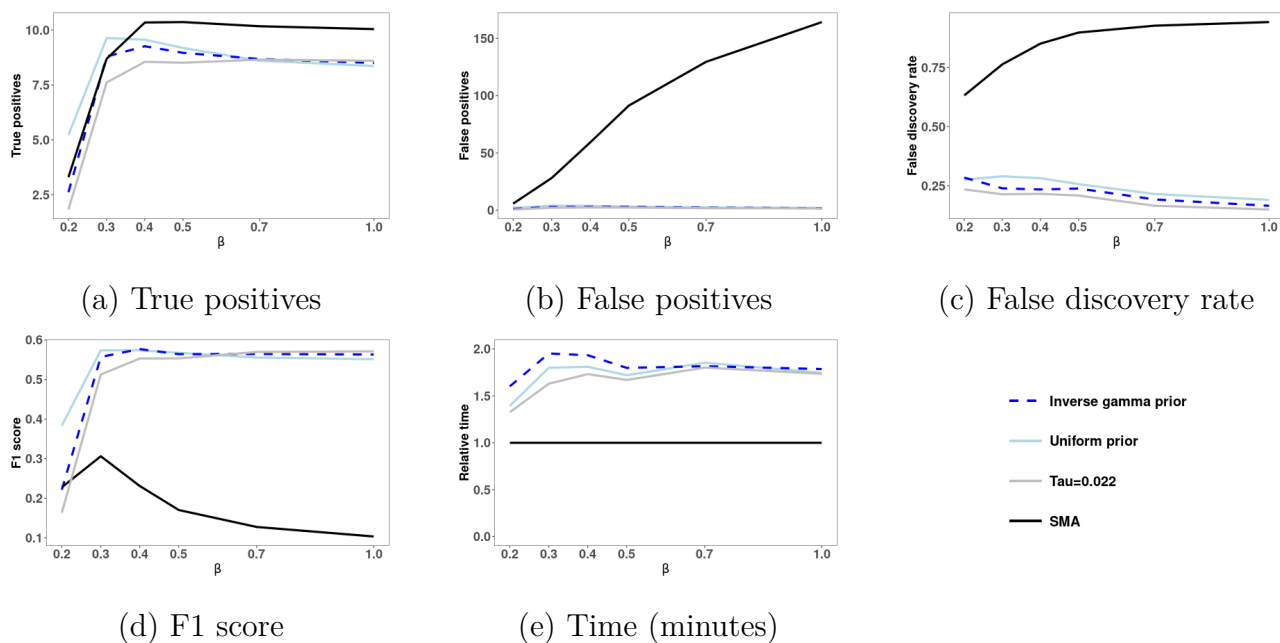


Figure 3.1: Results for simulated binary data. SNP search performance of four methods (SMA and three proposed BG2 methods: uniform prior for τ , inverse gamma prior for τ , and fixed $\tau = 0.022$) averaged over 100 datasets under each parameter setting $\beta = 0.2, 0.3, 0.4, 0.5, 0.7, 1$ respectively. Five criteria: True positives (TP), false positives (FP), false discovery rate (FDR), F1 score (F1) and computational time.

Finally, we have tested the robustness of BG2 to the case of binary GWAS data with no causal SNPs. Specifically, we have simulated 100 datasets with binary GWAS data from a Bernoulli GLMM with logit link function and linear predictor $\beta_0 + \alpha_i$. While BG2 with any of the ways to choose τ does not yield any false positive for 100 simulated datasets, SMA

has an average of 0.06 false positives. Therefore, BG2 performs better than SMA for binary GWAS data and is robust to the case when there are no causal SNPs.

3.4.2 Count data

We simulate count GWAS data using genotype information from The Arabidopsis Information Resource (TAIR9) (<https://www.arabidopsis.org/>). This simulation study is based on a case study on root-like structures in *A. Thaliana* that we present in Section 3.5.3.

Specifically, we use 188,980 SNPs with $MAF > 0.01$ from 152 ecotypes of *A. Thaliana*. This simulation study assumes 10 causal SNPs evenly located among all available SNPs. Of these 10 causal SNPs, 5 causal SNPs have fixed coefficients equal to 0.2, and the other 5 causal SNPs have the same coefficient β which varies in eight settings: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7 and 1. In addition, we set the intercept β_0 equal to 1. Further, we assume that there are two random effects: a kinship random effect α_1 with variance component κ_1 equal to 1; and an overdispersion random effect α_2 with variance component κ_2 equal to 0.3, which is close to the estimate obtained in the case study presented in Section 3.5.3. Let r_i be the number of replicates of ecotype i . Because in the case study most ecotypes have 12 replicates, in this simulation study we assume that all ecotypes have 12 replicates. In addition, the phenotype y_i for ecotype i is the total number of root-like structures of the r_i replicates. These phenotype count data are sampled from a Poisson GLMM with logarithm link function and linear predictor $\log(r_i) + \beta_0 + \sum_{i=1}^5 \beta x_{ij} + \sum_{i=6}^{10} 0.2 x_{ij} + \alpha_{1i} + \alpha_{2i}$.

Figure 3.2 shows for count data the performance of our BG2 method as well as the performance of the SMA method. These performances are averaged over 100 simulated datasets for each setting and plotted as functions of the varying regression coefficient β . In addition, Figure 3.2 shows the computational time. Our BG2 methods take about eight times longer

than SMA, but they still provide results in a feasible amount of time. Among the four ways considered to choose τ for BG2, estimating τ based on an inverse gamma prior provides larger average number of true positives and about the same FDR level. As a result, when compared to the other ways to choose τ , estimating τ based on an inverse gamma prior has higher F1 scores for most considered values of β . In addition, when compared to SMA, BG2 with an inverse gamma prior provides larger average number of true positives TP for most considered values of β . Further, BG2 with inverse gamma prior has about the same FDR level as SMA for $\beta \leq 0.5$ and a much smaller FDR level for $\beta > 0.5$. As a result, while BG2 with an inverse gamma prior has comparable F1 to SMA for small values of β , the F1 of BG2 with an inverse gamma prior becomes much larger than the F1 of SMA as β increases.

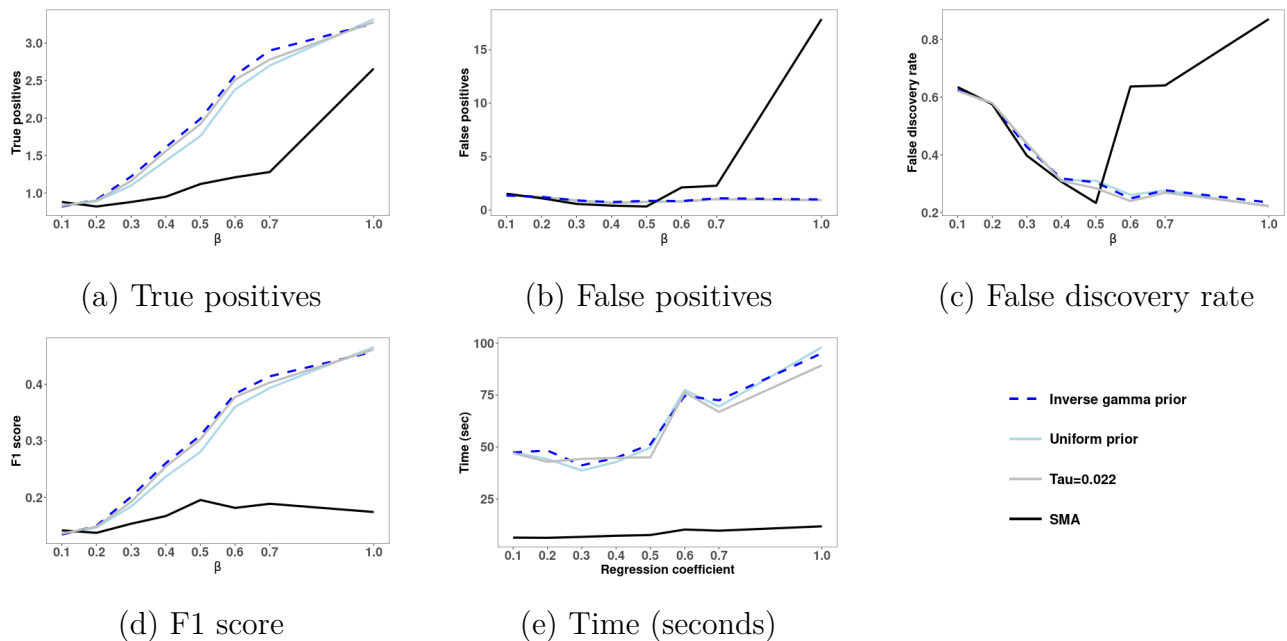


Figure 3.2: Results for simulated count data. SNP search performance of four methods (SMA and three proposed BG2 methods: uniform prior for τ , inverse gamma prior for τ , and fixed $\tau = 0.022$) averaged over 100 datasets under each parameter setting $\beta = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 1$ respectively. Five criteria: True positives (TP), false positives (FP), false discovery rate (FDR), F1 score (F1) and computational time.

In addition, we have tested the robustness of BG2 to the case of count GWAS data with no

causal SNPs. Specifically, we have simulated 100 datasets with count GWAS data from a Poisson GLMM with logarithm link function and linear predictor $\beta_0 + \alpha_{1i} + \alpha_{2i}$. The average number of false positives for all considered methods is 0. Thus, both SMA and BG2 methods perform well in the case of count GWAS data with no causal SNPs.

Therefore, because of its favorable performance when there are causal SNPs, for count GWAS data we recommend BG2 with an inverse gamma prior for τ .

3.5 Case studies

To illustrate the usefulness and flexibility of BG2, this section presents three case studies on cocaine dependence, alcohol consumption, and number of root-like structures in *A. Thaliana*.

3.5.1 Maximum number of alcoholic drinks

The Collaborative Study on the Genetics of Alcoholism (COGA) [5] was a large-scale family study that had as primary objective to identify genes related to alcohol dependence. Here, we perform a GWAS analysis of the maximum number of alcoholic drinks consumed in 24 hours. We analyze data on 2759 European Americans considering 846,076 SNPs with $MAF > 0.01$. To perform this analysis, we use the model for count data considered in Section 3.4.2.

While SMA detects 10 SNPs, BG2 detects only one SNP which is located in the protein-coding gene PTGER4 on chromosome 5. The protein encoded by PTGER4 is a receptor for prostaglandin E2 (PGE2). An increase in PGE2 is part of the inflammatory response to alcohol consumption, and the use of the PGE2-inhibitor tolfenamic acid significantly reduces the severity of several hangover symptoms [69].

3.5.2 Cocaine dependence

In this case study, we analyze the association between cocaine dependence and single nucleotide polymorphisms (SNPs). We analyze data from the Family Study of Cocaine Dependence (FSCD) [7], which was part of the Study of Addiction: Genetics and Environment. Specifically, we analyze data on 2,767 European Americans considering 846,076 SNPs with $MAF > 0.01$. Because males and females seem to have different behaviors with respect to cocaine, we include sex as a control covariate. To perform this analysis, we use the model for binary data considered in Section 3.4.1.

BG2 detects one SNP, which is located at the protein-coding gene *ABCC8* on chromosome 11. For this dataset, SMA only detects the same SNP. The protein encoded by this gene is a member of the superfamily of ATP-binding cassette (ABC) proteins which transport various molecules across extra-cellular and intra-cellular membranes. In addition, a quantitative transcriptomics analysis (RNA-Seq) has shown that this gene is overexpressed in the brain [18]. Further, cocaine increases expression of *ABCC1* (another gene that encodes an ABC protein) in mice [13]. Finally, *ABCC1*-siRNA (a silencer of *ABCC1*) blocks cocaine-induced place preference in mice [13].

3.5.3 Root-like structures in *A. Thaliana*

To illustrate the application of our method to count data, we analyze data on root-like structures in *A. Thaliana* from a study of plant regeneration from root explants [36]. Specifically, we note that [36] applied a square root transformation to count phenotype data. In contrast, we use the Poisson GLMM with overdispersion considered in Section 3.4.2 to analyze the original count data. We focus on the number of root-like structures after 21 days in which seedlings are under warm white light at 21°C following a 14/10 h light/dark regime. There

are 188,980 SNPs from TAIR9 with $MAF > 0.01$.

BG2 detects 3 SNPs. For this dataset, SMA detects the same 3 SNPs. These 3 SNPs are expressed in the root and are located in protein-coding genes AT1G20090, AT1G20100 and AT1G20720. Specifically, AT1G20100 encodes a DNA ligase-like protein involved in the regulation of metabolic processes. In addition, gene AT1G20720 encodes a RAD3-like DNA binding helicase protein that acts in the repair of double-strand breaks in DNA, and in nucleotide-excision repair. Finally, AT1G20090 encodes a ROP2 protein which is known to effect root hair initiation and tip growth [28].

3.6 Discussion

We have proposed BG2, a two-stage Bayesian SNP detection method for non-Gaussian GWAS data. BG2 uses a GLMM framework that includes kinship random effects and overdispersion random effects. BG2 has two steps: a screening step and a model selection step. The screening step performs a Bayesian SMA that selects a set of candidate SNPs. The model selection step then considers all possible GLMMs based on this set of candidate SNPs. To speed up computations, we develop a pseudo likelihood approach combined with P3D. Further, we develop a novel class of nonlocal priors for the regression coefficients specially tailored for GLMMs. Simulation studies show that, for both binary and count GWAS data, BG2 is much better than SMA in terms of FDR and F1.

There are several possible avenues for future research. One promising research direction is to adapt BG2 for application to biobank scale data. Another possible research direction is to implement BG2 with an iterative procedure that would allow smaller effect sizes to be detected. Finally, another possible research avenue is to develop BG2 for GWAS analysis when the phenotype is survival time.

3.7 Supplementary Material

3.7.1 The pseudo-likelihood approach

In this section, we explain in detail the pseudo-likelihood method for the analysis of non-Gaussian GWAS data. The main point is that our approach performs different estimation procedures for baseline models and for non-baseline models. In what follows, the first part provides details on the estimation procedure for baseline models and the second part provides details on the estimation procedure for non-baseline models.

The pseudo-likelihood approach for the baseline models

For fitting of baseline models, our approach uses an iterative pseudo-likelihood algorithm. The algorithm starts with some initial estimates of the parameters β_c , β_s , α_1 , and α_2 , typically by assuming a generalized linear model. Denote by $\hat{\beta}_c$, $\hat{\beta}_s$, $\hat{\alpha}_1$, and $\hat{\alpha}_2$ the current estimates of β_c , β_s , α_1 , and α_2 , respectively. The pseudo-likelihood method is based on a first-order Taylor expansion. Write the vector of observations \mathbf{y} as the sum of its mean vector $\boldsymbol{\mu}$ and an error vector $\boldsymbol{\epsilon}$, that is $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$. Here, the mean vector is $\boldsymbol{\mu} = B'(\boldsymbol{\eta}) = m(\boldsymbol{\eta})$ and the covariance matrix of the error vector is the diagonal matrix $V = \text{diag}(v(\boldsymbol{\eta}))$, where $v(\boldsymbol{\eta}) = (v_1, \dots, v_n)' = (B''(\eta_1), \dots, B''(\eta_n))'$. Let X_c be the matrix of control covariates and X_s be the matrix of SNPs. Expand the mean vector using a first-order Taylor expansion about $\hat{\beta}_c$, $\hat{\beta}_s$, $\hat{\alpha}_1$, and $\hat{\alpha}_2$. Then,

$$\begin{aligned} \mathbf{y} &= \boldsymbol{\mu} + \boldsymbol{\epsilon} \\ &= B'(\boldsymbol{\eta}) + \boldsymbol{\epsilon} \\ &\approx B'(\hat{\boldsymbol{\eta}}) + B''(\hat{\boldsymbol{\eta}})[X_c(\beta_c - \hat{\beta}_c) + X_s(\beta_s - \hat{\beta}_s) + \alpha_1 - \hat{\alpha}_1 + \alpha_2 - \hat{\alpha}_2] + \boldsymbol{\epsilon} \end{aligned}$$

$$\approx B'(X_c\widehat{\beta}_c + X_s\widehat{\beta}_s + \widehat{\alpha}_1 + \widehat{\alpha}_2) + B''(X_c\widehat{\beta}_c + X_s\widehat{\beta}_s + \widehat{\alpha}_1 + \widehat{\alpha}_2)[X_c(\beta_c - \widehat{\beta}_c) + X_s(\beta_s - \widehat{\beta}_s) + \alpha_1 - \widehat{\alpha}_1 + \alpha_2 - \widehat{\alpha}_2] + \epsilon. \quad (\text{S.3})$$

where $\widehat{\beta}_c$, $\widehat{\beta}_s$, $\widehat{\alpha}_1$, and $\widehat{\alpha}_2$ are the current estimates of β_c , β_s , α_1 , and α_2 , respectively.

Denote the current estimate of the mean vector μ evaluated at $\widehat{\beta}_c$, $\widehat{\beta}_s$, $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ by $\widehat{\mu}$. That is,

$$\widehat{\mu} = m(X_c\widehat{\beta}_c + X_s\widehat{\beta}_s + \widehat{\alpha}_1 + \widehat{\alpha}_2).$$

In addition, denote the current estimate of the covariance matrix of ϵ evaluated at $\widehat{\beta}_c$, $\widehat{\beta}_s$, $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ by \widehat{V} , that is

$$\widehat{V} = \text{diag}[v(X_c\widehat{\beta}_c + X_s\widehat{\beta}_s + \widehat{\alpha}_1 + \widehat{\alpha}_2)].$$

Reorganize Equation (S.3) by moving $\widehat{\mu}$ to the left side, pre-multiplying \widehat{V}^{-1} and then moving $X_c\widehat{\beta}_c$, $X_s\widehat{\beta}_s$, $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ to the left side. Let \mathbf{y}^* be equal to the left side of the resulting equation, that is

$$\mathbf{y}^* = \widehat{V}^{-1}(\mathbf{y} - \widehat{\mu}) + X_c\widehat{\beta}_c + X_s\widehat{\beta}_s + \widehat{\alpha}_1 + \widehat{\alpha}_2.$$

The vector \mathbf{y}^* is known as the vector of adjusted observations, which is computed as a function of the current estimates of fixed effects $\widehat{\beta}_c$ and $\widehat{\beta}_s$, and random effects $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$. The right side of the resulting equation is $X_c\beta_c + X_s\beta_s + \alpha_1 + \alpha_2 + \widehat{V}^{-1}\epsilon$. Hence, we obtain the following approximate model for the adjusted observations

$$\mathbf{y}^* \approx X_c\beta_c + X_s\beta_s + \alpha_1 + \alpha_2 + \widehat{V}^{-1}\epsilon.$$

Further, assuming that $\widehat{V}^{-1}V\widehat{V}^{-1} \approx \widehat{V}^{-1}$ and applying properties of expectation and vari-

ance, we get

$$\begin{aligned} E(\mathbf{y}^*) &= X_c \boldsymbol{\beta}_c + X_s \boldsymbol{\beta}_s, \\ \text{Var}(\mathbf{y}^*) &\approx \kappa_1 \Sigma + \kappa_2 I + \widehat{V}^{-1}. \end{aligned}$$

If we further assume that $\boldsymbol{\epsilon}$ has an approximate normal distribution, then

$$\mathbf{y}^* \sim N \left(X_c \boldsymbol{\beta}_c + X_s \boldsymbol{\beta}_s, \kappa_1 \Sigma + \kappa_2 I + \widehat{V}^{-1} \right).$$

As we explain below, estimates of $\boldsymbol{\beta}_c$, $\boldsymbol{\beta}_s$, $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, κ_1 and κ_2 are updated iteratively. After convergence, we have the final vector of adjusted observations \mathbf{y}^* and the approximate LMM

$$\begin{aligned} \mathbf{y}^* &\approx X_c \boldsymbol{\beta}_c + X_s \boldsymbol{\beta}_s + \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 + \widehat{V}^{-1} \boldsymbol{\epsilon}, \\ \boldsymbol{\alpha}_1 &\sim N(\mathbf{0}, \kappa_1 \Sigma), \\ \boldsymbol{\alpha}_2 &\sim N(\mathbf{0}, \kappa_2 I), \\ \boldsymbol{\epsilon} &\sim N(\mathbf{0}, V). \end{aligned}$$

The closed form of the likelihood function with respect to the unknown parameters is

$$\begin{aligned} L(\boldsymbol{\beta}_c, \boldsymbol{\beta}_s, \kappa_1, \kappa_2 | \mathbf{y}^*) &= (2\pi)^{-\frac{n}{2}} \left| \kappa_1 \Sigma + \kappa_2 I + \widehat{V}^{-1} \right|^{-\frac{1}{2}} \\ &\exp \left\{ -\frac{1}{2} (\mathbf{y}^* - X_c \boldsymbol{\beta}_c - X_s \boldsymbol{\beta}_s)^T (\kappa_1 \Sigma + \kappa_2 I + \widehat{V}^{-1})^{-1} (\mathbf{y}^* - X_c \boldsymbol{\beta}_c - X_s \boldsymbol{\beta}_s) \right\}. \end{aligned}$$

Let $\widehat{\kappa}_1$ and $\widehat{\kappa}_2$ be current estimates of the variance components. We update $\widehat{\boldsymbol{\beta}}_c$ and $\widehat{\boldsymbol{\beta}}_s$ with the conditional posterior mean of $\boldsymbol{\beta} = (\boldsymbol{\beta}'_c, \boldsymbol{\beta}'_s)'$, given by $\widehat{\boldsymbol{\beta}} = (X^T \widehat{H}^{-1} X)^{-1} X^T \widehat{H}^{-1} \mathbf{y}^*$, where $X = (X_c, X_s)$ and $\widehat{H} = \widehat{\kappa}_1 \Sigma + \widehat{\kappa}_2 I + \widehat{V}^{-1}$. Given current estimates $\widehat{\boldsymbol{\beta}}_c$ and $\widehat{\boldsymbol{\beta}}_s$, we use the

conditional posterior mean to update the estimates of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$, that is

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_1 &= E(\boldsymbol{\alpha}_1|\mathbf{y}^\star) \\ &= \hat{\kappa}_1 \Sigma \hat{H}^{-1}(\mathbf{y}^\star - X_c \hat{\boldsymbol{\beta}}_c - X_s \hat{\boldsymbol{\beta}}_s), \\ \hat{\boldsymbol{\alpha}}_2 &= \hat{\kappa}_2 \hat{H}^{-1}(\mathbf{y}^\star - X_c \hat{\boldsymbol{\beta}}_c - X_s \hat{\boldsymbol{\beta}}_s).\end{aligned}\tag{3.12}$$

And then, to update the estimates of κ_1 and κ_2 , we maximize the profile pseudo-likelihood

$$\begin{aligned}\log L(\kappa_1, \kappa_2|\mathbf{y}^\star) &\propto -\frac{1}{2} \log \left| \kappa_1 \Sigma + \kappa_2 I + \hat{V}^{-1} \right| \\ &\quad - \frac{1}{2} (\mathbf{y}^\star - X_c \hat{\boldsymbol{\beta}}_c - X_s \hat{\boldsymbol{\beta}}_s)^T \left(\kappa_1 \Sigma + \kappa_2 I + \hat{V}^{-1} \right)^{-1} (\mathbf{y}^\star - X_c \hat{\boldsymbol{\beta}}_c - X_s \hat{\boldsymbol{\beta}}_s),\end{aligned}$$

obtaining the estimates $\hat{\kappa}_1, \hat{\kappa}_2 = \mathit{argmax} \log L(\kappa_1, \kappa_2|\mathbf{y}^\star)$.

The pseudo-likelihood algorithm proceeds iteratively updating the parameters until convergence. Algorithm 1 summarizes the pseudo-likelihood approach for baseline models.

Model fitting for non-baseline models

In each BG2 step, a baseline model is fitted with the pseudo-likelihood approach presented in the first part of Section 3.7.1. This results in estimated variance parameters and a vector of adjusted observations \mathbf{y}^\star . After that, these variance parameter estimates and vector of adjusted observations \mathbf{y}^\star are used to fit the non-baseline models in that BG2 step. As a result, fitting a non-baseline model does not have any iteration, but just computes the estimate of the regression coefficients with the formula $\hat{\boldsymbol{\beta}} = (X^T \hat{H}^{-1} X)^{-1} X^T \hat{H}^{-1} \mathbf{y}^\star$. The matrix H , which is a function of the variance parameters, is estimated with the baseline model at the beginning of the respective BG2 step, and then it remains fixed for all non-baseline models. Hence, the eigen decomposition of H can be computed at the beginning of the BG2 step

Algorithm 2 Pseudo-likelihood approach for baseline models

```

procedure PSEUDO LIKELIHOOD( $\mathbf{y}, X_c, X_s$ )
  Initial values:  $\beta_c^{(0)}, \beta_s^{(0)}$  = estimates from GLM,  $\alpha_1^{(0)}, \alpha_s^{(0)} = \mathbf{0}$ ,  $\kappa_1^{(0)}, \kappa_2^{(0)} = 0$ .
  Calculate  $\mu^{(0)}, V^{(0)}, H^{(0)}$  and  $\mathbf{y}^{\star(0)}$ .
  while  $\beta_c, \beta_s, \kappa_1$  and  $\kappa_2$  not converge do
     $\beta_c^{(t)} = (X_c^T H^{(t-1)} - X_c) - X_c^T H^{(t-1)} - \mathbf{y}^{\star(t-1)}$ 
     $\beta_s^{(t)} = (X_s^T H^{(t-1)} - X_s) - X_s^T H^{(t-1)} - \mathbf{y}^{\star(t-1)}$ 
     $\alpha_1^{(t)} = \kappa_1^{(t-1)} \Sigma H^{(t-1)} - (\mathbf{y}^{\star(t-1)} - X_c \beta_c^{(t)} - X_s \beta_s^{(t)})$ 
     $\alpha_2^{(t)} = \kappa_2^{(t-1)} H^{(t-1)} - (\mathbf{y}^{\star(t-1)} - X_c \beta_c^{(t)} - X_s \beta_s^{(t)})$ 
     $\kappa_1^{(t)}, \kappa_2^{(t)} = \operatorname{argmax} \log L(\kappa_1, \kappa_2 | \mathbf{y}^{\star(t-1)})$ 
    Update  $\mu^{(t)}, V^{(t)}, H^{(t)}$  and  $\mathbf{y}^{\star(t)}$ 
  end while
end procedure

```

and the same linear algebra trick used in EMMAX can be used in BG2. Therefore, fitting non-baseline models is super fast.

The baseline model differs in the screening step and in the model selection step. In the screening step, the baseline model is the null model with no SNPs, that is, X_s is not included in the null model. In the model selection step, the baseline model is the full model with X_s containing all candidate SNPs identified in the screening step.

3.7.2 Results of simulation study of count data simulated with human genome

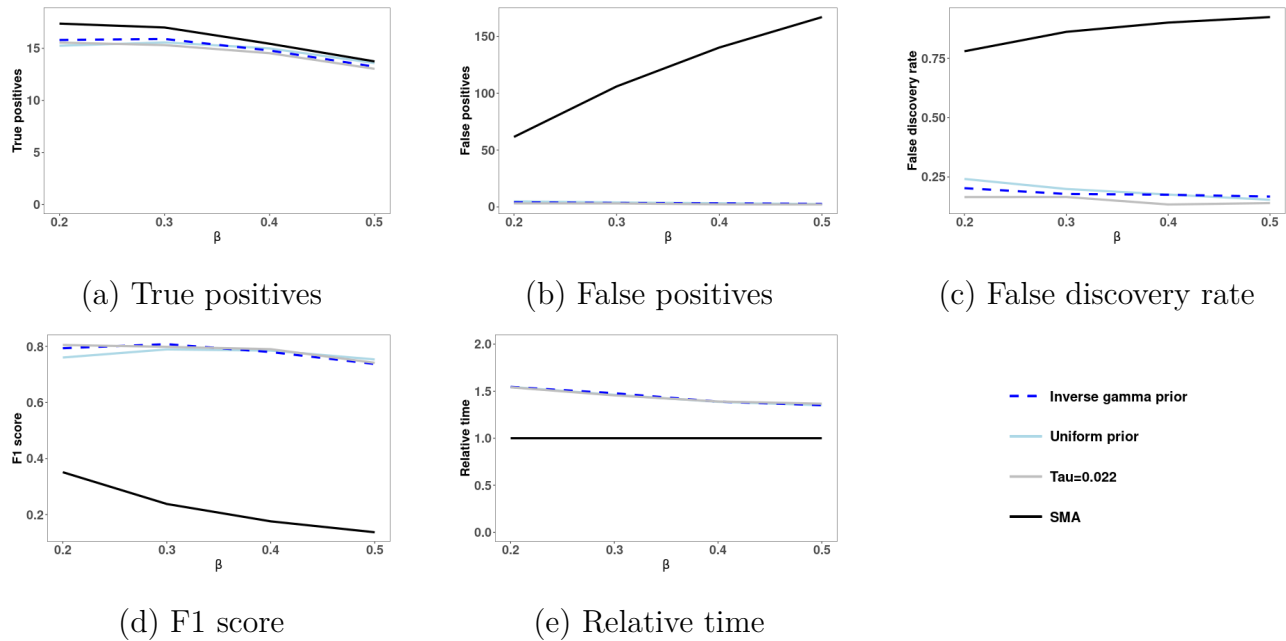


Figure 3.3: Results for simulated count data based on human genome. SNP search performance of four methods (SMA and three proposed BG^2 methods: uniform prior for τ , inverse gamma prior for τ , and fixed $\tau = 0.022$) averaged over 100 datasets under each parameter setting $\beta = 0.2, 0.3, 0.4, 0.5$ respectively. Intercept $\beta_0 = -0.5$, variance component for kinship random effects $\kappa_1 = 0.1$, variance component for overdispersion random effects $\kappa_2 = 0.05$. Five criteria: True positives (TP), false positives (FP), false discovery rate (FDR), F1 score (F1) and Relative time with respect to SMA.

3.7.3 Boxplots of TP, FP, FDR, and F1 in the simulation studies

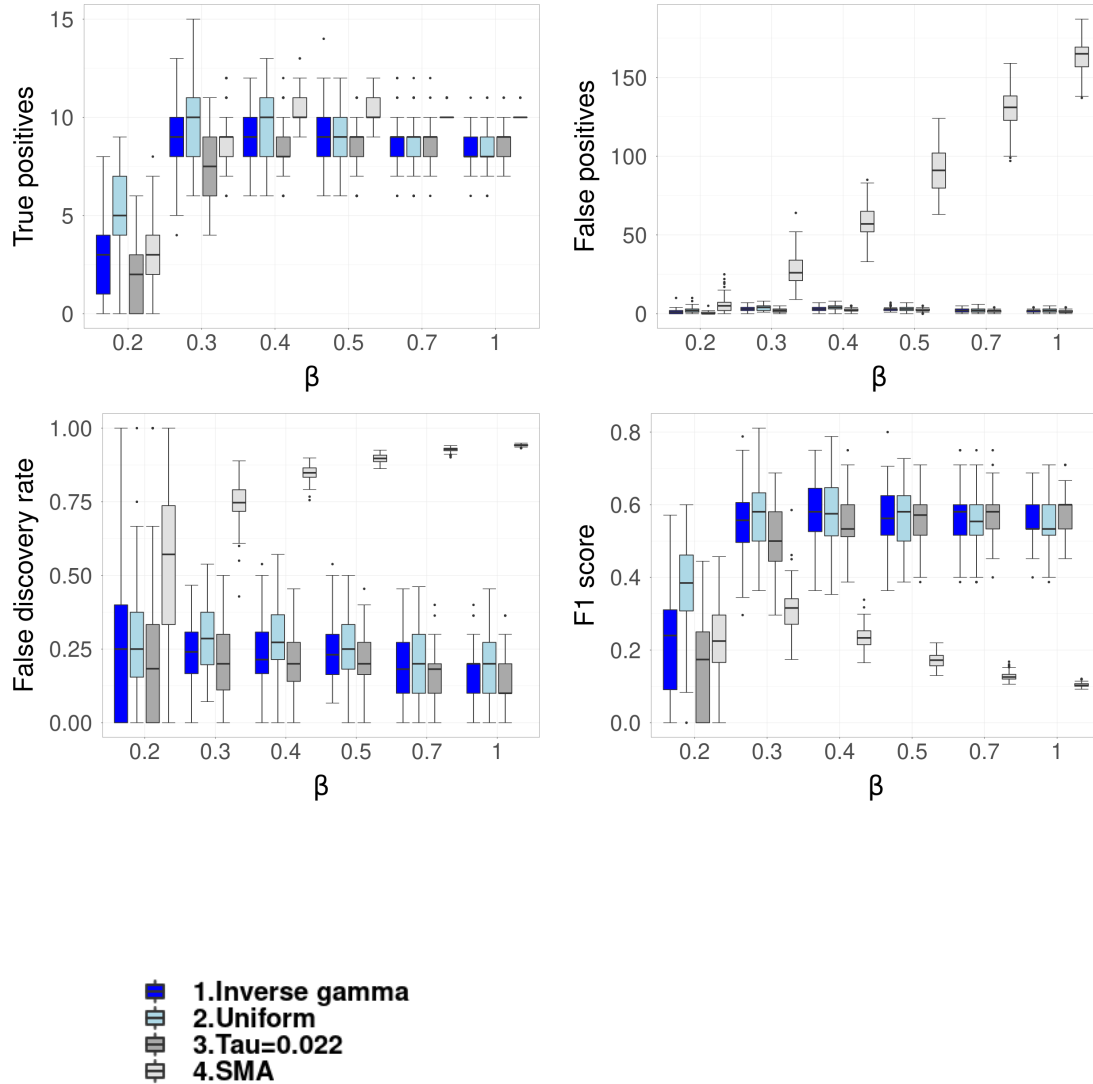


Figure 3.4: Results for simulated binary data based on human genome. SNP search performance of four methods (SMA and three proposed BG2 methods: uniform prior for τ , inverse gamma prior for τ , and fixed $\tau = 0.022$) for 100 simulated datasets under each parameter setting $\beta = 0.2, 0.3, 0.4, 0.5, 0.7, 1$ respectively. Boxplots of four criteria: True positives (TP), false positives (FP), false discovery rate (FDR), and F1 score (F1).

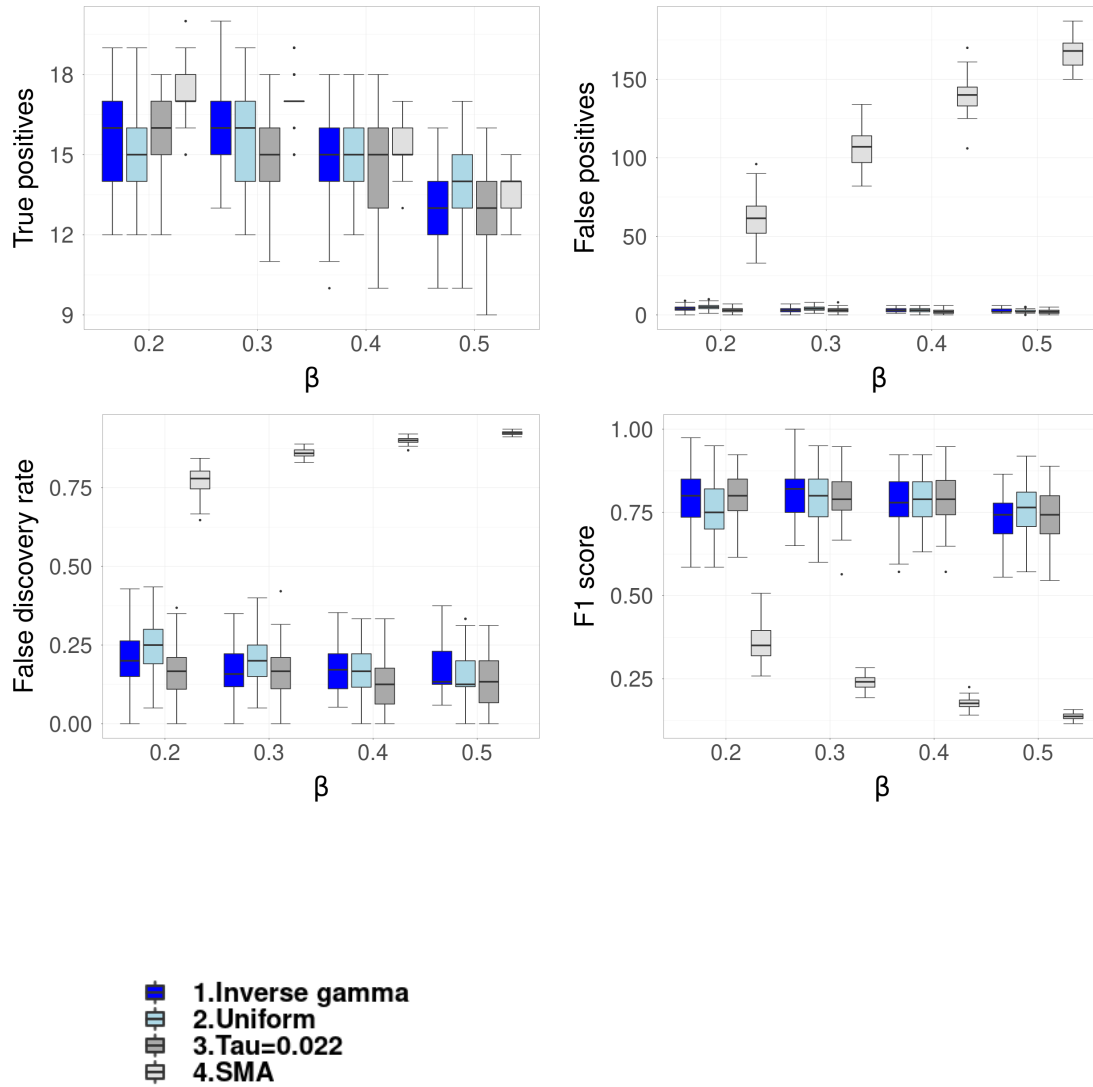


Figure 3.5: Results for simulated count data based on human genome. SNP search performance of four methods (SMA and three proposed BG2 methods: uniform prior for τ , inverse gamma prior for τ , and fixed $\tau = 0.022$) for 100 simulated datasets under each parameter setting $\beta = 0.2, 0.3, 0.4, 0.5$ respectively. Boxplots of four criteria: True positives (TP), false positives (FP), false discovery rate (FDR), and F1 score (F1).

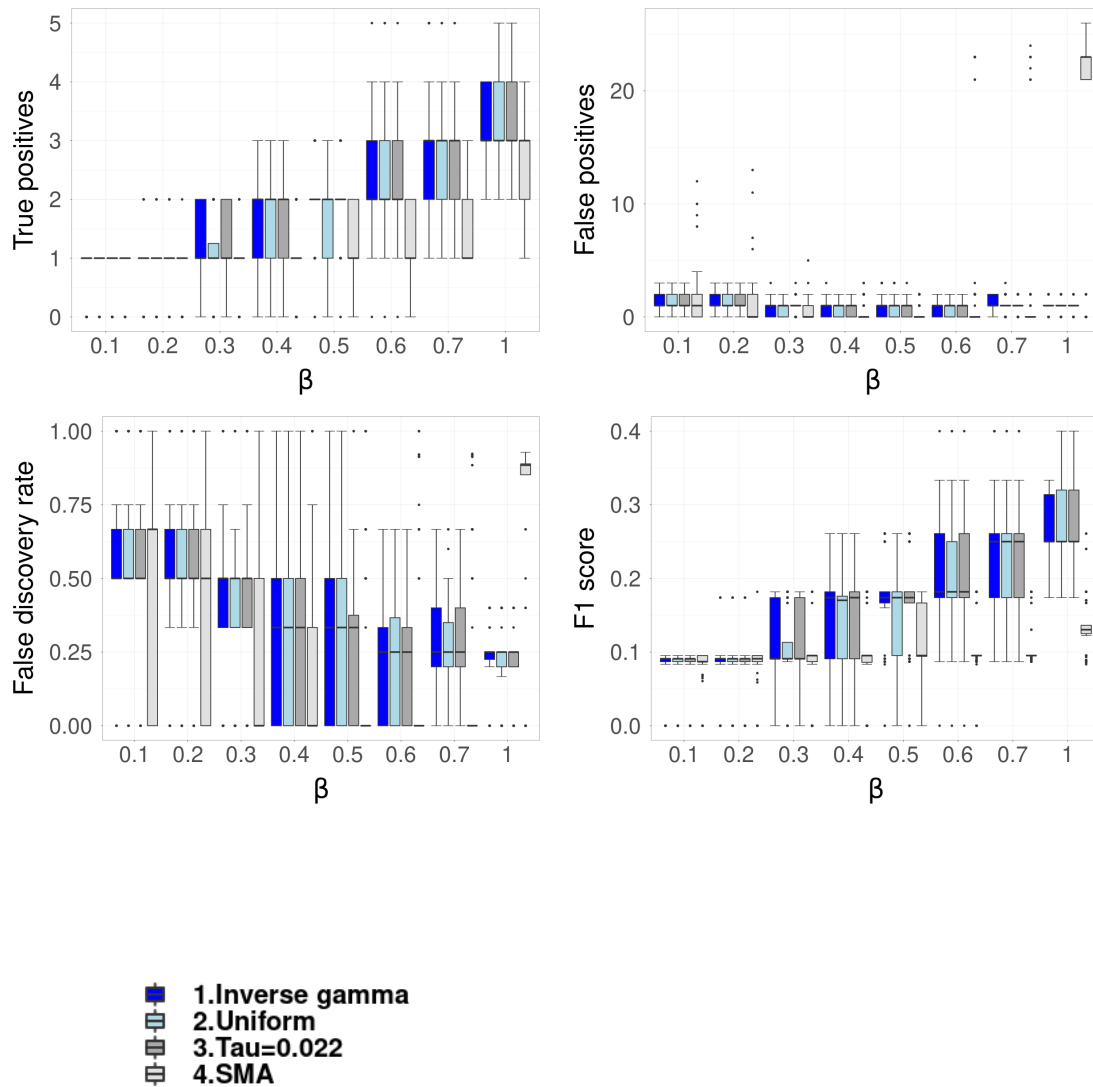


Figure 3.6: Results for simulated count data based on A. Thaliana genome. SNP search performance of four methods (SMA and three proposed BG2 methods: uniform prior for τ , inverse gamma prior for τ , and fixed $\tau = 0.022$) for 100 simulated datasets under each parameter setting $\beta = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 1$ respectively. Boxplots of four criteria: True positives (TP), false positives (FP), false discovery rate (FDR), and F1 score (F1).

3.7.4 Robustness of BG2 when dealing with imbalanced binary data or highly skewed count data

In the original simulation studies for count data presented in Section 3.4, we have highly skewed count data. To visualize that, Figure 3.7(a) shows that when $\beta = 0.1$ the count data are skewed. In addition, Figure 3.7(b) shows that when $\beta = 0.7$ the count data are tremendously skewed. Table 3.1 shows that skewed count data do not affect variable selection. As a matter of fact, when BG2 is applied to more skewed data ($\beta = 0.7$) the performance of BG2 improves with higher TP, lower FP and FDR, and higher F1 score.

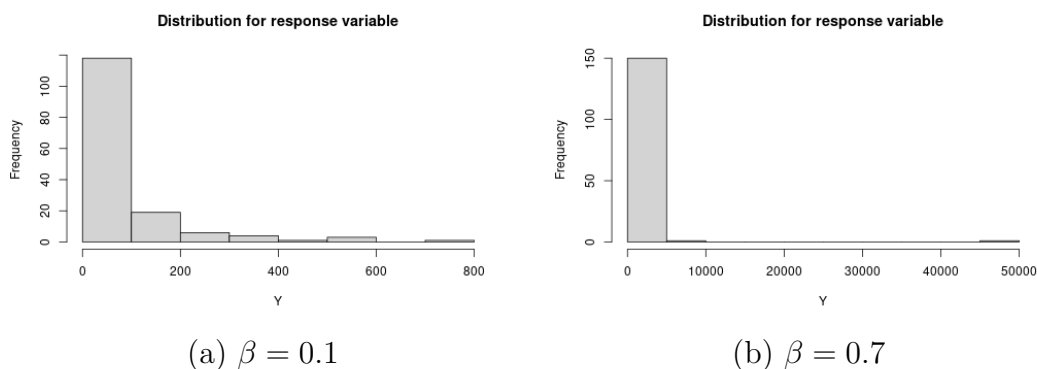


Figure 3.7: Histograms for two simulated count datasets from Section 3.4.2. The simulation setting has 10 Causal SNPs, 5 SNPs with coefficient 0.2, 5 SNPs with coefficient β . (a) $\beta = 0.1$. (b) $\beta = 0.7$.

	Method	TP	FP	FDR	F1
$\beta = 0.1$	Inverse gamma	0.82	1.37	0.63	0.13
	Uniform	0.82	1.36	0.62	0.13
	$\tau = 0.022$	0.83	1.36	0.62	0.14
	SMA	0.88	1.53	0.63	0.14
$\beta = 0.7$	Inverse gamma	2.90	1.11	0.28	0.41
	Uniform	2.70	1.03	0.28	0.39
	$\tau = 0.022$	2.78	1.02	0.27	0.40
	SMA	1.28	2.28	0.64	0.19

Table 3.1: Performance of BG2 and SMA when data are skewed or tremendously skewed. When $\beta = 0.1$, count data are skewed. When $\beta = 0.7$, count data are tremendously skewed.

The binary data we have in the original simulation study presented in Section 3.4.1 is almost balanced, with about 56% 0s and 44% 1s. We have added a new simulation study with $\beta_0 = 2$, which is highly imbalanced with about 29% 0s and 71% 1s. Table 3.2 shows that BG2 is robust to imbalanced data, and even performs slightly better when the data are imbalanced.

	Method	TP	FP	FDR	F1
$\beta_0 = -0.5$	Inverse gamma	8.50	1.68	0.17	0.84
	Uniform	8.36	1.96	0.19	0.82
	$\tau = 0.022$	8.60	1.51	0.15	0.86
	SMA	10.04	164.25	0.94	0.11
$\beta_0 = 2$	Inverse gamma	8.77	1.50	0.15	0.87
	Uniform	8.74	1.58	0.15	0.86
	$\tau = 0.022$	8.73	1.34	0.13	0.87
	SMA	10.01	156.53	0.94	0.11

Table 3.2: Performance of BG2 and SMA when data are balanced or imbalanced. When $\beta_0 = -0.5$, binary data are balanced. When $\beta_0 = 2$, binary data are imbalanced.

3.7.5 Robustness of BG2 to genome spacing of SNPs

To verify the robustness of BG2 to genome spacing of SNPs, we have added three new simulation studies that we name SIM1, SIM2, and SIM3. These simulation studies expand the simulation study from Section 3.4.1 for binary data based on human genome.

In the first simulation study SIM1, we generate data from 20 causal SNPs, which are from 4 clusters. In each cluster, there are 5 causal SNPs. Each cluster has a length of 30000 bp. In two clusters, two SNPs have large coefficient $\beta = 1$. In another two clusters, three SNPs have large coefficient $\beta = 1$. All the other SNPs have small coefficient 0.2 or -0.2 . For reference, simulation study SIM0 in the Table 3.3 is the simulation study in Section 3.4.1, which has the same parameter setting except that 20 SNPs are evenly spaced.

In the second simulation study SIM2, we generate data from 10 causal SNPs, which are from

2 clusters. In each cluster, there are 5 causal SNPs. Each cluster has a length of 30000 bp. In one cluster, two SNPs have large coefficient $\beta = 1$. In another two clusters, three SNPs have large coefficient $\beta = 1$. All the other SNPs have small coefficient 0.2 or -0.2 .

In the third simulation study SIM3, we generate data from 5 causal SNPs, which are from only one cluster. The length of the cluster is 30000 bp. Three SNPs have large coefficient $\beta = 1$. One SNP has coefficient 0.2, and another SNP has coefficient -0.2 .

Table 3.3 shows that BG2 can detect almost all SNPs with large coefficient. The number of causal SNPs and the position of SNPs do not alter the performance of the method BG2. Comparing SIM0 and SIM1, BG2 for clustering causal SNPs has lower FP and FDR, and higher F1. Comparing SIM1, SIM2, and SIM3, small number of causal SNPs make BG2 have lower FDR and higher F1 score.

Simulation	Method	TP	FP	FDR	F1
SIM0	Inverse gamma	8.50	1.68	0.17	0.56
	Uniform	8.36	1.96	0.19	0.55
	$\tau = 0.022$	8.60	1.51	0.15	0.57
	SMA	10.04	164.25	0.94	0.10
SIM1	Inverse gamma	8.68	0.36	0.04	0.60
	Uniform	8.66	0.44	0.05	0.60
	$\tau = 0.022$	8.56	0.34	0.04	0.59
	SMA	17.11	75.00	0.81	0.31
SIM2	Inverse gamma	4.94	0.02	0.004	0.66
	Uniform	4.92	0.02	0.006	0.66
	$\tau = 0.022$	4.91	0.03	0.004	0.66
	SMA	8.53	47.22	0.847	0.26
SIM3	Inverse gamma	2.88	0.01	0.003	0.73
	Uniform	2.92	0.01	0.003	0.74
	$\tau = 0.022$	2.79	0.00	0.000	0.72
	SMA	5.00	48.50	0.907	0.17

Table 3.3: Robustness to genome spacing of SNPs. Performance of BG2 and SMA when the number of causal SNPs decreases and the SNPs are clustered. SIM0: 20 evenly spaced causal SNPs. SIM1: 20 causal SNPs in four clusters. SIM2: 10 causal SNPs in two clusters. SIM3: 5 causal SNPs in one cluster.

3.7.6 Sensitivity of BG2 to parameter values

To study the sensitivity of BG2 to the values of parameters, we have added three new simulation studies.

The first of these simulation studies is SIM4 where, instead of 0.2 or -0.2, the regression coefficients for 10 causal SNPs are 0.4 and -0.4. Other 10 causal SNPs' coefficient β have six parameter settings: 0.2, 0.3, 0.4, 0.5, 0.7 and 1. Intercept $\beta_0 = -0.5$. Variance component for kinship random effects $\kappa = 0.15$. Figure 3.8 presents results for the SIM4 simulation study. Compared with Figure 3.1, Figure 3.8 shows that when all 20 causal SNPs' coefficient are equal to 0.4, BG2 can detect about 16 causal SNPs. Otherwise, BG2 can detect about 10 SNPs with relative large coefficient. In addition, BG2 has higher F1 score in Figure 3.8 than in Figure 3.1.

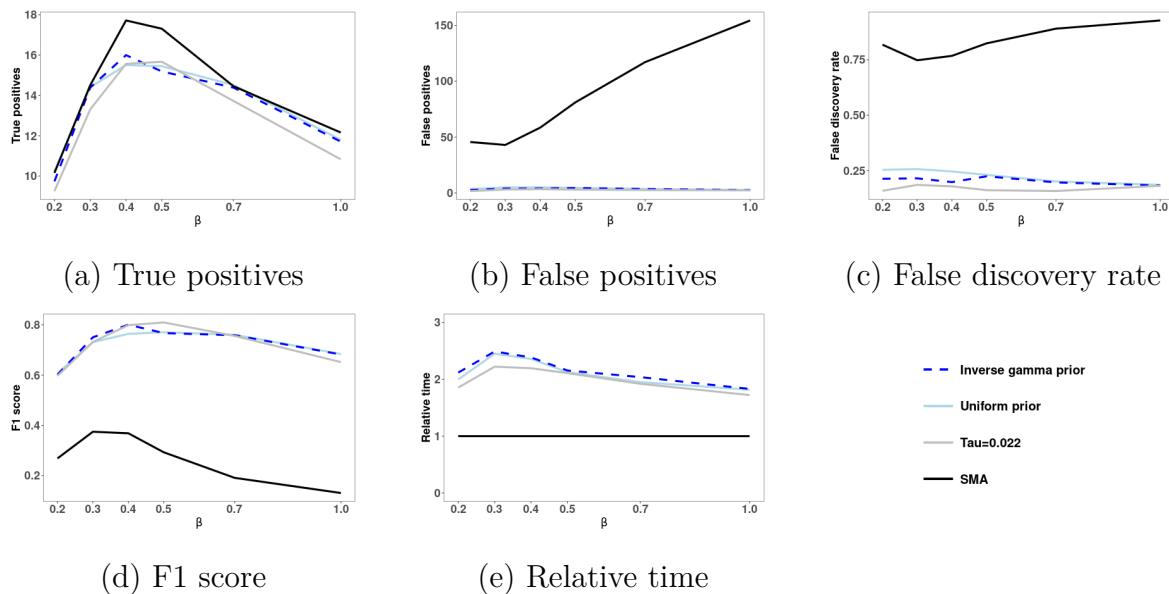


Figure 3.8: Results of simulation study SIM4. Performance of BG2 and SMA. Regression coefficients for 10 causal SNPs are 0.4 and -0.4. Another 10 causal SNPs' coefficient β have six parameter settings: 0.2, 0.3, 0.4, 0.5, 0.7 and 1. Intercept $\beta_0 = -0.5$. Variance component for kinship random effects $\kappa = 0.15$.

The second of these simulation studies is SIM5 where, instead of $\beta_0 = -0.5$, the intercept

is $\beta_0 = 1$. We have 20 causal SNPs. The regression coefficients for 10 causal SNPs are 0.2 and -0.2. Another 10 causal SNPs' coefficient β have six parameter settings: 0.2, 0.3, 0.4, 0.5, 0.7 and 1. Variance component for kinship random effects $\kappa = 0.15$. Figure 3.9 presents the results for SIM5. The results presented in Figure 3.9 look similar to those presented in Figure 3.1. Thus, BG2 does not seem to be sensitive to changes in β_0 .

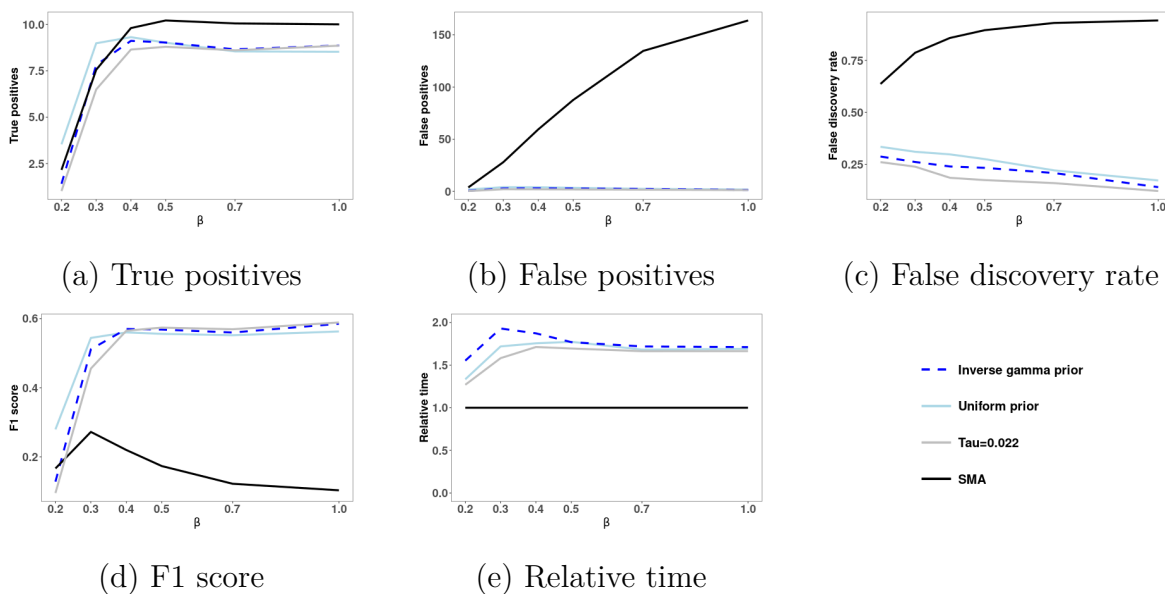


Figure 3.9: Results of simulation study SIM5. Performance of BG2 and SMA. Intercept $\beta_0 = 1$. Generate data from 20 causal SNPs. The regression coefficients for 10 causal SNPs are 0.2 and -0.2. Another 10 causal SNPs' coefficient β have six parameter settings: 0.2, 0.3, 0.4, 0.5, 0.7 and 1. Variance component for kinship random effects $\kappa = 0.15$.

The third of these simulation studies is SIM6 where, instead of $\kappa = 0.15$, the variance component of the kinship random effects is $\kappa = 0.3$. We have 20 causal SNPs. The regression coefficients for 10 causal SNPs are 0.2 and -0.2. Another 10 causal SNPs' coefficient β have six parameter settings: 0.2, 0.3, 0.4, 0.5, 0.7 and 1. Intercept $\beta_0 = -0.5$. Figure 3.10 presents the results for SIM6. The results in Figure 3.10 seem to be fairly similar to those in Figure 3.1. Thus, BG2 performs similarly when κ changes from 0.15 to 0.3.

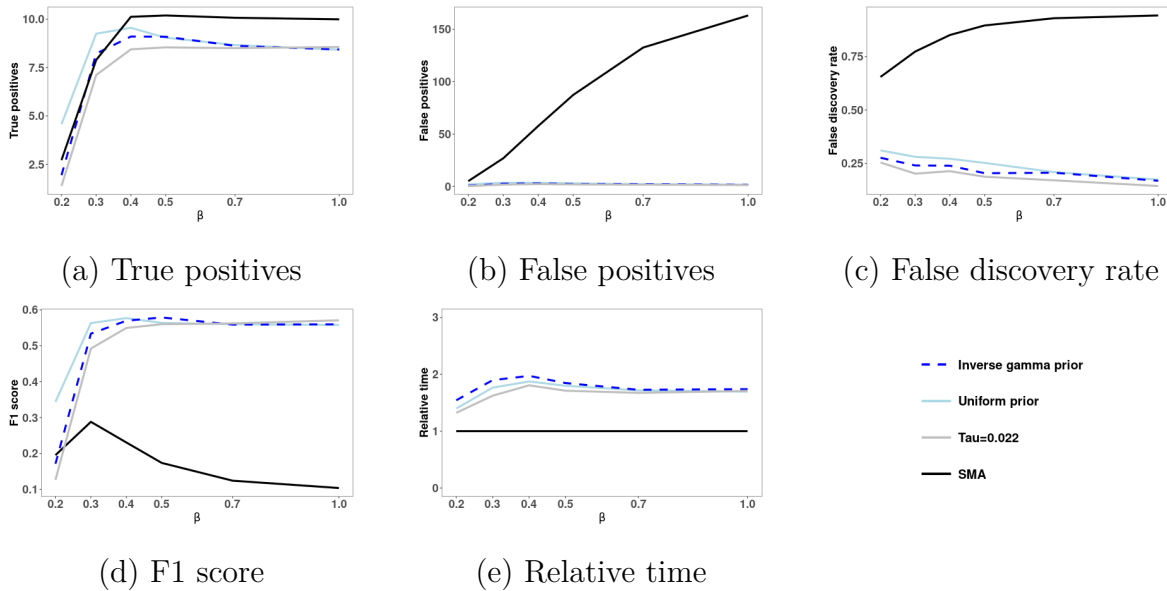


Figure 3.10: Results of simulation study SIM6. Performance of BG2 and SMA. Variance component for kinship random effects $\kappa = 0.3$. Generate data from 20 causal SNPs. The regression coefficients for 10 causal SNPs are 0.2 and -0.2. Another 10 causal SNPs' coefficient β have six parameter settings: 0.2, 0.3, 0.4, 0.5, 0.7 and 1. Intercept $\beta_0 = -0.5$.

3.7.7 Calibration of the pseudo-likelihood approach

The BG2 approach does not provide p-values. To check if the pseudo-likelihood approach is calibrated, we compute p-values based on the pseudo-likelihood approach for two datasets from Sections 3.4.1 and 3.4.2 that do not have any causal SNP. In this case, if the pseudo-likelihood approach is calibrated then the distribution of the p-values should be a uniform distribution. Figure 3.11 presents a Q-Q plot of p-values from one binary dataset in Section 3.4.1 whereas Figure 3.12 presents a Q-Q plot of p-values from one count dataset in Section 3.4.2. It is clear from the figures that in both cases the p-values have a uniform distribution. Therefore, the pseudo-likelihood approach is calibrated.

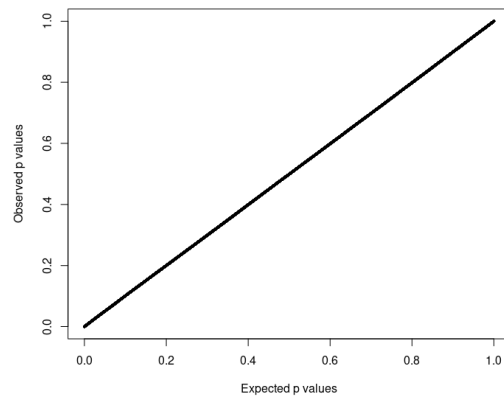


Figure 3.11: Calibration of the pseudo-likelihood approach. Binary data simulated from human genome data. Q-Q plot of p-values based the pseudo-likelihood approach.

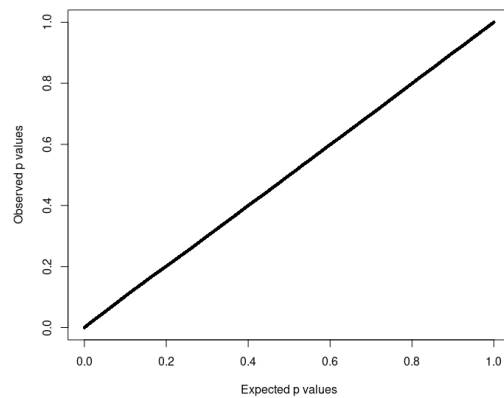


Figure 3.12: Calibration of the pseudo-likelihood approach. Count data simulated from *A. Thaliana* genome data. Q-Q plot of p-values based the pseudo-likelihood approach.

3.7.8 Histograms of the response variables in the case studies

Figures 3.13, 3.14, and 3.15 present the histograms of the response variables for each of the three case studies. From these figures, it is clear that the count response variables in the case studies presented in Sections 3.5.1 and 3.5.3 are skewed. However, Section 3.7.4 shows that BG2 can deal with imbalanced data without difficulties, and that the performance of BG2 improves as the level of skewness increases. In addition, the binary response variable from the case study presented in Section 3.5.2 is imbalanced. However, Section 3.7.4 shows that BG2 is robust to imbalanced data, and even performs slightly better when the data are imbalanced.

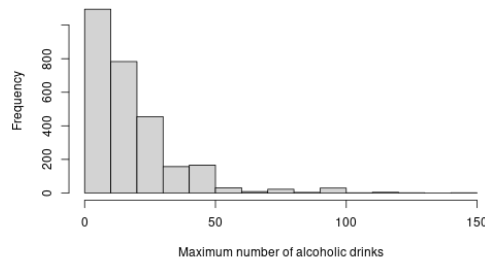


Figure 3.13: Case study from Section 3.5.1: Maximum number of alcoholic drinks. Histogram of the maximum number of alcoholic drinks.

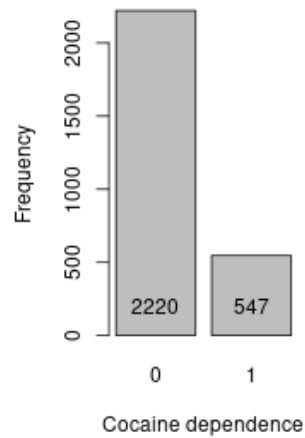


Figure 3.14: Case study from Section 3.5.1: Cocaine dependence. Histogram of the response variable cocaine dependence that is equal to 1 if the subject is cocaine dependent and 0 otherwise.

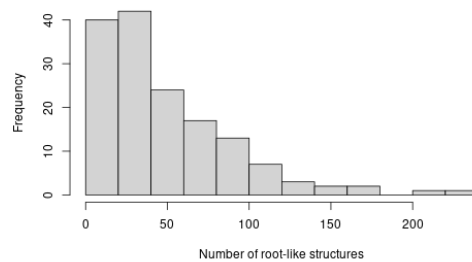


Figure 3.15: Case study from Section 3.5.3: Root-like structures in *A. Thaliana*. Histogram of number of root-like structures in *A. Thaliana*.

Chapter 4

IBG3: Iterative Bayesian fine-mapping for GLMMs and non-Gaussian GWAS data

4.1 Introduction

This chapter is based on the following manuscript that is in preparation for submitting to Nature Genetics: Shuangshuang Xu, Jacob Williams, Marco A.R. Ferreira, and Allison Tegge. IBG3: Iterative Bayesian fine-mapping for GLMMs and non-Gaussian GWAS data. Nature Genetics, In Preparation, 202X.

Genome-wide association studies (GWAS) generate useful data that provide genetic insight for many domains, such as diseases [8, 35, 38], human height [80], and salt tolerance in plants [82]. However, GWAS faces a challenge: the number of genetic variables, which often run into the millions, considerably outnumber the sample size. The predominant method employed in GWAS is Single Marker Analysis (SMA), which focuses on the separate analysis of one single nucleotide polymorphism (SNP) at a time. However, due to linkage disequilibrium (LD) and cryptic population structure, SNPs are highly correlated. As a result, SMA tends to have very high false discovery rate (FDR). In addition, SMA methods

have difficulty identifying SNPs with smaller effect sizes. Fine-mapping method is a post-GWAS method, which can solve LD problem and detect the most determinant SNPs from an associated genomic region found by GWAS methods [60]. In this paper, we propose a genome-wide fine-mapping method, combining screening among the whole genome and fine-mapping based on screening results. Furthermore, we iterate screening and fine-mapping steps to identify more SNPs with smaller effect sizes.

We propose iterative Bayesian GLMMs for GWAS with Zellner g prior (IBG3). IBG3 iterates two steps: a screening step and a fine-mapping step. IBG3 is initialized with a baseline model with no SNPs. Then, IBG3's screening step fits as many GLMMs as the number of possible SNPs, where each model includes the baseline model and one SNP. These model fits yield screening posterior probabilities for each SNP. After that, IBG3's screening step then uses Bayesian FDR control [46, 47] to choose a list of candidate SNPs. Next, the fine-mapping step performs a search through model space that includes all possible GLMMs obtained by combinations of the SNPs that are either in the list of candidate SNPs or in the baseline model. Then, the model with largest posterior probability found in this model search is declared the best model. If the best model is different from the current baseline model, then the baseline model is updated to be the same as the best model and another iteration of IBG3 starts with a new screening step. On another hand, if the best model is the same as the current baseline model, then IBG3 has converged. In that case, IBG3 reports the SNPs in the best model as the identified SNPs.

A useful consequence of the iterative nature of IBG3 is the ability to detect SNPs with smaller effect sizes. This contrasts with SMA methods and post-GWAS fine-mapping methods that are only able to detect SNPs with large effect sizes. As SMA considers only one SNP at one time, when fitting a model for a causal SNP the error term includes not only random variation but also the variation due to all the other causal SNPs. This causes SMA to have

diminished statistical power to detect SNPs with smaller effect sizes. Fine-mapping methods are applied to candidate SNPs found by SMA methods. Thus, fine-mapping methods also cannot detect the SNPs with small effect sizes which are not detected by SMA methods. In contrast, IBG3's first iteration finds SNPs with large effect sizes. These SNPs with large effect sizes are then included in the baseline model that is used in IBG3's second iteration. In this iteration, IBG3's screening step is performed conditional on the baseline model, which removes the variation due to the large effect size SNPs from the estimation of the variance of the error term thereby increasing statistical power to detect SNPs of smaller effect sizes. We note that two published Bayesian GWAS iterative procedures, GWASinlps [58] and BICOSS [71], focus on Gaussian data. To analyze non Gaussian GWAS data, we previously introduced BG2 [78], a Bayesian SNP detection method with the two steps of screening and fine-mapping but without iterations. As we show in [78], compared to SMA, BG2 yields much lower FDR. However, because BG2 does not iterate these two steps, BG2's recall of true causal SNPs is a bit lower. In contrast, a simulation study in Section 4 shows that, when compared to SMA methods and fine-mapping methods, IBG3 not only yields much lower FDR but also yields higher recall of true causal SNPs.

IBG3 uses a Bayesian hierarchical model and an empirical Bayes approach to estimate the hyperparameters of the prior distribution of the regression coefficients of GLMMs. IBG3 then uses Bayes Theorem to combine this prior distribution with the data to compute the posterior probabilities of the competing GLMMs [76, 78]. IBG3 uses a similar Bayesian ultra-high dimensional variable selection framework (p two orders of magnitude larger than n) for GLMMs applied to GWAS analysis as BG2. The main differences between IBG3 and BG2 are that BG2 is not iterative and uses a nonlocal prior whereas IBG3 is iterative and uses a Zellner g prior. We have performed a simulation study that compares a nonlocal prior [26], a unit information prior [32], Zellner's g prior, and Zellner-Siow prior [83] in an

iterative procedure. When considering statistical performance and computational time, the simulation study favors the Zellner g prior in an iterative procedure as implemented in IBG3. The remainder of this paper is organized as follows. Section 4.2 describes the GLMMs that we consider for non-Gaussian GWAS data. Section 4.3 describes the IBG3 method, including approximation methods for fast computations, as well as details for the screening step and the fine-mapping step. Section 4.4 presents results of two simulation studies: one simulation study that compared the performance of four different priors for the implementation of IBG3, and a simulation study that compares IBG3 with SMA methods (GLMM-based SMA method and GEMMA [88]) and fine-mapping methods (SuSiE-RSS [89] and BG2). Section 4.5 illustrates the application of IBG3 with applications to two case studies: alcohol consumption and breast cancer. Section 4.6 concludes with a discussion and future directions.

4.2 GLMMs

We consider generalized linear mixed models for non-Gaussian GWAS data. Here is the general model:

$$g(E(\mathbf{y}|\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)) = X_S\boldsymbol{\beta}_S + X_c\boldsymbol{\beta}_c + Z_1\boldsymbol{\alpha}_1 + Z_2\boldsymbol{\alpha}_2, \quad (4.1)$$

where \mathbf{y} is the vector of n observed phenotypes, X_S is the matrix of candidate SNPs, and X_c is the matrix of other control covariates (e.g., age, gender, and environmental factors). In addition, $\boldsymbol{\alpha}_1$ is the vector of kinship random effects, which follows a multivariate normal distribution $N(\mathbf{0}, \kappa_1\Sigma_1)$, where κ_1 is an unknown scalar and Σ_1 is a kinship matrix. Furthermore, if phenotypes are count data and follow Poisson distribution, we have $\boldsymbol{\alpha}_2$ in the model, which is a vector of overdispersion random effects following $N(\mathbf{0}, \kappa_2I)$. Conditional expecta-

tion of phenotypes $E(\mathbf{y}|\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ is linked to the linear predictor $X_S\boldsymbol{\beta}_S + X_c\boldsymbol{\beta}_c + Z_1\boldsymbol{\alpha}_1 + Z_2\boldsymbol{\alpha}_2$ by the link function $g(\cdot)$.

4.3 Iterative model selection

We propose a Bayesian iterative two-step fine-mapping method for non-Gaussian GWAS data. The first step is screening step. In the screening step, we have as many models as the number of SNPs. Each model has control covariates and only one SNP in it. We calculate the posterior probabilities for each model with the SNP, and use Bayesian false discovery rate control to select a set of candidate SNPs. In the fine-mapping step, we have all combinations of candidate SNPs from the screening step as candidate models. The model with the highest posterior probability is the best model. In the next iteration, we use the best model from the last iteration's fine-mapping step as the base model and repeat screening and fine-mapping step. We end the algorithm when there is no new SNP been selected in the screening step or the best models are identical in two consecutive iterations. The following sections are organized as: Section 4.3.1 presents the pseudo-likelihood approach and population parameters previously determined (P3D) approach, Section 4.3.2 introduces the screening step, and Section 4.3.3 presents the fine-mapping step.

4.3.1 Pseudo-likelihood method and Population Parameters Previously Determined approach

To calculate posterior probabilities for models in both the screening and the fine-mapping steps, we need the priors for the model space, the priors for parameters, and the integrated likelihood for parameters. However, integrated likelihood function $L(\boldsymbol{\beta}_S, \boldsymbol{\beta}_c, \kappa_1, \kappa_2|\mathbf{y})$ cannot

be solved analytically in the GLMMs. We use pseudo-likelihood approach to solve this problem.

Pseudo-likelihood method is an iterative procedure to estimate parameters $\boldsymbol{\beta}_S$, $\boldsymbol{\beta}_c$, κ_1 , and κ_2 , and random effects $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$. Meanwhile, pseudo-likelihood method provides the estimate of mean $\hat{\boldsymbol{\mu}}$ and covariance matrix \hat{V} for \mathbf{y} , and a vector of adjusted observations $\mathbf{y}^* = \hat{V}^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}) + X_S \hat{\boldsymbol{\beta}}_S + X_c \hat{\boldsymbol{\beta}}_c + Z_1 \hat{\boldsymbol{\alpha}}_1 + Z_2 \hat{\boldsymbol{\alpha}}_2$. The pseudo-likelihood algorithm iterates until convergence of these estimates. The adjusted observations \mathbf{y}^* can be modeled by LMM $\mathbf{y}^* = X_S \boldsymbol{\beta}_S + X_c \boldsymbol{\beta}_c + Z_1 \boldsymbol{\alpha}_1 + Z_2 \boldsymbol{\alpha}_2 + \hat{V}^{-1} \boldsymbol{\epsilon}$. In the LMMs, the integrated likelihood function $L(\boldsymbol{\beta}_S, \boldsymbol{\beta}_c, \kappa_1, \kappa_2 | \mathbf{y})$ has the closed form.

However, we have as many models as the number of SNPs in the screening, which has the magnitude of hundred thousands or millions, and models with all combinations of candidate SNPs in the fine-mapping step, which has the magnitude of hundreds or thousands. If we apply pseudo-likelihood methods to all GLMMs in both the screening step and the fine-mapping step, it is time-consuming. To speed up our method, we use population parameters previously determined approach (P3D). We apply pseudo-likelihood method to the base model once in the screening step and the fine-mapping step. From the base model, we obtain the adjusted observations and estimate of variance components for random effects, and assume they are the same for every model in the screening step or in the fine-mapping step. Here is the general form of base model:

$$g(E(\mathbf{y} | \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)) = X_B \boldsymbol{\beta}_B + Z_1 \boldsymbol{\alpha}_1 + Z_2 \boldsymbol{\alpha}_2, \quad (4.2)$$

where X_B contains covariates in the base model. For the screening step in the first iteration, the base model does not have any SNP in it. Thus, X_B only contains control covariates and intercept. For the fine-mapping step in the first iteration, the base model has all candidate

SNPs from the screening step. From the second iteration, X_B contains the SNPs identified in the last iteration additionally.

4.3.2 Screening step

In the screening step, we use the base model without any SNP to obtain the adjusted observations \mathbf{y}^* and estimate of $\boldsymbol{\beta}_c$, κ_1 , and κ_2 . We have as many models as the number of SNPs. In addition, we assume models with only one SNP have the same adjusted observations \mathbf{y}^* , estimate of coefficients of control covariates $\boldsymbol{\beta}_c$, and estimate of covariance matrix \hat{H} , where $\hat{H} = \hat{\kappa}_1 \Sigma + \hat{\kappa}_2 I + \hat{V}^{-1}$. Let \mathbf{x}_s be the vector of covariate SNP s , and β_s be the coefficient of SNP s . The general model with one SNP is:

$$\mathbf{y}^* = X_c \hat{\boldsymbol{\beta}}_c + \mathbf{x}_s \beta_s + \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 + \hat{V}^{-1} \boldsymbol{\epsilon} \quad (4.3)$$

The adjusted observations \mathbf{y}^* can be approximately modeled by a multivariate Gaussian distribution $N(X_c \hat{\boldsymbol{\beta}}_c + \mathbf{x}_s \beta_s, \hat{H})$.

Consider the spectral decomposition $\hat{H} = P D P^\top$. Let $\tilde{\mathbf{y}} = P^\top (\mathbf{y}^* - X_c \hat{\boldsymbol{\beta}}_c)$ and $\tilde{\mathbf{x}}_s = P^\top \mathbf{x}_s$. Then, an estimator of β_s is $\hat{\beta}_s = (\tilde{\mathbf{x}}_s^\top D^{-1} \tilde{\mathbf{x}}_s)^{-1} \tilde{\mathbf{x}}_s^\top D^{-1} \tilde{\mathbf{y}}$, which has an approximate distribution $N(\beta_s, \sigma_s^2)$, where $\sigma_s^2 = \text{var}(\hat{\beta}_s) = (\tilde{\mathbf{x}}_s^\top D^{-1} \tilde{\mathbf{x}}_s)^{-1}$.

Let π_0 be the probability that one SNP is not in the model. We use a mixture prior for β_s , that is

$$p(\beta_s | \pi_0, \Theta) = \pi_0 \delta(\beta_s = 0) + (1 - \pi_0) \pi(\beta_s | \Theta), \quad (4.4)$$

where $\pi(\beta_s | \Theta)$ is the prior for β_s when SNP s is in the model. We compare four priors in this paper: nonlocal prior, unit information prior, Zellner g prior, and Zellner-Siow prior. The

details of priors $\pi(\beta_s|\Theta)$ are in the Supplementary material. Then, the predictive density of $\hat{\beta}_s$ is

$$p(\hat{\beta}_s|\tau, \pi_0) = \int p(\hat{\beta}_s|\beta_s)p(\beta_s|\tau, \pi_0) d\beta_s. \quad (4.5)$$

To estimate parameters π_0 and Θ in the Equation (4.4), we obtain the likelihood function of π_0 and Θ , that is

$$L(\hat{\beta}_1, \dots, \hat{\beta}_p|\pi_0, \Theta) = \prod_{s=1}^p p(\hat{\beta}_s|\pi_0, \Theta). \quad (4.6)$$

Let $\pi(\pi_0)$ and $\pi(\Theta)$ be the prior densities of π_0 and Θ , respectively. We use flat prior for $\pi(\pi_0)$. As for $\pi(\Theta)$, we assign different priors for $\pi(\Theta)$ when we use different prior $\pi(\beta_s|\Theta)$ for β_s . More details about $\pi(\Theta)$ are in the Supplementary material.

Thus, the posterior distribution of π_0 and Θ is

$$\pi(\pi_0, \Theta|\hat{\beta}_1, \dots, \hat{\beta}_p) \propto \pi(\pi_0)\pi(\Theta)L(\hat{\beta}_1, \dots, \hat{\beta}_p|\pi_0, \Theta). \quad (4.7)$$

We use the posterior mode $\hat{\pi}_0$ and $\hat{\Theta}$ as the estimate of π_0 and Θ . After estimating π_0 and Θ , we calculate the posterior probability for each regressor SNP s in the screening step, that is

$$P(\beta_s \neq 0|\hat{\beta}_s, \hat{\pi}_0, \hat{\Theta}) = 1 - \frac{\hat{\pi}_0 N(\hat{\beta}_s|0, \sigma_s^2)}{p(\hat{\beta}_s|\hat{\pi}_0, \hat{\Theta})}. \quad (4.8)$$

Then, we apply Bayesian false discovery rate (FDR) control to the posterior probabilities of all SNPs with nominal FDR at 5% and select a list of candidate SNPs.

4.3.3 Fine-mapping step

In the fine-mapping step, we consider all combinations of candidate SNPs from the screening step as the candidate models. Assume we obtain k candidate SNPs from the screening step. The fine-mapping consider $S = 2^k$ possible models. Let M_m be the m^{th} model with p_m SNPs, $m = 1, \dots, S$.

As mentioned in the Section 4.3.1, we use base model to obtain the adjusted observations \mathbf{y}^* and estimate of β_c , κ_1 and κ_2 . In the model selecting step, the base model is the full model with all k candidate SNPs. Then, we assume all possible models have the same adjusted observations. Let X_m be the matrix of SNPs in the model M_m , and β_m be the corresponding vector of regression coefficients. The general model is:

$$\mathbf{y}^* = X_c \hat{\beta}_c + X_m \beta_m + \alpha_1 + \alpha_2 + \hat{V}^{-1} \epsilon \quad (4.9)$$

The adjusted observations \mathbf{y}^* can be approximately modeled by a multivariate Gaussian distribution $N(X_c \hat{\beta}_c + X_m \beta_m, \hat{H})$, where $\hat{H} = \hat{\kappa}_1 \Sigma + \hat{\kappa}_2 I + \hat{V}^{-1}$. Consider the spectral decomposition $\hat{H} = P D P^\top$. Let $\tilde{\mathbf{y}} = P^\top (\mathbf{y}^* - X_c \hat{\beta}_c)$ and $\tilde{X}_m = P^\top X_m$. Then, $\tilde{\mathbf{y}}$ conditional on β_m approximately follows distribution $N(\tilde{X}_m \beta_m, D)$. We consider four kinds of prior distribution $\pi(\beta_m | M_m)$ for β_m conditional on model M_m . Then, by integrating out β_m we can obtain the marginal density $m(\tilde{\mathbf{y}} | M_m)$ of $\tilde{\mathbf{y}}$. Details about priors for β_m and marginal density $m(\tilde{\mathbf{y}} | M_m)$ are in the supplementary material. The prior probability for model M_m is $P(M_m) = \hat{\pi}_0^{k-p_m} (1 - \hat{\pi}_0)^{p_m}$, where $\hat{\pi}_0$ is the estimate from the screening step in the first iteration. Then, the posterior probability of model M_m is

$$P(M_m | \tilde{\mathbf{y}}) \propto P(M_m) m(\tilde{\mathbf{y}} | M_m). \quad (4.10)$$

We select SNPs in the model with the highest posterior probability.

4.4 Simulation Studies

In this section, we conduct simulation studies to evaluate the performance of our proposed method, IBG3, when implemented using various priors, in comparison to the SMA methods (GLMM-based SMA method and GEMMA) and fine-mapping methods (SuSiE-RSS and BG2). To simulate challenges commonly encountered in Genome-Wide Association Studies (GWAS), we employ genotype data identical to those used in our case studies to generate phenotype data for the simulations. Our assessment relies on four key criteria: True Positives (TP), False Positives (FP), False Discovery Rate (FDR), and the F1 score. Additionally, we consider computational time as an important factor. All criteria are calculated as the mean values across 100 simulated datasets.

We simulate GWAS phenotype data using genotype information from the Study of Addiction: Genetics and Environment (SAGE) which is part of the National Human Genome Research Institute’s Gene Environment Association Study Initiative [Database for Genotypes and Phenotypes (dbGaP) study accession phs000092.v1.p1]. There are 2,772 European Americans and 800,000 SNPs with minor allele frequency (MAF) larger than 0.01. From these 800,000 SNPs, we selected 20 evenly spaced SNPs to be the causal SNPs, where 5 SNPs have relatively large coefficients β_l and 15 SNPs have small coefficients β_s . We have four parameter settings: (a) $\beta_l = 1.2$, $\beta_s = 0.3$; (b) $\beta_l = 1.6$, $\beta_s = 0.4$; (c) $\beta_l = 2$, $\beta_s = 0.5$; (d) $\beta_l = 2.4$, $\beta_s = 0.6$. We set the intercept $\beta_0 = -0.5$, and the variance component κ of the kinship random effects $\boldsymbol{\alpha}$ is equal to 0.15. Thus, the phenotype data are simulated from a Bernoulli

GLMM, that is

$$\text{logit}(p_i) = \beta_0 + \sum_{j=1}^5 \beta_l x_{ij} + \sum_{j=6}^{20} \beta_s x_{ij} + \alpha_i, \quad (4.11)$$

$$\boldsymbol{\alpha} \sim N(\mathbf{0}, \kappa \boldsymbol{\Sigma}). \quad (4.12)$$

where p_i is the expectation of y_i , and $\boldsymbol{\Sigma}$ is the kinship matrix.

GLMM-based SMA method only takes 2 minutes, which is much faster than GEMMA. However, the performance of GLMM-based SMA method is a little bit worse than GEMMA. Thus, we do not include GLMM-based SMA method in the table. From the Table 4.1, we can see that IBG3 methods can detect all SNPs with large coefficients and part of SNPs with small coefficients. When coefficients are larger, the number of true positives increases. Overall, IBG3 with Zellner g prior and nonlocal prior always detect the most SNPs correctly. IBG3 with Zellner-Siow prior can detect more SNPs correctly when coefficients increase. GEMMA and SuSiE-RSS cannot detect the SNPs with smaller coefficients. BG2 can detect one or two SNPs with small coefficient on average. Table 4.2 and 4.3 show that IBG3 and BG2 methods have very low false positives and false discovery rate, compared to GEMMA and SuSiE-RSS. GEMMA has more than 100 false positives. Though SuSiE-RSS reduces some false positives, there are still more than 30 false positives. In addition, SuSiE-RSS, as a post-GWAS method, detects less true positives than GEMMA. IBG3 methods and BG2 have only 3 false positives at most. Table 4.4 shows that according to F1 score, IBG3 with Zellner g prior has the best performance detecting SNPs. IBG3 with nonlocal prior's F1 scores are a little bit lower than IBG3 with Zellner g prior. However, IBG3 with nonlocal prior takes more time than IBG3 with Zellner g prior, which is shown in Table 4.5. Though SuSiE-RSS only takes 3 minutes, it needs GEMMA to help do screening first. The total process of GEMMA and SuSiE-RSS needs more than 30 minutes. IBG3 is faster than GEMMA+SuSiE-RSS.

Methods	Set1	Set2	Set3	Set4
IBG3, Zellner's g prior	8.75	10.67	11.19	11.29
IBG3, Nonlocal prior	8.77	10.62	10.86	11.20
IBG3, Unit information prior	6.63	9.07	9.04	9.10
IBG3, Zellner-Siow prior	5.62	7.41	9.86	11.52
GEMMA	5.65	6.06	6.33	6.78
SuSiE-RSS	5.64	6.00	6.30	6.72
BG2	6.42	6.93	7.52	7.81

Table 4.1: True positives (TP) for IBG3, GEMMA, SuSiE-RSS, and BG2. Four parameter settings: (1) $\beta_l = 1.2$, $\beta_s = 0.3$; (2) $\beta_l = 1.6$, $\beta_s = 0.4$; (3) $\beta_l = 2$, $\beta_s = 0.5$; (4) $\beta_l = 2.4$, $\beta_s = 0.6$.

Methods	Set1	Set2	Set3	Set4
IBG3, Zellner's g prior	2.90	2.78	2.86	2.51
IBG3, Nonlocal prior	2.99	3.03	3.06	2.82
IBG3, Unit information prior	1.67	2.23	1.99	1.67
IBG3, Zellner-Siow prior	1.04	1.61	2.28	2.25
GEMMA	119.77	128.85	133.22	136.13
SuSiE-RSS	33.96	34.8	34.70	34.59
BG2	1.20	1.39	1.30	1.09

Table 4.2: False positives (FP) for IBG3, GEMMA, SuSiE-RSS, and BG2. Four parameter settings: (1) $\beta_l = 1.2$, $\beta_s = 0.3$; (2) $\beta_l = 1.6$, $\beta_s = 0.4$; (3) $\beta_l = 2$, $\beta_s = 0.5$; (4) $\beta_l = 2.4$, $\beta_s = 0.6$.

4.5 Case studies

To illustrate the applicability of the IBG3 methods to non-Gaussian GWAS data, we present two case studies in this section: alcohol consumption (count data), and breast cancer diagnosis (count data). In addition, we compare IBG3 methods with GEMMA, SuSiE-RSS, and BG2.

Methods	Set1	Set2	Set3	Set4
IBG3, Zellner’s g prior	0.249	0.207	0.203	0.182
IBG3, Nonlocal prior	0.254	0.222	0.220	0.201
IBG3, Unit information prior	0.201	0.197	0.180	0.155
IBG3, Zellner-Siow prior	0.156	0.178	0.188	0.163
GEMMA	0.955	0.955	0.955	0.953
SuSiE-RSS	0.858	0.853	0.846	0.837
BG2	0.157	0.167	0.147	0.122

Table 4.3: False discovery rate (FDR) for IBG3, GEMMA, SuSiE-RSS, and BG2. Four parameter settings: (1) $\beta_l = 1.2, \beta_s = 0.3$; (2) $\beta_l = 1.6, \beta_s = 0.4$; (3) $\beta_l = 2, \beta_s = 0.5$; (4) $\beta_l = 2.4, \beta_s = 0.6$.

Methods	Set1	Set2	Set3	Set4
IBG3, Zellner’s g prior	0.553	0.638	0.657	0.668
IBG3, Nonlocal prior	0.553	0.631	0.640	0.658
IBG3, Unit information prior	0.469	0.579	0.583	0.591
IBG3, Zellner-Siow prior	0.422	0.510	0.614	0.682
GEMMA	0.078	0.078	0.079	0.083
SuSiE-RSS	0.189	0.197	0.207	0.219
BG2	0.465	0.489	0.522	0.540

Table 4.4: F1 score for IBG3, GEMMA, SuSiE-RSS, and BG2. Four parameter settings: (1) $\beta_l = 1.2, \beta_s = 0.3$; (2) $\beta_l = 1.6, \beta_s = 0.4$; (3) $\beta_l = 2, \beta_s = 0.5$; (4) $\beta_l = 2.4, \beta_s = 0.6$.

4.5.1 Maximum number of alcoholic drinks

The Collaborative Study on the Genetics of Alcoholism (COGA) [5] was a large-scale family research project, primarily aiming to pinpoint genes linked to alcohol dependence. In this context, we conducted a Genome-Wide Association Study (GWAS) to analyze the maximum number of alcoholic beverages consumed within a 24-hour span. Our analysis encompasses data from 2,759 European Americans, evaluating 846,076 SNPs with a Minor Allele Frequency (MAF) larger than 0.01. For our analysis, we employed Poisson Generalized Linear Mixed Models (GLMM). Within the GLMMs, the 846,076 SNPs are treated as potential explanatory variables. Additionally, we incorporated a kinship random effects vector to accommodate the genetic structure across the 2,759 participants and another vector for

Methods	Set1	Set2	Set3	Set4
IBG3, Zellner's g prior	11.95	16.68	19.03	20.70
IBG3, Nonlocal prior	14.55	24.98	23.44	31.33
IBG3, Unit information prior	9.95	14.32	18.94	18.52
IBG3, Zellner-Siow prior	15.83	19.82	25.04	24.64
GEMMA	30.86	31.19	33.64	31.35
GEMMA + SuSiE-RSS	33.07	33.70	35.99	33.88
BG2	4.12	3.71	3.63	3.92

Table 4.5: Time (min) for IBG3, GEMMA, SuSiE-RSS, and BG2. Four parameter settings: (1) $\beta_l = 1.2$, $\beta_s = 0.3$; (2) $\beta_l = 1.6$, $\beta_s = 0.4$; (3) $\beta_l = 2$, $\beta_s = 0.5$; (4) $\beta_l = 2.4$, $\beta_s = 0.6$.

overdispersion random effects to account for any overdispersion present.

The IBG3 methods identify only one SNP residing within the coding region of the PTGER4 gene on chromosome 5. This finding was corroborated by the BG2 method, which isolated the identical SNP within PTGER4. The PTGER4 gene encodes a receptor for prostaglandin E2 (PGE2), a molecule implicated in the body's inflammatory response to alcohol intake. Notably, PGE2's involvement in inflammation is consistent with the observation that tolfenamic acid, a PGE2 inhibitor, significantly mitigates a range of hangover symptoms, as detailed in the work of [69]. Conversely, the GEMMA identifies a total of 9 SNPs, and SuSiE-RSS method identifies the same 9 SNPs as GEMMA. GLMM-based SMA method also found these 9 SNPs as GEMMA and one more SNP.

4.5.2 Breast cancer

GWAS has played an important role in identifying genetic variants associated with breast cancer. Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) was a project which is part of the NCI's Genetic Associations and Mechanisms in Oncology (GAME-ON) initiative (<http://epi.grants.cancer.gov/gameon/>), whose goal was to foster an intra-disciplinary and collaborative approach to the translation of promising research leads

deriving from the initial wave of cancer GWAS. The main objective of DRIVE is to determine genetic variants associated with breast cancer. DRIVE has 60015 breast cancer cases and controls, which is composed by 10 consent groups. In this case, we only focus on consent group 8, which is general research use (GRU) group. After doing quality control, GRU has 21653 individuals with 410854 SNPs with a MAF larger than 0.01, missing data rate less than 5%.

In this case, IBG3 methods with different priors have different findings. IBG3 method with Zellner's g prior identifies 16 SNPs. IBG3 method with nonlocal prior also identifies 16 SNPs. IBG3 method with Zellner-Siow prior identifies 4 SNPs, and IBG3 method with unit information prior identifies 6 SNPs. These SNPs are also identified by IBG3 method with Zellner's g prior and nonlocal prior. BG2 method finds 13 SNPs. GEMMA finds 64 SNPs. SuSiE-RSS finds 24 SNPs. GLMM-based SMA method finds 134 SNPs.

Based on the IBG3 methods' performance in the simulation study, we recommend IBG3 method with Zellner g prior. Thus, we look into the result of IBG3 with Zellner g prior here. IBG3 method with Zellner g prior identifies a total of 16 SNPs that are associated with breast cancer. These SNP includes rs1657220307, rs10995190, rs2981584, rs10829706, rs75296154, rs734148, rs7204722, rs4784227, rs9901120, rs62078752, rs1954098297, rs3827256, chr21_16563640_C_T, chr3_30684907_C_T, rs10941679, and chr5_56212595_C_T. Comparing these SNPs with previously reported breast cancer susceptibility loci that are identified from GWAS, there are 11 SNPs been reported by other papers. Out of 11 SNPs, there are 10 SNPs which are located in or close to gene SRGAP2C, ZNF365, FGFR2, LINC01488, LINC01234, MAP1LC3B, CASC16, FASN, TNS1, and PFKL respectively. These 10 genes and other 1 SNP were proven to affect breast cancer by themselves or interacting with other genes. SRGAP2C is overexpressed in pancreas and breast from HIPED (Health in prisons European database) [44]. ZNF365 is associated with the risk of breast cancer in a group of the Iranian popula-

tion [22]. Mutations on FGFR2 have been identified in both ER+ and ER-BCs [37]. SNP rs75296154 is close to gene LINC01488. There is a significant correlation between LINC01488 and CCND1 expression, and [17] demonstrated a strong correlation between CCND1 amplification and its protein expression in breast cancer. LINC01234 is negatively related to miR-190b, and miR-190b was reported to be down-regulated in breast cancer and was related to estrogen receptor [24]. MAP1LC3B and SQSTM1 co-expression plays an important role in breast invasive ductal carcinoma, which is a common type of breast cancer [40]. CASC16 is significantly related to breast cancer susceptibility in a Northwest Chinese female population [90]. FASN is overexpressed in breast from HIPED. [12] reviews a rationale for the EGFR-mediated pathways interacting with FASN, communion of these two biomarkers with breast cancer. TNS1 is regulated by MaTAR25 lncRNA to impact breast cancer progression [11]. [86] found that PFKL is one of glycolysis genes which compose a glycolysis signature that could predict the survival rate of patients with breast cancer. SNP rs10941679 is located at chromosome 5p12. SNP rs10941679 is associated with increased expression of FGF10 and MRPS30, which are two candidate genes for breast cancer pathogenesis [23]. In addition, we identify 5 novel associated loci that may affect breast cancer. They are rs10829706, rs9901120, chr21_16563640_C_T, chr3_30684907_C_T, and chr5_56212595_C_T.

The Figure 4.1 shows that in the screening step of the first iteration, IBG3 found 137 significant SNPs. Due to the high correlation, some SNPs are clustering in a region, such as a cluster in the chromosome 5. However, only 13 SNPs were included in the best model by IBG3 in the fine-mapping step. Conditional on this best model, some new SNPs were shown up in the screening step in the iteration 2. Besides 13 SNPs found in the iteration 1, the fine-mapping step in the iteration 2 found 2 more SNPs in the best model. In the end, when IBG3 did not find more new SNPs, there are 16 SNPs in the best model.

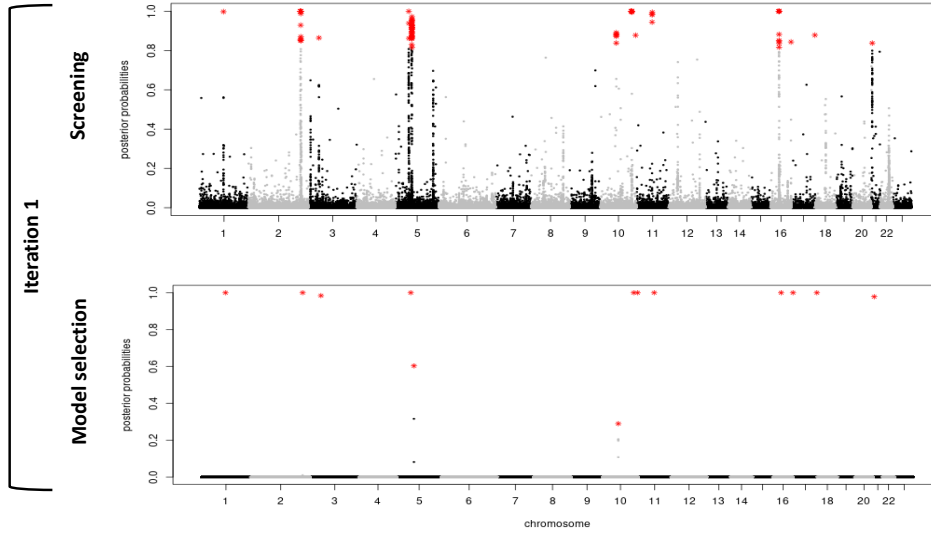


Figure 4.1: The posterior probabilities of all SNPs in Iteration 1 for breast cancer data. The red dots are the significant SNPs in the screening step under Bayesian false discovery control.

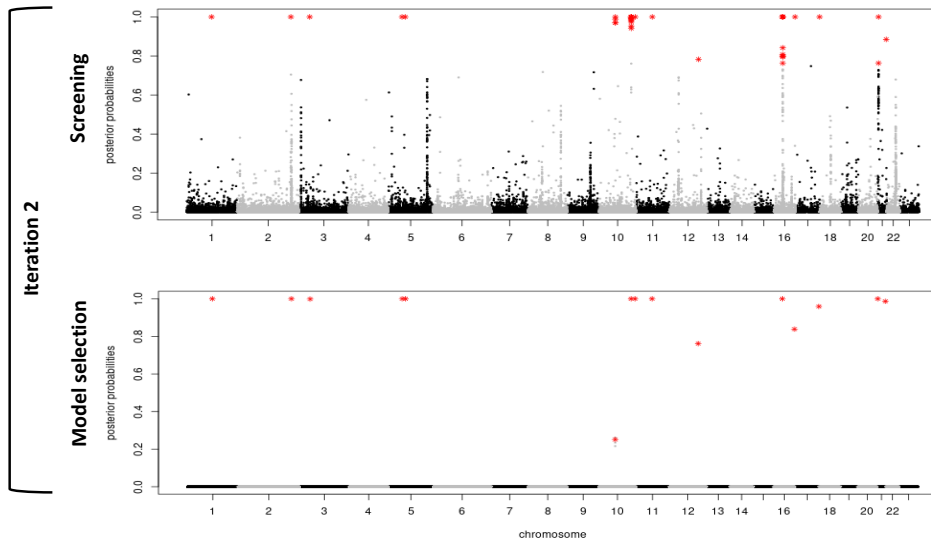


Figure 4.2: The posterior probabilities of all SNPs in Iteration 2 for breast cancer data. The red dots are the significant SNPs in the screening step under Bayesian false discovery control.

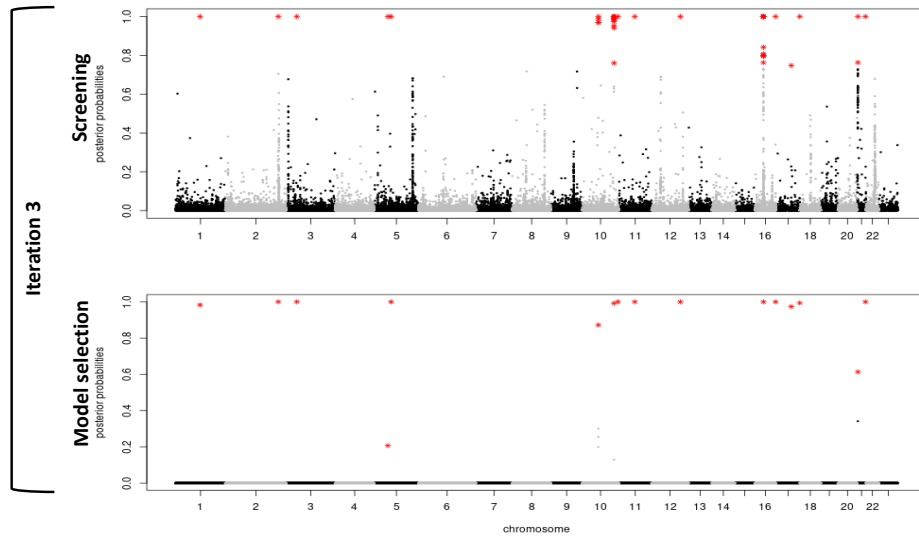


Figure 4.3: The posterior probabilities of all SNPs in Iteration 3 for breast cancer data. The red dots are the significant SNPs in the screening step under Bayesian false discovery control.

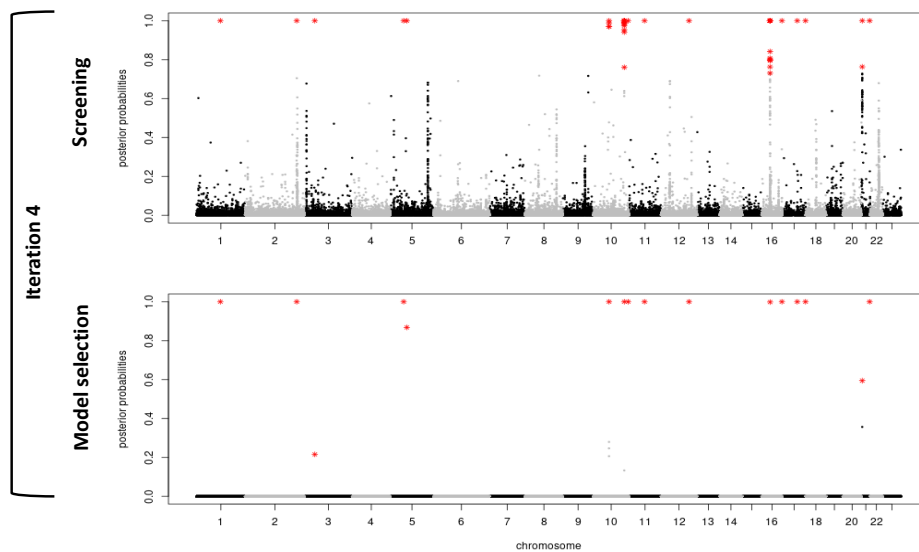


Figure 4.4: The posterior probabilities of all SNPs in Iteration 4 for breast cancer data. The red dots are the significant SNPs in the screening step under Bayesian false discovery control.

4.6 Discussion

We propose an iterative Bayesian variable selection method for GLMMs and non-Gaussian GWAS data (IBG3). IBG3 is a genome-wide fine-mapping method. The IBG3 framework operates through a iterative two-step process. Each iteration comprises a screening step followed by a fine-mapping step. During the screening step, IBG3 has as many models as the number of SNPs, identifies a subset of SNPs exhibiting the strongest association with the phenotype. These SNPs are candidate SNPs for the next step. Subsequently, IBG3 considers all combinations of candidate SNPs in the fine-mapping step. Within this step, IBG3 select the SNPs in the model which has the highest posterior probability. With each successive iteration, previously selected SNPs are fixed in each GLMM model. By Conditioning on the fixed SNPs, IBG3 do selection among the rest of SNPs. This iterative process continues until there is no new SNPs been selected in the new iteration.

IBG3 methods solve most GWAS methods' problem. The widely used method for GWAS is SMA, for example GEMMA. SMA has only one SNP in a model each time. In this way, SMA can solve $p \gg n$ problem in GWAS analysis, where p is the number of variables and n is the sample size. However, SMA ignore the interaction between SNPs, which affects diseases and traits in reality. IBG3's screening step is like SMA method to solve high-dimension problem. In addition, SMA has high false discovery rate problem because of LD. IBG3's fine-mapping step helps reduce false positives. Furthermore, IBG3's fine-mapping step takes care of the interaction effect among SNPs. In addition, IBG3 uses iteration procedure to detect SNPs with small effect size, which are ignored by SMA and BG2 since SNPs with small effect size are not significant if the model considers SNPs with large effect size are not true covariates and are merged in the error term. IBG3 takes more steps and iterations than SMA to find more SNPs. We use P3D method to approximate computation and reduce time to a doable range.

IBG3 methods can detect more true positives than SMA methods and non-iterative fine-mapping methods, and much less false discovery rate than GEMMA and SuSiE-RSS, which is shown by our simulation study. The simulation study is based on real-world genotype data to mimic GWAS analysis’s difficulties. IBG3 methods and BG2 method have much less false positives than GEMMA and SuSiE-RSS. IBG3 methods and BG2 method have less than 3 false positives. However, GEMMA has more than 100 false positives, and SuSiE-RSS has more than 30 false positives. IBG3 with Zellner’s g prior and nonlocal prior always identify more SNPs correctly than GEMMA, SuSiE-RSS, and BG2. IBG3 with Zellner’s g prior takes less time than IBG3 with nonlocal prior. We recommend to use IBG3 with Zellner’s g prior for GWAS data analysis.

In the future, we can improve computation of IBG3 further. The IBG3’s computation timing is highly related to the sample size. Now, IBG3 can handle $10^3 \sim 10^4$ magnitude sample size. We can develop IBG3 to deal with larger sample size, such as biobank-scale data.

4.7 Supplementary Material

4.7.1 Priors for IBG3 method

Screening step

In this section, we provide the details of priors $\pi(\beta_s|\Theta)$ and $\pi(\Theta)$. In the screening step, we assume the estimator $\hat{\beta}_s$ has an approximate normal distribution:

$$\hat{\beta}_s|\beta_s \sim N(\beta_s, (\mathbf{x}_s^\top D^{-1} \mathbf{x}_s)^{-1}). \quad (4.13)$$

Let $\sigma_s^2 = (\mathbf{x}_s^\top D^{-1} \mathbf{x}_s)^{-1}$. The prior $\pi(\beta_s | \Theta, \pi_0)$ is defined as mixture of a Dirac delta prior and $\pi(\beta_s | \Theta)$, where Θ is the general parameter sets for $\pi(\beta_s | \Theta)$. We have four types of priors for $\pi(\beta_s | \Theta)$, Non local priro, unit information prior, Zellner's g prior, and Zellner-Siow prior.

The first prior we propose in our iterative algorithm is the nonlocal prior. The prior $\pi(\beta_s | \Theta)$ is

$$\pi(\beta_s | \tau) = \frac{\beta_s^2}{n\tau\sigma_s^2} N(\beta_s | 0, n\tau\sigma_s^2). \quad (4.14)$$

The prior for τ is uniform prior.

The second prior is unit information prior, that is

$$\pi(\beta_s) = N(\beta_s | 0, n\sigma_s^2). \quad (4.15)$$

The third prior is Zellner's g prior, that is

$$\pi(\beta_s | g) = N(\beta_s | 0, g\sigma_s^2), \quad (4.16)$$

where g follows a uniform prior.

Tha fourth prior is Zellner-Siow prior, that is

$$\pi(\beta_s) = \int_g \pi(g) N(\beta_s | 0, g\sigma_s^2) dg, \quad (4.17)$$

where $\pi(g)$ follows inverse gamma distribution with shape 0.5 and scale $n/2$. n is the sample size.

Model selection step

In the model selection step, we consider all combinations of candidate SNPs from the screening step. Assume there are S candidate models. Let M_m be the m^{th} model with covariate matrix X_m , and there are p_m covariates. $\boldsymbol{\beta}_m$ is the corresponding vector of regression coefficients.

In this section, we provide four priors for $\boldsymbol{\beta}_m$ and marginal density $m(\tilde{\mathbf{y}}|M_m)$. The first prior is nonlocal prior. The prior for $\boldsymbol{\beta}_m$ is

$$p(\boldsymbol{\beta}_m|M_m) = d_m(2\pi)^{-p_m/2}(\hat{\tau}n)^{-3p_m/2} \left| \tilde{X}_m^\top D^{-1} \tilde{X}_m \right|^{3/2} \exp \left[-\frac{1}{2\hat{\tau}n} \boldsymbol{\beta}_m^\top \tilde{X}_m^\top D^{-1} \tilde{X}_m \boldsymbol{\beta}_m \right] \prod_{i=1}^{p_m} \beta_{mi}^2.$$

where d_m is a normalizing constant. $\hat{\tau}$ is estimate from screening step. Let $C_m = \tilde{X}_m^\top D^{-1} \tilde{X}_m (1 + (\hat{\tau}n)^{-1})$, $\hat{\boldsymbol{\beta}}_m = C_m^{-1} \tilde{X}_m^\top D^{-1} \tilde{\mathbf{y}}$, and $R_m = \tilde{\mathbf{y}}^\top D^{-1} \tilde{\mathbf{y}} - \tilde{\mathbf{y}}^\top D^{-1} \tilde{X}_m \hat{\boldsymbol{\beta}}_m$. The marginal density is:

$$m(\tilde{\mathbf{y}}|M_m) = (2\pi)^{-n/2} |D|^{-1/2} (1 + \hat{\tau}n)^{-p_m/2} \exp \left(-\frac{R_m}{2} \right) \frac{E_2(\prod_{i=1}^{p_m} \beta_{mi}^2)}{E_1(\prod_{i=1}^{p_m} \beta_{mi}^2)}.$$

$E_1(\prod_{i=1}^{p_m} \beta_{mi}^2)$ is the expectation of $\prod_{i=1}^{p_m} \beta_{mi}^2$ with respect to $N(\mathbf{0}, (1 + \hat{\tau}n)C_m^{-1})$. $E_2(\prod_{i=1}^{p_m} \beta_{mi}^2)$ is the expectation of $\prod_{i=1}^{p_m} \beta_{mi}^2$ with respect to $N(\hat{\boldsymbol{\beta}}_m, C_m^{-1})$.

The second prior is unit information prior, that is

$$\pi(\boldsymbol{\beta}_m|M_m) \sim N(\boldsymbol{\beta}_m|\mathbf{0}, n(\tilde{X}_m^\top D^{-1} \tilde{X}_m)^{-1}). \quad (4.18)$$

The marginal density is:

$$\begin{aligned}
 m(\tilde{\mathbf{y}}|M_m) &= \int N(\tilde{\mathbf{y}}|\tilde{X}_m\boldsymbol{\beta}_m, D)N(\boldsymbol{\beta}_m|\mathbf{0}, n(\tilde{X}_m^\top D^{-1}\tilde{X}_m)^{-1}) d\boldsymbol{\beta}_m \\
 &= (2\pi)^{-n/2}(n+1)^{-p_m/2}|D|^{-1/2} \\
 &\quad \exp\left[-\frac{1}{2}\tilde{\mathbf{y}}^\top D^{-1}\tilde{\mathbf{y}} + \frac{1}{2}\tilde{\mathbf{y}}^\top D^{-1}\tilde{X}_m\left(\frac{n+1}{n}\tilde{X}_m^\top D^{-1}\tilde{X}_m\right)^{-1}\tilde{X}_m^\top D^{-1}\tilde{\mathbf{y}}\right].
 \end{aligned} \tag{4.19}$$

The third prior is Zellner's g prior, that is

$$\pi(\boldsymbol{\beta}_m|M_m, \hat{g}) \sim N(\boldsymbol{\beta}_m|\mathbf{0}, \hat{g}(\tilde{X}_m^\top D^{-1}\tilde{X}_m)^{-1}). \tag{4.20}$$

The marginal density is:

$$\begin{aligned}
 m(\tilde{\mathbf{y}}|M_m, \hat{g}) &= \int N(\tilde{\mathbf{y}}|\tilde{X}_m\boldsymbol{\beta}_m, D)N(\boldsymbol{\beta}_m|\mathbf{0}, \hat{g}(\tilde{X}_m^\top D^{-1}\tilde{X}_m)^{-1}) d\boldsymbol{\beta}_m \\
 &= (2\pi)^{-n/2}(\hat{g}+1)^{-p_m/2}|D|^{-1/2} \\
 &\quad \exp\left[-\frac{1}{2}\tilde{\mathbf{y}}^\top D^{-1}\tilde{\mathbf{y}} + \frac{1}{2}\tilde{\mathbf{y}}^\top D^{-1}\tilde{X}_m\left(\frac{\hat{g}+1}{\hat{g}}\tilde{X}_m^\top D^{-1}\tilde{X}_m\right)^{-1}\tilde{X}_m^\top D^{-1}\tilde{\mathbf{y}}\right],
 \end{aligned} \tag{4.21}$$

where \hat{g} is estimate from the screening step.

The fourth prior is Zellner-Siow prior, that is

$$\pi(\boldsymbol{\beta}_m|M_m) \sim \int \pi(g)N(\boldsymbol{\beta}_m|\mathbf{0}, g(\tilde{X}_m^\top D^{-1}\tilde{X}_m)^{-1}) dg. \tag{4.22}$$

The marginal density is:

$$\begin{aligned}
 m(\tilde{\mathbf{y}}|M_m) &= (2\pi)^{-n/2}|D|^{-1/2} \\
 &\quad \int \pi(g)(g+1)^{-\frac{p_m}{2}} \exp\left[-\frac{1}{2}\tilde{\mathbf{y}}^\top D^{-1}\tilde{\mathbf{y}} + \frac{1}{2}\tilde{\mathbf{y}}^\top D^{-1}\tilde{X}_m\left(\frac{g+1}{g}\tilde{X}_m^\top D^{-1}\tilde{X}_m\right)^{-1}\tilde{X}_m^\top D^{-1}\tilde{\mathbf{y}}\right] dg.
 \end{aligned}$$

4.7.2 Plots for simulation study

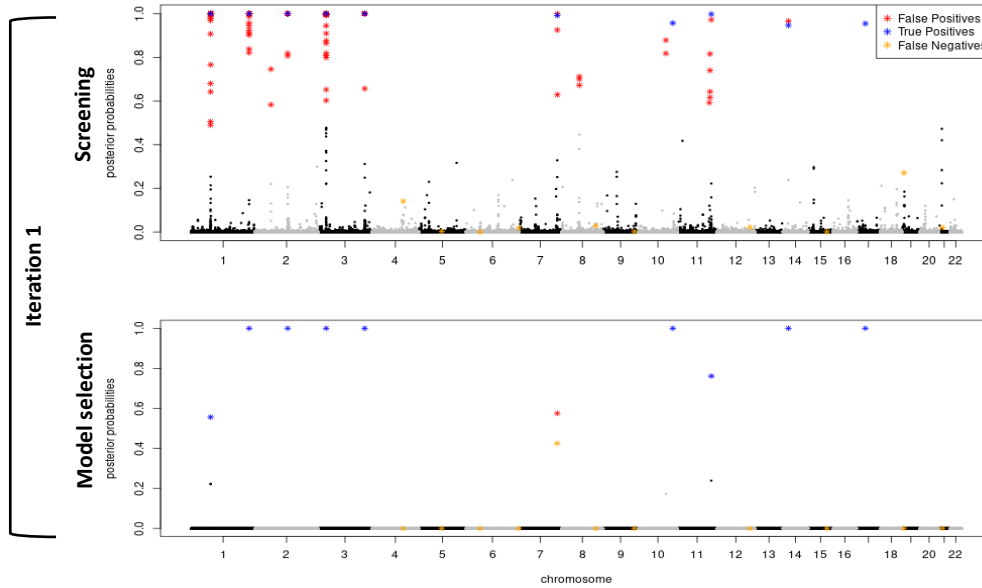


Figure 4.5: The posterior probabilities of all SNPs in the screening step of Iteration 1 for one dataset in the simulation study. The red dots are false positives. The blue dots are true positives. The yellow dots are false negatives.

In the screening step for the iteration 1, IBG3 screens hundreds of SNPs (red dots and blue dots) from 800000 SNPs. Due to the high correlation, there are several false positives around the true causal SNPs. However, in the model selection step for in the iteration 1, IBG3 selects the model with 10 SNPs, 9 out of 10 are causal SNPs, which extremely decreases the number of false positives. In the screening step for the iteration 2, conditional on the best model which IBG3 found in the last iteration, IBG3's screening step in the iteration 2 screens some new SNPs. Compared to the best model in the iteration 1, the best model in the iteration 2 contains more true positives (16 SNPs) and no more false positives (1 SNP). Since SNPs in the the best model change, IBG3 does not converge in this iteration. There is one more iteration. In the iteration 3, IBG3 found no more new SNPs. Then, the algorithm stops.

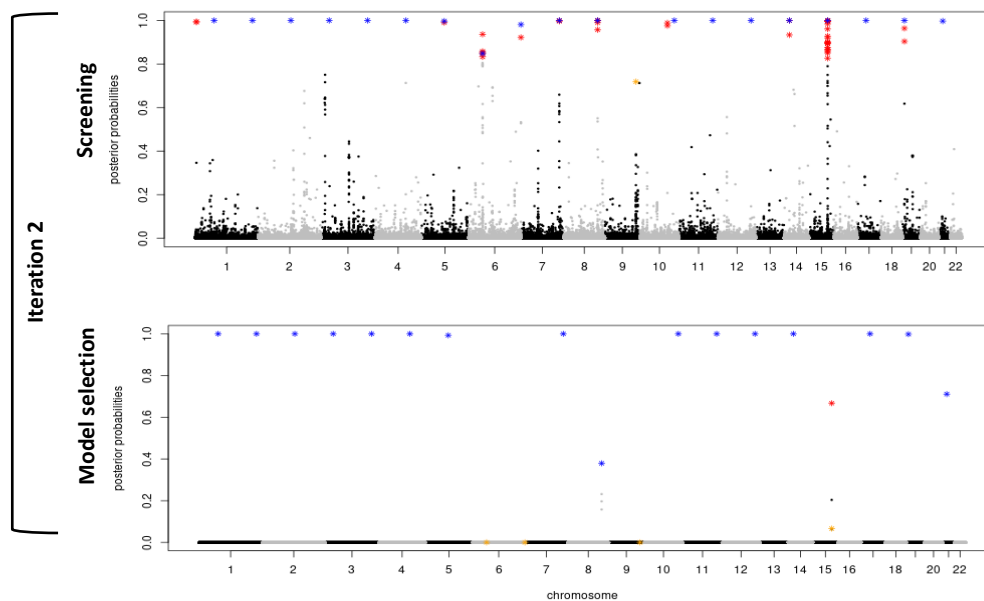


Figure 4.6: The posterior probabilities of all SNPs in the screening step of Iteration 2 for one dataset in the simulation study. The red dots are false positives. The blue dots are true positives. The yellow dots are false negatives.

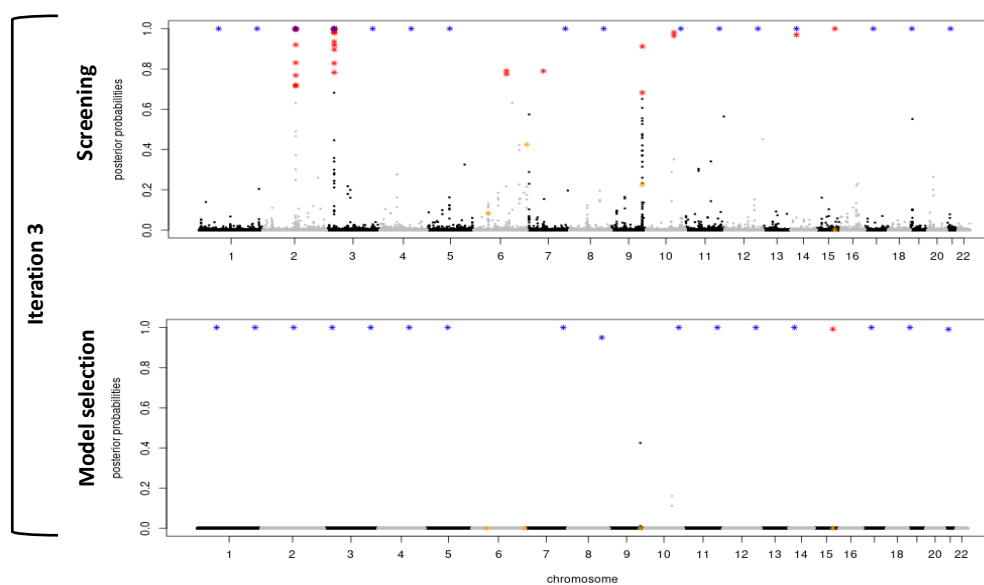


Figure 4.7: The posterior probabilities of all SNPs in the screening step of Iteration 3 for one dataset in the simulation study. The red dots are false positives. The blue dots are true positives. The yellow dots are false negatives.

Chapter 5

Conclusions

In this dissertation, we presents two novel Bayesian model selection methods for GLMMs (HCM and ARM) and two Bayesian variable selection methods for non-Gaussian GWAS data (BG2 and IBG3). Bayesian variable selection methods for GLMMs have difficulty to obtain the marginal likelihood function analytically. To solve this problem, our methods use pseudo likelihood approach to approximate non-Gaussian data in GLMMs by computing adjusted observations that are modeled by approximate Gaussian LMMs. Bayesian variable selection methods for non-Gaussian GWAS data have to solve $p \gg n$ problem in the GLMM structure since the number of SNPs p in GWAS data is from 10^5 to 10^6 , however the sample size n is only around 10^3 . In addition, SNPs are highly correlated. For ultra-high dimensional and highly correlated variable selection, BG2 and IBG3 employ two-step procedure, screening step and fine-mapping step, to reduce the high false discovery rate. In addition, IBG3 iterates these two steps to detect more SNPs with more small effect sizes. Simulation studies with real GWAS data show that BG2 can detect true positives as other popular GWAS methods but much less false positives. IBG3 can detect more true positives which cannot be detected by SMA methods and other fine-mapping methods.

The first two methods presented in this dissertation are HCM and ARM. HCM and ARM are two Bayesian model selection methods for GLMMs, which can select fixed effects and random effects simultaneously. HCM and ARM use pseudo likelihood approach to approximate non-Gaussian data in GLMMs and obtain the marginal likelihood function. HCM and ARM

use flat prior for the fixed effects. Since flat prior is improper, we develop an FBF approach to obtain posterior probabilities for candidate models. The simulation studies show that pseudo likelihood approach can help to obtain the marginal likelihood function, and using the posterior probabilities of models based on this marginal likelihood function can correctly select the true fixed effects and random effects. The second method presented in this dissertation is BG2. BG2 is a Bayesian two-step variable selection method for non-Gaussian GWAS data and GLMMs. BG2 also uses pseudo likelihood approach to solve marginal likelihood function problem in GLMMs. BG2 is designed for high-dimensional highly correlated data. BG2 has two steps, screening step and fine-mapping step. The screening step can filter a subset of the most associated SNPs. Then, the fine-mapping step considers the interaction between SNPs and reduces false discovery rate. In BG2, we propose nonlocal priors for coefficients in GLMMs. Nonlocal priors have been proven to be advantageous for high-dimensional problems in GLMs. Simulation studies show that BG2 with nonlocal priors for GLMMs can detect as many SNPs as other popular GWAS methods but less false positives. The third method presented in this dissertation is IBG3. IBG3 is a genome-wide fine-mapping method for non-Gaussian GWAS data and GLMMs. IBG3 expands on BG2 by iterating screening step and fine-mapping step. In the IBG3, we compare nonlocal prior, unit information prior, Zellner-g prior, and Zellner-Siow prior for coefficient in GLMMs. The simulation studies show that IBG3 with nonlocal prior and IBG3 with Zellner-g prior can detect more true positives and less false positives. However, IBG3 with Zellner-g prior takes less time than IBG3 with nonlocal prior.

In the future, there are two points where we can improve Bayesian variable selection method for GWAS data based on IBG3. First, all methods in this dissertation rely on the pseudo likelihood method. In the pseudo likelihood method, there are some matrix computations, such as inverse of matrix and spectrum decomposition. The computation complexity is

$O(n^3)$. The computation timing depends on the sample size. Therefore, finding a faster way to code pseudo likelihood method can benefit all methods in this dissertation. Second, after we solve the computation timing problem, we can extend our methods to analyze biobank-scale data. The sample size in biobanks is about 10^5 to 10^6 in magnitude. Larger sample size can provide more statistical power to analyze causal SNPs for interested phenotypes.

Bibliography

- [1] D. Altomare, G. Consonni, and L. La Rocca. Objective bayesian search of gaussian directed acyclic graphical models for ordered variables with non-local priors. *Biometrics*, 69(2):478–487, 2013.
- [2] S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall / CRC, Boca Raton, 2nd edition, 2014.
- [3] M. Baragatti and D. Pommeret. A study of variable selection using g-prior distribution with ridge parameter. *Computational Statistics and Data Analysis*, 56(6):1920–1934, 2012. ISSN 0167-9473. doi: <https://doi.org/10.1016/j.csda.2011.11.017>.
- [4] M. M. Barbieri and J. O. Berger. Optimal predictive model selection. *The Annals of Statistics*, 32(3):870 – 897, 2004.
- [5] H. Begleiter, T. Reich, V. Hesselbrock, B. Porjesz, T.-K. Li, M. A. Schuckit, H. J. Edenberg, J. P. Rice, et al. The collaborative study on the genetics of alcoholism. *Alcohol Health and Research World*, 19:228–228, 1995.
- [6] J. O. Berger and L. R. Pericchi. The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122, 1996.
- [7] L. J. Bierut, J. R. Strickland, J. R. Thompson, S. E. Afful, and L. B. Cottler. Drug use and dependence in cocaine dependent subjects, community-based individuals, and their siblings. *Drug and Alcohol Dependence*, 95(1-2):14–22, 2008.
- [8] L. K. Billings and J. C. Florez. The genetics of type 2 diabetes: what have we learned from gwas? *Annals of the New York Academy of Sciences*, 1212(1):59–77, 2010.

- [9] N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993.
- [10] B. Cai and D. B. Dunson. Bayesian covariance selection in generalized linear mixed models. *Biometrics*, 62(2):446–457, 2006.
- [11] K.-C. Chang, S. D. Diermeier, A. T. Yu, L. D. Brine, S. Russo, S. Bhatia, H. Alsudani, K. Kostroff, T. Bhuiya, E. Brogi, et al. Matar25 lncrna regulates the tensin1 gene to impact breast cancer progression. *Nature communications*, 11(1):6438, 2020.
- [12] S. Chaturvedi, M. Biswas, S. Sadhukhan, and A. Sonawane. Role of egfr and fasn in breast cancer progression. *Journal of Cell Communication and Signaling*, pages 1–34, 2023.
- [13] L. Chen, H. Chen, Y. Xing, and J. Li. ABCC1 regulates cocaine-associated memory, spine plasticity and GluA1 and GluA2 surface expression. *NeuroReport*, 32(10):833–839, 2021.
- [14] D. Clayton and J. Kaldor. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, pages 671–681, 1987.
- [15] M. Clyde. *BAS: Bayesian Variable Selection and Model Averaging using Bayesian Adaptive Sampling*, 2022. R package version 1.6.4.
- [16] S. Cui, S. Guha, M. A. R. Ferreira, and A. N. Tegge. hmmseq: A hidden Markov model for detecting differentially expressed genes from RNA-seq data. *The Annals of Applied Statistics*, 9(2):901–925, 2015.
- [17] S. Elsheikh, A. R. Green, M. A. Aleskandarany, M. Grainge, C. E. Paish, M. B. Lambros, J. S. Reis-Filho, and I. O. Ellis. Ccnd1 amplification and cyclin d1 expression in breast

- cancer and their relation with proteomic subgroups and patient outcome. *Breast cancer research and treatment*, 109:325–335, 2008.
- [18] L. Fagerberg, B. M. Hallström, P. Oksvold, C. Kampf, D. Djureinovic, J. Odeberg, M. Habuka, S. Tahmasebpoor, A. Danielsson, K. Edlund, et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & Cellular Proteomics*, 13(2):397–406, 2014.
- [19] M. A. R. Ferreira and V. De Oliveira. Bayesian reference analysis for Gaussian Markov random fields. *Journal of Multivariate Analysis*, 98(4):789–812, 2007.
- [20] M. A. R. Ferreira, E. M. Porter, and C. T. Franck. Fast and scalable computations for Gaussian hierarchical models with intrinsic conditional autoregressive spatial random effects. *Computational Statistics and Data Analysis*, 162:107264, 2021.
- [21] A. Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, 1(3):515–534, 2006.
- [22] S. R. Ghadamgahi, L. Hosseinzadeh, S. A. Khales, M. Nassiri, M. Alidoust, S. Etemadrezaei, A. K. Shamshiri, F. H. Shandiz, A. Pasdar, and F. Afzaljavan. Potential role of zinc finger 365 rs10822013 and rs10995190 in mammographic density, sporadic breast cancer risk, and prognosis. *Iranian journal of medical sciences*, 48(6):551, 2023.
- [23] M. Ghoussaini, J. D. French, K. Michailidou, S. Nord, J. Beesley, S. Canisus, K. M. Hillman, S. Kaufmann, H. Sivakumaran, M. M. Marjaneh, et al. Evidence that the 5p12 variant rs10941679 confers susceptibility to estrogen-receptor-positive breast cancer through fgf10 and mrps30 regulation. *The American Journal of Human Genetics*, 99(4):903–911, 2016.

- [24] W. Guo, Q. Wang, Y. Zhan, X. Chen, Q. Yu, J. Zhang, Y. Wang, X.-j. Xu, and L. Zhu. Transcriptome sequencing uncovers a three–long noncoding rna signature in predicting breast cancer survival. *Scientific reports*, 6(1):27931, 2016.
- [25] L. Jiang, Z. Zheng, H. Fang, and J. Yang. A generalized linear mixed model association tool for biobank-scale data. *Nature Genetics*, 53(11):1616–1621, 2021.
- [26] V. E. Johnson and D. Rossell. On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):143–170, 2010.
- [27] V. E. Johnson and D. Rossell. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660, 2012.
- [28] M. A. Jones, J.-J. Shen, Y. Fu, H. Li, Z. Yang, and C. S. Grierson. The arabidopsis rop2 gtpase is a positive regulator of both root hair initiation and tip growth. *The Plant Cell*, 14(4):763–776, 2002.
- [29] M. M. Julkowska, I. T. Koevoets, S. Mol, H. Hoefsloot, R. Feron, M. A. Tester, J. J. Keurentjes, A. Korte, M. A. Haring, G.-J. de Boer, et al. Genetic components of root architecture remodeling in response to salt stress. *The Plant Cell*, 29(12):3198–3213, 2017.
- [30] H. M. Kang, N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, 2008.
- [31] H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S.-y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354, 2010.

- [32] R. E. Kass and L. Wasserman. A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the american statistical association*, 90(431):928–934, 1995.
- [33] M. J. Keefe, M. A. R. Ferreira, and C. T. Franck. On the formal specification of sum-zero constrained intrinsic conditional autoregressive models. *Spatial Statistics*, 24:54–65, 2018. ISSN 2211-6753.
- [34] M. J. Keefe, M. A. R. Ferreira, and C. T. Franck. Objective Bayesian analysis for Gaussian hierarchical models with intrinsic conditional autoregressive priors. *Bayesian Analysis*, 14(1):181 – 209, 2019.
- [35] P. Kraft and C. A. Haiman. Gwas identifies a common breast cancer risk allele among brca1 carriers. *Nature genetics*, 42(10):819–820, 2010.
- [36] R. Lardon, E. Wijnker, J. Keurentjes, and D. Geelen. The genetic framework of shoot regeneration in Arabidopsis comprises master regulators and conditional fine-tuning factors. *Communications Biology*, 3(1):1–13, 2020.
- [37] H. Lei and C.-X. Deng. Fibroblast growth factor receptor 2 signaling in breast cancer. *International journal of biological sciences*, 13(9):1163, 2017.
- [38] T. Lencz and A. Malhotra. Targeting the schizophrenia genome: a fast track strategy from gwas to clinic. *Molecular psychiatry*, 20(7):820–826, 2015.
- [39] I. Leppik, F. Dreifuss, R. Porter, T. Bowman, N. Santilli, M. Jacobs, C. Crosby, J. Cloyd, J. Stackman, N. Graves, et al. A controlled study of progabide in partial seizures: methodology and results. *Neurology*, 37(6):963–963, 1987.
- [40] P.-F. Liu, C.-W. Shu, H.-C. Yang, C.-H. Lee, H.-H. Liou, L.-P. Ger, Y.-D. T. Tzeng, and

- W.-C. Wang. Combined evaluation of map1lc3b and sqstm1 for biological and clinical significance in ductal carcinoma of breast cancer. *Biomedicines*, 9(11):1514, 2021.
- [41] Z. Liu, V. J. Berrocal, A. J. Bartsch, and T. D. Johnson. Pre-surgical fMRI data analysis using a spatially adaptive conditionally autoregressive model. *Bayesian Analysis*, 11: 599–625, 2016.
- [42] P. McCullagh and J. A. Nelder. *Generalized Linear Models*, volume 37. CRC Press, 1989.
- [43] S. Meyer, L. Held, and M. Höhle. Spatio-temporal analysis of epidemic phenomena using the R package surveillance. *Journal of Statistical Software*, 77, 2017.
- [44] S. H. Mueller, A. G. Lai, M. Valkovskaya, K. Michailidou, M. K. Bolla, Q. Wang, J. Dennis, M. Lush, Z. Abu-Ful, T. U. Ahearn, et al. Aggregation tests identify new gene associations with breast cancer in populations with diverse ancestry. *Genome medicine*, 15(1):1–18, 2023.
- [45] D. Müller, F. Technow, and A. E. Melchinger. Shrinkage estimation of the genomic relationship matrix can improve genomic estimated breeding values in the training set. *Theoretical and Applied Genetics*, 128(4):693–703, 2015.
- [46] P. Muller, G. Parmigiani, and K. Rice. FDR and Bayesian multiple comparisons rules. In J. M. Bernardo, J. O. M. J. Bayarri, Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics 8*. Oxford: Oxford University Press, 2007.
- [47] M. A. Newton, A. Noueir, D. Sarkar, and P. Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176, 2004.

- [48] A. Nikooienejad, W. Wang, and V. E. Johnson. Bayesian variable selection for binary outcomes in high-dimensional genomic studies using non-local priors. *Bioinformatics*, 32(9):1338–1345, 2016.
- [49] P. Nouvellet, S. Bhatia, A. Cori, K. E. Ainslie, M. Baguelin, S. Bhatt, A. Boonyasiri, N. F. Brazeau, L. Cattarino, L. V. Cooper, et al. Reduction in mobility and COVID-19 transmission. *Nature Communications*, 12(1):1–9, 2021.
- [50] A. O’Hagan. Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):99–118, 1995.
- [51] J. Poland, J. Endelman, J. Dawson, J. Rutkoski, S. Wu, Y. Manes, S. Dreisigacker, J. Crossa, H. Sánchez-Villeda, M. Sorrells, et al. Genomic selection in wheat breeding using genotyping-by-sequencing. *The Plant Genome*, 5(3), 2012.
- [52] N. G. Polson and J. G. Scott. On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902, 2012.
- [53] E. M. Porter, C. T. Franck, and M. A. R. Ferreira. Objective Bayesian model selection for spatial hierarchical models with intrinsic conditional autoregressive priors. *Bayesian Analysis*, 2023. to appear.
- [54] D. Rossell and D. Telesca. Nonlocal priors for high-dimensional estimation. *Journal of the American Statistical Association*, 112(517):254–265, 2017.
- [55] D. Rossell, D. Telesca, and V. E. Johnson. High-dimensional Bayesian classifiers using non-local priors. In *Statistical Models for Data Analysis*, pages 305–313. Springer, 2013.
- [56] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

- [57] N. Sanyal and M. A. R. Ferreira. Bayesian wavelet analysis using nonlocal priors with an application to fMRI analysis. *Sankhya B*, 79(2):361–388, 2017.
- [58] N. Sanyal, M.-T. Lo, K. Kauppi, S. Djurovic, O. A. Andreassen, V. E. Johnson, and C.-H. Chen. GWASinlps: non-local prior based iterative SNP selection tool for genome-wide association studies. *Bioinformatics*, 35(1):1–11, 2019.
- [59] R. Sauter and L. Held. Network meta-analysis with integrated nested Laplace approximations. *Biometrical Journal*, 57(6):1038–1050, 2015.
- [60] D. J. Schaid, W. Chen, and N. B. Larson. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8):491–504, 2018.
- [61] J. G. Scott and J. O. Berger. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, pages 2587–2619, 2010.
- [62] L. Scrucca. GA: A Package for Genetic Algorithms in R. *Journal of Statistical Software*, 53(4):1–37, 2013. doi: 10.18637/jss.v053.i04.
- [63] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.
- [64] B. C. Sutradhar and R. P. Rao. On quasi-likelihood inference in generalized linear mixed models with two components of dispersion. *The Canadian Journal of Statistics*, pages 415–435, 2003.
- [65] R. Tawiah, G. J. McLachlan, and S. K. Ng. A bivariate joint frailty model with mixture framework for survival analysis of recurrent events with dependent censoring and cure fraction. *Biometrics*, 76(3):753–766, 2020.

- [66] P. Ten Eyck and J. E. Cavanaugh. An alternate approach to pseudo-likelihood model selection in the generalized linear mixed modeling framework. *Sankhya B*, 80(1):98–122, 2018.
- [67] P. F. Thall and S. C. Vail. Some covariance models for longitudinal count data with overdispersion. *Biometrics*, pages 657–671, 1990.
- [68] A. T. Tredennick, G. Hooker, S. P. Ellner, and P. B. Adler. A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology*, 102(6): e03336, 2021.
- [69] A. J. van de Loo, M. Mackus, O. Kwon, I. M. Krishnakumar, J. Garsen, A. D. Kranefeld, A. Scholey, and J. C. Verster. The inflammatory response to alcohol consumption and its role in the pathology of alcohol hangover. *Journal of Clinical Medicine*, 9(7): 2081, 2020.
- [70] S. Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(12):3571–3594, 2010.
- [71] J. Williams, M. A. R. Ferreira, and T. Ji. BICOSS: Bayesian iterative conditional stochastic search for GWAS. *BMC Bioinformatics*, 23:1–14, 2022.
- [72] R. Wolfinger and M. O’Connell. Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48(3-4):233–243, 1993.
- [73] H.-H. Wu, M. A. R. Ferreira, M. Elkhoully, and T. Ji. Hyper nonlocal priors for variable selection in generalized linear models. *Sankhya A*, 82(1):147–185, 2020.
- [74] J. Xie, T. Ji, M. A. R. Ferreira, Y. Li, B. N. Patel, and R. M. Rivera. Modeling allele-

- specific expression at the gene and SNP levels simultaneously by a Bayesian logistic mixed regression model. *BMC Bioinformatics*, 20(1):1–13, 2019.
- [75] D. Xu, A. Chatterjee, and M. Daniels. A note on posterior predictive checks to assess model fit for incomplete data. *Statistics in Medicine*, 35(27):5029–5039, 2016.
- [76] S. Xu, M. A. Ferreira, E. M. Porter, and C. T. Franck. Bayesian model selection for generalized linear mixed models. *Biometrics*, 2023.
- [77] S. Xu, M. A. R. Ferreira, E. M. Porter, and C. T. Franck. The GLMMselect package. <https://CRAN.R-project.org/package=GLMMselect>, 2023. [accessed 20-April-2023].
- [78] S. Xu, J. Williams, and M. A. Ferreira. Bg2: Bayesian variable selection in generalized linear mixed models with nonlocal priors for non-gaussian gwas data. *BMC bioinformatics*, 24(1):343, 2023.
- [79] S. Xu, J. Williams, and M. A. R. Ferreira. The BG2 package. <https://bioconductor.org/packages/BG2>, 2023. [accessed 2023].
- [80] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569, 2010.
- [81] J. Yu, G. Pressoir, W. H. Briggs, I. Vroh Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208, 2006.
- [82] J. Yuan, X. Wang, Y. Zhao, N. U. Khan, Z. Zhao, Y. Zhang, X. Wen, F. Tang, F. Wang,

- and Z. Li. Genetic basis and identification of candidate genes for salt tolerance in rice by gwas. *Scientific reports*, 10(1):9958, 2020.
- [83] A. Zellner and A. Siow. Posterior odds ratios for selected regression hypotheses. *Trabajos de estadística y de investigación operativa*, 31:585–603, 1980.
- [84] H. Zhang, T. U. Ahearn, J. Lecarpentier, D. Barnes, J. Beesley, G. Qi, X. Jiang, T. A. O’Mara, N. Zhao, M. K. Bolla, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nature Genetics*, 52(6):572–581, 2020.
- [85] X. Zhang, W. Ding, D. Xue, X. Li, Y. Zhou, J. Shen, J. Feng, N. Guo, L. Qiu, H. Xing, et al. Genome-wide association studies of plant architecture-related traits and 100-seed weight in soybean landraces. *BMC Genomic Data*, 22(1):1–14, 2021.
- [86] X. Zhang, C. Liu, and C. Sun. A novel glycolysis-related four-mrna signature for predicting the survival of patients with breast cancer. *Frontiers in genetics*, 12:606937, 2021.
- [87] Z. Zhang, E. Ersoz, C.-Q. Lai, R. J. Todhunter, H. K. Tiwari, M. A. Gore, P. J. Bradbury, J. Yu, D. K. Arnett, J. M. Ordovas, et al. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 42(4):355–360, 2010.
- [88] X. Zhou and M. Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821–824, 2012.
- [89] Y. Zou, P. Carbonetto, G. Wang, and M. Stephens. Fine-mapping from summary data with the “sum of single effects” model. *PLoS Genetics*, 18(7):e1010299, 2022.
- [90] X. Zuo, H. Wang, Y. Mi, Y. Zhang, X. Wang, Y. Yang, and S. Zhai. The association

of *cas16* variants with breast cancer risk in a northwest chinese female population.
Molecular Medicine, 26(1):1–10, 2020.