THE UNIVERSITY OF CHICAGO

A NEW MODELING FRAMEWORK FOR MULTI-TRAIT MAPPING OF BINARY AND QUANTITATIVE PHENOTYPES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
YI WEI

CHICAGO, ILLINOIS
MARCH 2025

Dedicated to my parents, Zhen Zhen and Ruifeng Wei.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

On a cold Chicago winter night, as the world outside lay still, my heart grows warm as the memories of my PhD journey over the past years come vividly to life. This journey has been filled with challenges and growth, made possible by the guidance, support, and kindness of so many, whom I wish to acknowledge and deeply thank here. I am especially grateful to the Department of Statistics at the University of Chicago for providing me with an inspiring and supportive environment, and to the university itself for being the foundation of my academic and personal growth during these transformative years.

First and foremost, I wish to express my heartfelt gratitude to my advisor, Professor Mary Sara McPeek, who has been my academic mentor and guiding light throughout my PhD journey. I am forever indebted to her for her unwavering support, kindness, patience, and professionalism. She always believed in me, encouraging me to persevere and grow. From her, I have not only learned what it takes to be a good researcher but also how to be a better person. She helped me develop a mindset that allows me to face challenges with confidence and resilience, unlocking my potential. Her work ethic, passion for research, dedication, and brilliant ideas have shaped my approach to conducting research with professionalism and integrity. I cannot find the words to fully express my gratitude for everything she has done to support me. Without her encouragement and guidance, this thesis would not exist, and her impact on my life and career is unforgettable. For this, I will always hold her in the highest regard and carry her lessons with me in everything I do.

I would also like to sincerely thank my committee members, Professor Matthew Stephens and Professor Dan Nicolae, for their invaluable guidance and insightful feedback throughout my PhD studies. Their thoughtful advice and suggestions were crucial in shaping this thesis, and their support during my proposal and defense was greatly appreciated. I admire them as brilliant scholars, and their encouragement helped me build confidence. My gratitude also extends to all the faculty and staff in the department for their support and dedication.

I would like to extend my gratitude to my friends and fellow doctoral students for their support, assistance, and engaging discussions: Huanlin Zhou, Yi Wang, Chih-Hsuan Wu, Zehao Niu, Wanrong Zhu, and many others. Their companionship made the challenges of academic life more manageable and even enjoyable. A special mention goes to Huanlin Zhou, my long-time roommate, whose care, help, and the joyful moments we shared brought warmth to my PhD journey. I am profoundly thankful to Joelle Mbatchou, who assisted me with my research questions even after her graduation. I deeply appreciate her willingness to share her expertise. I am also grateful to my best friends of more than ten years: Ge Feng, Xiaoxue Yang, Xiaoyan Chen, and Jiajing Chen. Being so far away, their consistent support and belief in me have given me so much solace and strength.

Lastly, but certainly not least, I want to thank my family. I am grateful to my aunt Ruihong Wei and my cousin Jing Yang for their care and support throughout this journey. I am especially thankful for my wonderful parents: my mom, Zhen Zhen, and my dad, Ruifeng Wei. Their unconditional love and steadfast support have been the foundation of everything I have achieved. They have always stood by my side, encouraging me to pursue my dreams fearlessly and trusting every decision I make. They are the anchor of my life. They always tell me that they are proud of me, and I want to let them know that I am so honored to be their daughter. Despite the 10,000 kilometers that separate us, our hearts remain deeply connected. Their encouragement lifted me whenever I doubted myself and empowered me to strive for better and never give up. I love them more than words can express. With their love and support, I believe I can continue to be brave and keep growing.

ABSTRACT

Joint modeling of multiple closely related quantitative traits with genetic variants is widely applied in genetics to increase power for detecting associations. Linear mixed models (LMMs) are one of the most commonly used approaches. However, when considering a small number of disease-related traits, it is common for one or more of the traits to be binary, not quantitative. Previous work has found that using LMM to analyze binary traits suffers from substantial power loss if covariate effects are important. Generalized linear mixed-model methods could, in principle provide a solution to this problem, but in practice the penalized quasi-likelihood estimation methods that make them computationally feasible are too inaccurate to provide reliable type I error control. Furthermore, assessing the significance of multi-trait associations with single or multiple genetic variants is challenging, particularly in samples with population structure and related individuals. There is a lack of methods capable of jointly modeling both binary and quantitative traits in the presence of population structure or relatedness, while also accommodating multiple genetic variants and remaining robust to ascertainment and model misspecification. To address these limitations, we developed BCMAP (Binary and Continuous Multi-trait Association test with Population structure), a novel modeling framework for multi-trait mapping of a combination of binary and quantitative phenotypes, which is based on a mixed-effects quasi-likelihood framework. BCMAP accommodates covariates, population structure, and relatedness, capturing the dichotomous nature of binary traits, and is suitable for testing with both single and multiple genetic variants. Our test employs a retrospective approach, ensuring robustness to both ascertainment and misspecification of the phenotype model. Additionally, we integrate the recently proposed genetic association test method, JASPER (Joint Association analysis in Structured samples based on approximating a PERmutation distribution). JASPER is a fast, powerful, and robust genetic association test that effectively accounts for population structure, enhancing the accuracy and reliability of our analysis. Parameter estimation for the binary trait(s) in this setting presents additional challenges beyond those for the quantitative trait case. As part of estimating the correlation matrix, we explore a recently proposed parametrization which enforces the positive (semi) definiteness and which can be viewed as a multivariate generalization of Fisher's Z-transformation of a single correlation. In simulations, BCMAP achieves accurate type I error calibration and demonstrates improved power over existing methods. We apply BCMAP to analyze the genetic associations of genetic variants with diabetes and BMI in the Framingham Heart Study.

CHAPTER 1

INTRODUCTION

For identifying genetic variants associated with a trait, the use of a univariate association test has achieved many interesting results. For studying genetic associations with multiple (often correlated) traits, univariate testing combined with multiple testing corrections has been commonly employed due to its computational efficiency. However, this method is not as powerful or efficient as a joint modeling of multiple traits method to detect association between the traits and genetic variants [1]. Joint modeling of multiple closely-related quantitative traits using linear mixed models (LMMs) has been proposed as one solution. One major application of these models is to increase power to detect genetic variants associated with multiple traits or associated with one trait and not others [2]. Furthermore, in genome-wide association studies (GWASs), analyzing multiple traits simultaneously provides valuable insights into the genetic architecture of complex traits [3, 4, 5, 6, 7] and enhances the prediction of comorbidities using genetic data [8, 9, 10]. Joint analysis of multiple traits and multiple variants also increases the power to identify gene-based associations [5].

In practice, with multiple closely-related traits, it is common for one or more of the traits to be binary, not quantitative. For example, weight and diabetes are closely related traits, but weight is a quantitative trait and diabetes is a binary trait. Researchers have demonstrated that using LMM to analyze binary traits can lead to incorrect type I error due to the violation of the homoscedasticity assumption in binary trait data [1]. Additionally, LMMs can suffer from substantial power loss when covariate effects are important [11].

There is a clear need for efficient methods that can perform multi-trait genetic association testing while accommodating a combination of binary and quantitative traits. To address this gap, we propose a new model in this thesis for multi-trait mapping of both binary and quantitative phenotypes. The proposed model is based on a mixed-effects quasi-likelihood framework and can incorporate covariates and population structure and relatedness. It effectively captures the dichotomous nature of binary traits and supports both multi-trait single-variant and multi-trait multi-variant tests. The test built on our model is based on a retrospective approach, making it robust to misspecification of the phenotype model and ascertainment. We name the proposed method BCMAP (Binary and Continuous Multi-trait Association test with Population structure). In this thesis, we first address the parameter estimation problem for a model involving multiple binary traits, which presents additional challenges compared to the quantitative traits case. We then demonstrate, through simulation results for the multi-trait single-variant association testing scenario, that BCMAP achieves accurate type I error calibration and exhibits improved power compared to existing methods. Furthermore, we extend the method to the multi-trait multi-variant association testing scenario and use simulation results to show that BCMAP maintains accurate type I error calibration and demonstrates robust power. Finally, we apply BCMAP to analyze the genetic associations of genetic variants with diabetes and BMI in the Framingham Heart Study.

1.1 Multi-Trait Single-Variant Association Testing

For single genetic variant testing, there are some past works related to our topic. CARAT [11] is a univariate binary-trait association approach, which accounts for relevant covariate information, controls for unknown population structure, employs a logistic mean structure, and maintains the necessary mean-variance relationship for a binary trait. Because CARAT uses a retrospective approach, it is more robust to misspecification of the phenotype model and ascertainment. CERAMIC [12] extends the CARAT method to allow samples with related individuals and to incorporate partially missing data, which makes it more powerful than many methods when the sample includes related individuals with some missing data. GMMAT [13] is a computationally efficient logistic mixed model approach for binary-trait association testing, which effectively controls for population structure and relatedness. However, GMMAT is based on a prospective testing approach, which makes it result in worse type I error and lower power than CERAMIC when the phenotype model is misspecified [12]. Furthermore, GMMAT relies on a penalized quasi-likelihood approach that sacrifices accuracy for computational speed [14].

These methods focus on univariate association tests for single genetic variant and binary traits. Zhou and Stephens [15] proposed efficient multivariate linear mixed model (mvLMM) algorithms for genome-wide association studies, which involves modeling multiple correlated quantitative traits jointly and accounting for relatedness among samples and the method is working for single genetic variant testing. However, their method does not accommodate binary traits. Wang, Meigs, and Dupuis [1] proposed an efficient bivariate robust score test for one binary trait and one quantitative trait with single genetic variant, which is applicable for both family-based and unrelated samples, but they do not analyze with more traits. Our proposed method BCMAP builds on the strengths of these approaches while addressing their limitations.

1.2 Multi-Trait Multi-Variant Association Testing

Several methods have been proposed for multi-trait multi-variant association testing. Tools like MTAR [16], MTaSPUsSet [17], metaCCA [18] and MGAS [19] use summary statistics for association testing between multiple variants and multiple traits. GAMuT [20] employed a dual-kernel distance-covariance method that compares similarity in multiple phenotypes to similarity in multiple genetic variants to analyze cross-phenotype associations of rare variants. MSKAT [21] is a score-based sequence kernel association test that assesses the joint effects of multiple variants and multiple traits. DKAT [22] builds on GAMuT's dual-kernel approach but delivers more robust performance in scenarios where the number of phenotypes is high relative to the sample size. KMU [23] uses a kernel-based multivariate U-statistics approach that is applicable to both binary and quantitative traits.

However, none of the above methods account for population relatedness and sub-structure. Failing to adjust for sample structure can result in reduced power and an increased type I error rate [24]. Multi-SKAT [25] uses multivariate kernel regression to test association of multiple variants with multiple continuous phenotypes while accounting for related individuals. However, this method is not applicable when there are binary traits in the analysis.

Our proposed method BCMAP for multi-trait multi-variant association testing addresses these challenges. It accounts for population structure and relatedness, incorporates covariates, and is applicable to scenarios where some traits are binary and others are quantitative. This approach offers a robust and flexible solution for complex genetic association studies.

CHAPTER 2

QUASI-LIKELIHOOD MODEL FOR MULTIPLE BINARY TRAITS AND PARAMETER ESTIMATION

In this chapter, we propose a novel quasi-likelihood model for analyzing multiple binary traits in association with a single genetic variant. The proposed method incorporates covariates and accounts for related individuals and population structure within the sample. We provide a detailed illustration of the parameter estimation procedure, highlighting the additional challenges it presents compared to the case of quantitative traits.

2.1 Quasi-Likelihood Model for Multiple Binary Traits and Single Genetic Variant

Assume there are n individuals and p traits. Denote phenotype by a $p \times n$ matrix $Y_{p \times n}$, whose i^{th} row $Y_{i\cdot}$ contains the phenotypes of all of the n individuals for trait i, and whose j^{th} column $Y_{\cdot j}$ contains the p phenotypes for individual j, so the $(i,j)^{th}$ element of Y, Y_{ij} is the value of the i^{th} phenotype for j^{th} individual. $G_{n \times 1} = [G_1, G_2, ..., G_n]^T$ denotes the vector of genotypes for the n individuals at the variant to be tested, where G_j equals the minor allele account (0,1 or 2) of individual j at the variant. Suppose we observe $k-1 \geqslant 0$ covariates, let $X_{k \times n}$ be an $k \times n$ covariate matrix, whose j^{th} column $X_{\cdot j}$ contains an intercept term represented as 1 and the values of $k-1$ non-constant covariates for individual j, so the first row of $X_{k \times n}$ is a row of 1s.

To model the phenotype matrix Y conditional on X and G, we take a quasi-likelihood approach. We specify only the conditional mean and variance structures of Y. For mean structure, we assume that, for $i = 1, ..., p$ and $j = 1, ..., n$,

$$E(Y_{ij}|X, G) = \mu_{ij}, \quad g(\mu_{ij}) = (\beta X)_{ij} + (\gamma G^T)_{ij}, \tag{2.1}$$

where $g(\cdot)$ is a known function, β is a $p \times k$ matrix of the unknown fixed effects of covariates and γ is a vector of length p representing fixed effects of tested variant on the phenotypes. We take $g(\cdot)$ to be the logit link function given by:

$$g(\mu_{ij}) = log \frac{\mu_{ij}}{1 - \mu_{ij}}.$$

We choose the logit link function because in this case, the linear coefficients can be interpreted as the size of an additive effect on the log odds scale [11]. For the conditional variance structure, we have:

$$\Omega := Var(vec(Y)|X, G) = \Gamma^{1/2} \Sigma \Gamma^{1/2}. \tag{2.2}$$

Note for an $n \times m$ matrix A, $vec(A)$ is a linear transformation of A to an nm-dimensional vector by stacking columns of A on top of one another [26]. Γ is an np-dimensional diagonal matrix, with s^{th} diagonal element, where $s = p(j-1) + i$, given by $\Gamma_{ss} = Var(Y_{ij}|X, G)$, which is equal to $\mu_{ij}(1 - \mu_{ij})$. Σ is an $np \times np$ correlation matrix (defined as a positive semi-definite matrix with 1s on the diagonal) that does not involve the mean structure. Under this specification, the marginal conditional variance is determined by the conditional mean according to $Var(Y_{ij}|X, G) = E(Y_{ij}|X, G)(1 - E(Y_{ij}|X, G))$ and is consistent with the dichotomous nature of the binary traits, because for binary random matrix Y, regardless of the joint distribution, the marginal distribution of each Y_{ij} is always a Bernoulli distribution. Σ is defined as follows:

$$\Sigma = K \otimes (D^{1/2} C_g D^{1/2}) + I_{n \times n} \otimes (\tilde{D}^{1/2} C_e \tilde{D}^{1/2}), \tag{2.3}$$

where $\otimes$ represents Kronecker product, K is an $n \times n$ genetic relationship matrix or kinship matrix, C_g is a $p \times p$ correlation matrix (defined as a positive semi-definite matrix with 1s on the diagonal) for the random effects of a single SNP on each of the p traits, C_e is a $p \times p$ correlation matrix (defined as a positive semi-definite matrix with 1s on the diagonal) for the non-genetic individual random effects on each of the p traits for a single person. D is a p-dimensional diagonal matrix, with i^{th} diagonal element $0 \leqslant d_i \leqslant 1$ representing the proportion of trait i's residual variance that is due to additive polygenic effects, and $\tilde{D} = I_{p\times p} - D$. $D^{1/2}C_gD^{1/2}$ is the genetic variance component, and $\tilde{D}^{1/2}C_e\tilde{D}^{1/2}$ is an environmental variance component.

The unknown parameters are γ, the parameter of interest, and the nuisance parameters: the entries of β, and the variance components: lower off-diagonal elements of C_g and C_e, and the diagonal elements of D. The number of unknown nuisance parameters is $p(p+k+1)-p = p^2+pk$ and is p fewer than in a standard multi-trait LMM for quantitative traits because of the mean-variance relationship for the binary trait. We denote the variance components by a vector Θ of length p^2.

2.2 Parameter Estimation for Coefficients

To get the estimates of the parameters, we form a system of estimating equations and solve them. The procedure to obtain the estimating equations is described now. At this section, we focus on the problem of parameter estimation for coefficients.

We bind the notations for genotype vector and covariate matrix to define $\tilde{X} = \begin{bmatrix} X \\ G^T \end{bmatrix}$ and $\tilde{\beta} = [\beta, \gamma]$. To estimate $\tilde{\beta}$ given the variance components parameters (Θ), we propose to generalize the method in CARAT [11] paper.

From the quasi-likelihood model we defined before we have:

$$\mu = E(Y|X,G) = logit^{-1}(\tilde{\beta}\tilde{X}), \tag{2.4}$$

$$\Omega := Var(vec(Y)|X,G) = \Gamma^{1/2}\Sigma\Gamma^{1/2}, \tag{2.5}$$

where Σ defined by equation (2.3).

Note for matrices A and B of dimensions $k \times l$, $l \times m$, we have:

$$vec(AB) = (B^T \otimes I_{k\times k})vec(A). \tag{2.6}$$

Therefore, we can get the conditional expectation of $vec(Y)$ from equation (2.4):

$$vec(\mu) = E(vec(Y)|X, G) = logit^{-1}(vec(\tilde{\beta}\tilde{X})) = logit^{-1}((\tilde{X}^T \otimes I_{p\times p})vec(\tilde{\beta}))). \tag{2.7}$$

Equations (2.5) and (2.7) build the quasi-likelihood model for $vec(Y)$. For known variance components(Θ), the quasi-likelihood function for $vec(\tilde{\beta})$ can be differentiated to obtain the quasi-score function for $vec(\tilde{\beta})$:

$$U(vec(\tilde{\beta})) = M^T\Omega^{-1}(vec(Y) - vec(\mu)), \tag{2.8}$$

where $M = M(vec(\tilde{\beta})) = \frac{\partial vec(\mu)}{\partial vec(\tilde{\beta})} = \Gamma(\tilde{X}^T \otimes I_{p\times p})$. Setting $U(\tilde{\beta}) = 0$ gives the generalized estimating equation for $vec(\tilde{\beta})$:

$$(\tilde{X} \otimes I_{p\times p})\Gamma^{1/2}\Sigma^{-1}\Gamma^{-1/2}(vec(Y) - vec(\mu)) = 0. \tag{2.9}$$

To solve for $vec(\tilde{\beta})$, a modified Newton-Raphson algorithm with Fisher scoring is used, which involves iteratively updating $vec(\beta)$ by:

$$\begin{aligned} vec(\tilde{\beta})_{(j+1)} = vec(\tilde{\beta})_{(j)} + \{M^T(vec(\tilde{\beta})_{(j)})\Omega^{-1}M(vec(\tilde{\beta})_{(j)})\}^{-1}\cdot \\ \{M^T(vec(\tilde{\beta})_{(j)})\Omega^{-1}[vec(Y) - vec(\mu(vec(\tilde{\beta})_{(j)}))]\}. \end{aligned} \tag{2.10}$$

Since the i^{th} row of Y represents the phenotype of the n individuals for trait i, and the i^{th} row of $\tilde{\beta}$ represents the fixed effects of the covariates and genotype for trait i, we can estimate each row of $\tilde{\beta}$ under the single binary trait model incorporating related individuals proposed in CARAT [11] independently and stack the estimators together to form the initial value for $vec(\tilde{\beta})$. Starting at the initial values for $vec(\tilde{\beta})$ estimated from CARAT [11], we run this iterative procedure until convergence to obtain $vec(\tilde{\beta})$, for fixed variance components(Θ). Plug in the equation for M, equation (2.10) becomes

$$\begin{aligned} vec(\tilde{\beta})_{(j+1)} = vec(\tilde{\beta})_{(j)} + \{(\tilde{X} \otimes I_{p\times p})\Gamma^{1/2}(vec(\tilde{\beta})_{(j)})\Sigma^{-1}\Gamma^{1/2}(vec(\tilde{\beta})_{(j)})(\tilde{X}^T \otimes I_{p\times p})\}^{-1} \\ \cdot\{(\tilde{X} \otimes I_{p\times p})\Gamma^{1/2}(vec(\beta)_{(j)})\Sigma^{-1}\Gamma^{-1/2}(vec(\tilde{\beta})_{(j)})[vec(Y) - vec(\mu(vec(\tilde{\beta})_{(j)}))]\}. \end{aligned} \tag{2.11}$$

Equation (2.11) involve Σ^{-1}, but Σ is an $np \times np$ matrix, when n and p are large, it will be computationally expensive to invert Σ, so we do not want to invert it directly. We get the spectral decomposition of K: $K = UVU^T$, where U is orthogonal and $V = diag(\delta_1, \delta_2, ..., \delta_n)$, then we have:

$$\Sigma = (U \otimes I_{p\times p})[V \otimes (D^{1/2}C_gD^{1/2}) + I_{n\times n} \otimes (\tilde{D}^{1/2}C_e\tilde{D}^{1/2})](U^T \otimes I_{p\times p}), \tag{2.12}$$

note both V and $I_{n\times n}$ are diagonal matrices, so $[V \otimes (D^{1/2}C_gD^{1/2}) + I_{n\times n} \otimes (\tilde{D}^{1/2}C_e\tilde{D}^{1/2})]$ is in a block diagonal form. From equation (2.12) we have:

$$\Sigma^{-1} = (U \otimes I_{p\times p})[V \otimes (D^{1/2}C_gD^{1/2}) + I_{n\times n} \otimes (\tilde{D}^{1/2}C_e\tilde{D}^{1/2})]^{-1}(U^T \otimes I_{p\times p}), \tag{2.13}$$

and since $[V \otimes (D^{1/2}C_gD^{1/2}) + I_{n\times n} \otimes (\tilde{D}^{1/2}C_e\tilde{D}^{1/2})]$ is in a block diagonal form, we have:

$$\Sigma^{-1} = (U \otimes I_{p\times p}) \begin{bmatrix} F_1^{-1} & & \\ & \ddots & \\ & & F_n^{-1} \end{bmatrix} (U^T \otimes I_{p\times p}), \tag{2.14}$$

where $F_l = \delta_l D^{1/2} C_g D^{1/2} + \tilde{D}^{1/2} C_e \tilde{D}^{1/2}$ for $l = 1, 2, ..., n$. So to invert Σ, we only need to invert n $(p \times p)$ symmetric matrices. For one study, the spectral decomposition of K can be reused, which guaranteed the efficiency of our method.

Now consider

$$\Gamma^{1/2}(U \otimes I_{p\times p}) = \Gamma^{1/2} \begin{bmatrix} U_{11} I_{p\times p} & \cdots & U_{1n} I_{p\times p} \\ \vdots & \ddots & \vdots \\ U_{n1} I_{p\times p} & \cdots & U_{nn} I_{p\times p} \end{bmatrix}. \tag{2.15}$$

Set $\tilde{\Gamma}(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij})$. This $\tilde{\Gamma}$ is a function. Then $\Gamma = diag(vec(\tilde{\Gamma}(\mu)))$. Equation (2.15) becomes:

$$\begin{bmatrix} U_{11}\Gamma^{1/2}_{person\ 1} & \cdots & U_{1n}\Gamma^{1/2}_{person\ 1} \\ U_{21}\Gamma^{1/2}_{person\ 2} & \cdots & U_{2n}\Gamma^{1/2}_{person\ 2} \\ \vdots & \ddots & \vdots \\ U_{n1}\Gamma^{1/2}_{person\ n} & \cdots & U_{nn}\Gamma^{1/2}_{person\ n} \end{bmatrix} = \begin{bmatrix} U_{1\cdot} \otimes \Gamma^{1/2}_{person\ 1} \\ U_{2\cdot} \otimes \Gamma^{1/2}_{person\ 2} \\ \vdots \\ U_{n\cdot} \otimes \Gamma^{1/2}_{person\ n} \end{bmatrix}, \tag{2.16}$$

where for $j = 1, ..., n$, $U_{j\cdot}$ denotes the j^{th} row of matrix U and $\Gamma^{1/2}_{person\ j} = diag(vec(\tilde{\Gamma}(\mu_{person\ j})))$ where $\mu_{person\ j} = (\mu_{1j}, \mu_{2j}, ..., \mu_{pj})^T$. Define$(\tilde{X} \otimes I_{p\times p})\Gamma^{1/2}(U \otimes I_{p\times p}) = (*)$ We have:

$$\begin{aligned} (*) &= \begin{bmatrix} \sum_{l=1}^n X_{1l}U_{l1}\Gamma^{1/2}_{person\ l} & \sum_{l=1}^n \tilde{X}_{1l}U_{l2}\Gamma^{1/2}_{person\ l} & \cdots & \sum_{l=1}^n \tilde{X}_{1l}U_{ln}\Gamma^{1/2}_{person\ l} \\ \sum_{l=1}^n \tilde{X}_{2l}U_{l1}\Gamma^{1/2}_{person\ l} & \sum_{l=1}^n \tilde{X}_{2l}U_{l2}\Gamma^{1/2}_{person\ l} & \cdots & \sum_{l=1}^n \tilde{X}_{2l}U_{ln}\Gamma^{1/2}_{person\ l} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{l=1}^n \tilde{X}_{kl}U_{l1}\Gamma^{1/2}_{person\ l} & \sum_{l=1}^n \tilde{X}_{kl}U_{l2}\Gamma^{1/2}_{person\ l} & \cdots & \sum_{l=1}^n \tilde{X}_{kl}U_{ln}\Gamma^{1/2}_{person\ l} \end{bmatrix} \\ &= \begin{bmatrix} \nu_{11} & \nu_{12} & \cdots & \nu_{1n} \\ \nu_{21} & \nu_{22} & \cdots & \nu_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \nu_{k1} & \nu_{k2} & \cdots & \nu_{kn} \end{bmatrix}. \end{aligned} \tag{2.17}$$

The $(i,j)^{th}$ block of $(*)$ is ν_{ij} and $\nu_{ij} = \sum_{l=1}^{n} \tilde{X}_{il} U_{lj} \Gamma^{1/2}_{person\ l}$ is a $p \times p$ matrix. Define $[\sim] = (\tilde{X} \otimes I_{p\times p})\Gamma^{1/2}\Sigma^{-1}\Gamma^{1/2}(\tilde{X}^T \otimes I_{p\times p})$, Then

$$[\sim] = (*) \begin{bmatrix} F_1^{-1} & & \\ & \ddots & \\ & & F_n^{-1} \end{bmatrix} (*)^T. \tag{2.18}$$

So $[\sim]$ has $(i,j)^{th}$ block $\sum_{l=1}^{n} \nu_{il} F_l^{-1} \nu_{jl}$ and $[\sim]$ is a $pk \times pk$ matrix, we need to invert this matrix. Denote $\Gamma^{-1/2} vec(Y) = \tilde{Y}_{vec}$ and $\Gamma^{-1/2} vec(\mu) = \tilde{\mu}_{vec}$, then we have $\Gamma^{-1/2}[vec(Y) - vec(\mu(vec(\beta)))] = \tilde{Y}_{vec} - \tilde{\mu}_{vec}$. Use $\tilde{Y}_{vec}$ to form a $(p \times n)$ matrix as follows: first p elements of $\tilde{Y}_{vec}$ form the first column, second p elements of $\tilde{Y}_{vec}$ form the second column, so on so forth, finally, there will be a $(p \times n)$ matrix, we denote it as $\tilde{Y}$. So $\tilde{Y}$ is a $(p \times n)$ matrix whose j^{th} column consists of elements $p(j-1)+1$ through pj of the vector $\tilde{Y}_{vec}$. Form a $(p \times n)$ matrix $\tilde{\mu}$ use $\tilde{\mu}_{vec}$ as the same way. Then by equation (2.6) we have

$$(U^T \otimes I_{p\times p})\Gamma^{-1/2}(vec(Y) - vec(\mu)) = vec((\tilde{Y} - \tilde{\mu}) \cdot U) = vec(Z), \tag{2.19}$$

where we denote matrix $(\tilde{Y} - \tilde{\mu}) \cdot U$ to be Z, then we have:

$$\begin{bmatrix} F_1^{-1} & & \\ & \ddots & \\ & & F_n^{-1} \end{bmatrix} vec((\tilde{Y} - \tilde{\mu}) \cdot U) = \begin{bmatrix} F_1^{-1} Z_{\cdot 1} \\ F_2^{-1} Z_{\cdot 2} \\ \vdots \\ F_n^{-1} Z_{\cdot n} \end{bmatrix}, \tag{2.20}$$

where $Z_{\cdot i}$ is the i^{th} column of Z. From equations (2.17) and (2.20) we have:

$$
\begin{aligned}
&(X \otimes I_{p\times p})\Gamma^{1/2}(U \otimes I_{p\times p}) \begin{bmatrix} F_1^{-1} & & \\ & \ddots & \\ & & F_n^{-1} \end{bmatrix} (U^T \otimes I_{p\times p})\Gamma^{-1/2}\big(vec(Y) - vec(\mu)\big) \\
&= \begin{bmatrix} \nu_{11} & \cdots & \nu_{1n} \\ \vdots & \ddots & \vdots \\ \nu_{k1} & \cdots & \nu_{kn} \end{bmatrix} \begin{bmatrix} F_1^{-1} Z_{\cdot 1} \\ \vdots \\ F_n^{-1} Z_{\cdot n} \end{bmatrix} \\
&= \begin{bmatrix} \sum_{l=1}^n \nu_{1l} F_l^{-1} Z_{\cdot l} \\ \vdots \\ \sum_{l=1}^n \nu_{kl} F_l^{-1} Z_{\cdot l} \end{bmatrix}.
\end{aligned}
\tag{2.21}
$$

Equations (2.18) and (2.21) will help solve equation (2.11).

2.3 Parameter Estimation for Variance Components

2.3.1 Estimating Equations for Variance Components

Variance components estimating equations are motivated by maximum likelihood estimators of variance components in the case when $\tilde{\beta}$ is known, assuming a multivariate normal distribution for $vec(Y)$ with the same mean and covariance as the quasi-likelihood model. Note $E(vec(Y)|X, G) = vec(\mu)$ and $Var(vec(Y)|X, G) = \Gamma^{1/2}\Sigma\Gamma^{1/2}$. We denoted the variance components parameters to be Θ. So we have the "log-likelihood function" $l(\Theta; vec(Y))$ for Θ as follows:

$$
\begin{aligned}
l(\Theta; \mathrm{vec}(Y)) = &- 0.5np \cdot \log(2\pi) - 0.5 \log |\Gamma^{1/2}\Sigma\Gamma^{1/2}| \\
&- 0.5(\mathrm{vec}(Y) - \mathrm{vec}(\mu))^T (\Gamma^{1/2}\Sigma\Gamma^{1/2})^{-1}(\mathrm{vec}(Y) - \mathrm{vec}(\mu)).
\end{aligned}
\tag{2.22}
$$

So the system of estimating equations are calculated from:

$$\frac{\partial l(\Theta; \mathrm{vec}(Y))}{\partial D_{ii}} = 0, for\ i = 1, ..., p \tag{2.23}$$

$$\frac{\partial l(\Theta; \mathrm{vec}(Y))}{\partial Cg_{ij}} = 0, for\ i < j, i = 1, ..., p-1 \tag{2.24}$$

$$\frac{\partial l(\Theta; \mathrm{vec}(Y))}{\partial Ce_{ij}} = 0, for\ i < j, i = 1, ..., p-1 \tag{2.25}$$

We obtain the following set of estimating equations for the variance where u_j denotes the vector of length p with the j-th element equal to 1 and all other elements 0:

$$\mathrm{vec}(Y-\mu)^T \Gamma^{-1/2} \Sigma^{-1} \Psi \Sigma^{-1} \Gamma^{-1/2} \mathrm{vec}(Y-\mu) - \mathrm{tr}(\Sigma^{-1}\Psi) = 0, \quad \text{for } 1 \leqslant i < j \leqslant p, \tag{2.26}$$

We consider three choices for the matrix Ψ, each corresponding to estimating equations for specific variance components:

1. Estimating equations for lower-off diagonal elements of C_g:

$$\Psi = K \otimes [d_i^{1/2} d_j^{1/2} (u_i u_j^T + u_j u_i^T)]$$

2. Estimating equations for lower-off diagonal elements of C_e:

$$\Psi = I_{n\times n} \otimes [(1-d_i)^{1/2}(1-d_j)^{1/2}(u_i u_j^T + u_j u_i^T)]$$

3. Estimating equations for diagonal elements D:

$$\Psi = K \otimes S_j^g + I_{n\times n} \otimes S_j^e$$

where S_j^g is the $p \times p$ matrix with (i,l)th element:

$$S_j^g[i,l] = \begin{cases} 0, & \text{if } i \neq j \text{ and } l \neq j, \\ \frac{1}{2} d_i^{1/2} d_j^{-1/2} C_g[i,j], & \text{if } i \neq j \text{ and } l = j, \\ \frac{1}{2} d_l^{1/2} d_j^{-1/2} C_g[l,j], & \text{if } i = j \text{ and } l \neq j, \\ 1, & \text{if } i = j \text{ and } l = j, \end{cases}$$

and S_j^e is the $p \times p$ matrix with (i,l)th element:

$$S_j^e[i,l] = \begin{cases} 0, & \text{if } i \neq j \text{ and } l \neq j, \\ -\frac{1}{2}(1-d_i)^{1/2}(1-d_j)^{-1/2} C_e[i,j], & \text{if } i \neq j \text{ and } l = j, \\ -\frac{1}{2}(1-d_l)^{1/2}(1-d_j)^{-1/2} C_e[l,j], & \text{if } i = j \text{ and } l \neq j, \\ -1, & \text{if } i = j \text{ and } l = j. \end{cases}$$

(2.26) together with the estimating equation for $vec(\tilde{\beta})$ (equation (2.9)), we have a system of estimating equations for $vec(\tilde{\beta})$ and variance components. We need to solve the system of estimating equations. To estimate the parameters, we follow an iterative procedure. For fixed values of the coefficients, we estimate the variance components. Then, using the estimated variance components, we fix their values and estimate the coefficients. This process is repeated iteratively, alternating between estimating coefficients and variance components, until the estimating equations for both are satisfied.

2.3.2 Problem of Log-Likelihood Maximization

For fixed value of coefficients, we use an EM algorithm incorporating the Newton-Raphson algorithm to solve the problem of log-likelihood maximization for estimating variance components. Recall $E(\Gamma^{-1/2}\text{vec}(Y) \mid X, G) = \Gamma^{-1/2}\text{vec}(\mu)$ and $\text{Var}(\Gamma^{-1/2}\text{vec}(Y) \mid X, G) = \Sigma$.

We denote $\Gamma^{-1/2}\text{vec}(Y) = \tilde{Y}_{\text{vec}}$, $\Gamma^{-1/2}\text{vec}(\mu) = \tilde{\mu}_{\text{vec}}$, and let

$$\tilde{\tilde{Y}} = \text{unvec}(\tilde{Y}_{\text{vec}} - \tilde{\mu}_{\text{vec}}, p, n),$$

that is, a $p \times n$ matrix created by reshaping the vector $\tilde{Y}_{\text{vec}} - \tilde{\mu}_{\text{vec}}$ of length pn. The entries of $\tilde{\tilde{Y}}$ are filled column-wise from the vector. If we assume a multivariate normal distribution of $\tilde{Y}_{\text{vec}} - \tilde{\mu}_{\text{vec}}$ with mean zero and variance Σ (note we do not assume the real data follows this distribution, we only make this assumption for the purpose of paramter estimations), and based on the multivariate linear mixed model in [15], we have

$$\tilde{\tilde{Y}} = \tilde{\tilde{G}}+\tilde{\tilde{E}}; \quad \tilde{\tilde{G}} \sim MN_{p\times n}\left(0, D^{\frac{1}{2}}C_gD^{\frac{1}{2}}, K\right), \quad \tilde{\tilde{E}} \sim MN_{p\times n}\left(0, \tilde{D}^{\frac{1}{2}}C_e\tilde{D}^{\frac{1}{2}}, I_{n\times n}\right). \tag{2.27}$$

Where $\tilde{\tilde{Y}}$ is a p by n transformation of residuals of the phenotype after accounting for the effects of covariates. $\tilde{\tilde{G}}$ is a p by n random effect matrix, $\tilde{\tilde{E}}$ is a p by n residual errors. $MN_{p\times n}(0, V_1, V_2)$ is a p by n matrix normal distribution with mean 0, row covariance p by p matrix V_1 and column covariance n by n matrix V_2. Note we do not assume the data follow the distribution in (2.27), we just use it to calculate the estimation equations of the variance components. Following the calculation in [27, 28, 29], using the spectral decomposition of K we used before ($K = UVU^T$), we obtained new transformation of residuals of the phenotype after accounting for the effects of covariates $\tilde{\tilde{\tilde{Y}}} = \tilde{\tilde{Y}}U$, random effects $\tilde{\tilde{\tilde{G}}} = \tilde{\tilde{G}}U$ and residual errors $\tilde{\tilde{\tilde{E}}} = \tilde{\tilde{E}}U$ and we will have

$$\tilde{\tilde{\tilde{Y}}} = \tilde{\tilde{\tilde{G}}} + \tilde{\tilde{\tilde{E}}}; \quad \tilde{\tilde{\tilde{G}}} \sim MN_{p\times n}\left(0, D^{\frac{1}{2}}C_gD^{\frac{1}{2}}, V\right), \quad \tilde{\tilde{\tilde{E}}} \sim MN_{p\times n}\left(0, \tilde{D}^{\frac{1}{2}}C_e\tilde{D}^{\frac{1}{2}}, I_{n\times n}\right) \tag{2.28}$$

and this is equivalent to

$$y = g + e; \quad g \sim MVN(0, V \otimes (D^{\frac{1}{2}}C_gD^{\frac{1}{2}})), \quad e \sim MVN(0, I_{n\times n} \otimes \tilde{D}^{\frac{1}{2}}C_e\tilde{D}^{\frac{1}{2}}) \tag{2.29}$$

where $y = vec(\tilde{\tilde{\tilde{Y}}})$, $g = vec(\tilde{\tilde{\tilde{G}}})$ and $e = vec(\tilde{\tilde{\tilde{E}}})$ and MVN denotes multivariate normal distribution. Therefore, for each individual l, the new transformation of residuals of the phenotype after accounting for the effects of covariates is assumed to follow independent but not identical multivariate normal distribution to calculate the estimating equations:

$$y_l = g_l + e_l; \quad g_l \sim MVN(0, \delta_l D^{\frac{1}{2}} C_g D^{\frac{1}{2}}), \quad e_l \sim MVN(0, \tilde{D}^{\frac{1}{2}} C_e \tilde{D}^{\frac{1}{2}}). \tag{2.30}$$

The variance for l^{th} individual is $V_l = \delta_l D^{\frac{1}{2}} C_g D^{\frac{1}{2}} + \tilde{D}^{\frac{1}{2}} C_e \tilde{D}^{\frac{1}{2}}$ and y_l is the l^{th} column vector of $\tilde{\tilde{\tilde{Y}}}$, g_l is the l^{th} column vector of $\tilde{\tilde{\tilde{G}}}$, and e_l is the l^{th} column vector of $\tilde{\tilde{\tilde{E}}}$, for $\forall l = 1, ..., n$. Based on (2.29), we have the incomplete data log-likelihood function as follows:

$$\begin{aligned} \log \ell(\tilde{\tilde{\tilde{Y}}} \mid D, C_g, C_e) = \sum_{\ell=1}^{n} \Bigg[& -\frac{p}{2} \log(2\pi) - \frac{1}{2} \log |\tilde{D}^{\frac{1}{2}} C_e \tilde{D}^{\frac{1}{2}} + \delta_l D^{\frac{1}{2}} C_g D^{\frac{1}{2}}| \\ & - \frac{1}{2} y_\ell^T (\tilde{D}^{\frac{1}{2}} C_e \tilde{D}^{\frac{1}{2}} + \delta_l D^{\frac{1}{2}} C_g D^{\frac{1}{2}})^{-1} y_\ell \Bigg]. \end{aligned} \tag{2.31}$$

Based on (2.30), we view $\tilde{\tilde{\tilde{G}}}$ as missing values, then we can get the complete data log-likelihood function as follows:

$$\begin{aligned} \log \ell(\tilde{\tilde{\tilde{Y}}}, \tilde{\tilde{\tilde{G}}} \mid D, C_g, C_e) = \sum_{\ell=1}^{n} \Bigg[& - p \log(2\pi) - \frac{1}{2} \log |\tilde{D}^{\frac{1}{2}} C_e \tilde{D}^{\frac{1}{2}}| - \frac{1}{2} \log |\delta_l D^{\frac{1}{2}} C_g D^{\frac{1}{2}}| \\ & - \frac{1}{2} e_\ell^T (\tilde{D}^{\frac{1}{2}} C_e \tilde{D}^{\frac{1}{2}})^{-1} e_\ell - \frac{1}{2} g_\ell^T (\delta_l D^{\frac{1}{2}} C_g D^{\frac{1}{2}})^{-1} g_\ell \Bigg]. \end{aligned} \tag{2.32}$$

We can rewrite (2.32) as follows:

$$\log \ell(\tilde{\tilde{\tilde{Y}}}, \tilde{\tilde{\tilde{G}}} \mid D, C_g, C_e) = \log \ell_1(\tilde{\tilde{\tilde{Y}}}, \tilde{\tilde{\tilde{G}}} \mid D, C_g) + \log \ell_2(\tilde{\tilde{\tilde{Y}}}, \tilde{\tilde{\tilde{G}}} \mid D, C_e) \tag{2.33}$$

where the components are defined as:

$$\begin{aligned}\log \ell_1(\tilde{\tilde{\tilde{Y}}}, \tilde{\tilde{\tilde{G}}} \mid D, C_g) = &- \frac{np}{2} \log(2\pi) - n \log |D^{\frac{1}{2}}| - \frac{p}{2} \sum_{\ell=1}^{n} \log \delta_\ell - \frac{n}{2} \log |C_g| \\ &- \frac{1}{2} \mathrm{Tr} \left(T_1 (D^{\frac{1}{2}} C_g D^{\frac{1}{2}})^{-1} \right)\end{aligned} \tag{2.34}$$

and

$$\begin{aligned}\log \ell_2(\tilde{\tilde{\tilde{Y}}}, \tilde{\tilde{\tilde{G}}} \mid D, C_e) = &- \frac{np}{2} \log(2\pi) - n \log |\tilde{D}^{\frac{1}{2}}| - \frac{n}{2} \log |C_e| \\ &- \frac{1}{2} \mathrm{Tr} \left(T_2 (\tilde{D}^{\frac{1}{2}} C_e \tilde{D}^{\frac{1}{2}})^{-1} \right).\end{aligned} \tag{2.35}$$

Here $T_1 = \sum_{\ell=1}^{n} \delta_\ell^{-1} g_\ell g_\ell^T$, $T_2 = \sum_{\ell=1}^{n} e_\ell e_\ell^T$ are the sufficient statistics.

Maximizing the log-likelihood defined in (2.22) is equivalent to maximizing the log-likelihood defined in (2.31). This optimization problem can be effectively solved using an EM algorithm combined with the Newton-Raphson algorithm. Specifically, the E-step involves computing the expected value of the complete data log-likelihood function defined in (2.33) with respect to the conditional distribution of $\tilde{\tilde{\tilde{G}}}$ given $\tilde{\tilde{\tilde{Y}}}$ and current values of variance components, while the M-step uses the Newton-Raphson algorithm to maximize this expected value. Since the complete data log-likelihood function can be expressed in terms of sufficient statistics, we need to compute the expected values of these sufficient statistics to obtain the expected value of the complete data log-likelihood function. The conditional distribution of $\tilde{\tilde{\tilde{G}}}$ given $\tilde{\tilde{\tilde{Y}}}$ and current values of variance components follows

$$g_\ell \mid y_\ell, D, C_g, C_e \sim \mathrm{MVN}(\hat{g}_\ell, \hat{\Sigma}_\ell), \tag{2.36}$$

$$\hat{g}_\ell = \delta_\ell D^{\frac{1}{2}} C_g D^{\frac{1}{2}} \left(\delta_\ell D^{\frac{1}{2}} C_g D^{\frac{1}{2}} + \tilde{D}^{\frac{1}{2}} C_e \tilde{D}^{\frac{1}{2}} \right)^{-1} y_\ell, \tag{2.37}$$

$$\hat{\Sigma}_\ell = \delta_\ell D^{\frac{1}{2}} C_g D^{\frac{1}{2}} \left(\delta_\ell D^{\frac{1}{2}} C_g D^{\frac{1}{2}} + \tilde{D}^{\frac{1}{2}} C_e \tilde{D}^{\frac{1}{2}}\right)^{-1} \tilde{D}^{\frac{1}{2}} C_e \tilde{D}^{\frac{1}{2}}. \tag{2.38}$$

Therefore we have:

$$\begin{aligned} E(T_1 \mid y_\ell, D, C_g, C_e) &= \sum_{\ell=1}^{n} \delta_\ell^{-1} E(g_\ell g_\ell^T \mid y_\ell, D, C_g, C_e) \\ &= \sum_{\ell=1}^{n} \delta_\ell^{-1} (\hat{g}_\ell \hat{g}_\ell^T + \hat{\Sigma}_\ell). \end{aligned} \tag{2.39}$$

$$\begin{aligned} E(T_2 \mid y_\ell, D, C_g, C_e) &= \sum_{\ell=1}^{n} E(e_\ell e_\ell^T \mid y_\ell, D, C_g, C_e) \\ &= \sum_{\ell=1}^{n} E((y_\ell - g_\ell)(y_\ell - g_\ell)^T \mid y_\ell, D, C_g, C_e) \\ &= \sum_{\ell=1}^{n} y_\ell y_\ell^T - 2\hat{g}_\ell y_\ell^T + \hat{g}_\ell \hat{g}_\ell^T + \hat{\Sigma}_\ell. \end{aligned} \tag{2.40}$$

For simplicity we denote $E(T_1 \mid y_\ell, D, C_g, C_e) = S_1$ and $E(T_2 \mid y_\ell, D, C_g, C_e) = S_2$. Then we have the expected value of the complete data log-likelihood function defined in (2.33) with respect to the conditional distribution of $\tilde{\tilde{\tilde{G}}}$ given $\tilde{\tilde{\tilde{Y}}}$ and current values of variance components as follows:

$$\begin{aligned} &E_{\tilde{\tilde{\tilde{G}}}|\tilde{\tilde{\tilde{Y}}},D,C_g,C_e}\left[\log \ell(\tilde{\tilde{\tilde{Y}}}, \tilde{\tilde{\tilde{G}}} \mid D, C_g, C_e)\right] \\ &\qquad = E_{\tilde{\tilde{\tilde{G}}}|\tilde{\tilde{\tilde{Y}}},D,C_g}\left[\log \ell_1(\tilde{\tilde{\tilde{Y}}}, \tilde{\tilde{\tilde{G}}} \mid D, C_g)\right] \\ &\qquad\quad + E_{\tilde{\tilde{\tilde{G}}}|\tilde{\tilde{\tilde{Y}}},D,C_e}\left[\log \ell_2(\tilde{\tilde{\tilde{Y}}}, \tilde{\tilde{\tilde{G}}} \mid D, C_e)\right]. \end{aligned} \tag{2.41}$$

where the components are defined as:

$$\begin{aligned} E_{\tilde{\tilde{\tilde{G}}}|\tilde{\tilde{\tilde{Y}}},D,C_g}\left[\log \ell_1(\tilde{\tilde{\tilde{Y}}}, \tilde{\tilde{\tilde{G}}} \mid D, C_g)\right] &= -\frac{np}{2}\log(2\pi) - n \log|D^{\frac{1}{2}}| - \frac{p}{2}\sum_{\ell=1}^{n} \log \delta_\ell - \frac{n}{2}\log|C_g| \\ &\quad - \frac{1}{2}\mathrm{Tr}\left(S_1 (D^{\frac{1}{2}} C_g D^{\frac{1}{2}})^{-1}\right). \end{aligned} \tag{2.42}$$

and

$$E_{\tilde{\tilde{\tilde{G}}}|\tilde{\tilde{\tilde{Y}}},D,C_e}\left[\log \ell_2(\tilde{\tilde{\tilde{Y}}},\tilde{\tilde{\tilde{G}}} \mid D, C_e)\right] = -\frac{np}{2}\log(2\pi) - n\log|\tilde{D}^{\frac{1}{2}}| - \frac{n}{2}\log|C_e| - \frac{1}{2}\mathrm{Tr}\left(S_2(\tilde{D}^{\frac{1}{2}}C_e\tilde{D}^{\frac{1}{2}})^{-1}\right). \tag{2.43}$$

We need to maximize (2.41) at each iteration. This is a constraint maximum likelihood problem, since we need the correlation matrices to be positive (semi) definite and the elements of the D matrix must be restricted to the range (0, 1). To relieve the constraint problem of the elements of the D matrix, we take the following parametrization:

$$a_{ii} = log(-log(D_{ii})), for\ i = 1, ..., p \tag{2.44}$$

This parameterization maps values from the interval $(0, 1)$ to the entire real line, $\mathbb{R}$. And the target function for elements in D matrix will be:

$$\frac{\partial E_{\tilde{\tilde{\tilde{G}}}|\tilde{\tilde{\tilde{Y}}},D,C_g,C_e}\left[\log \ell(\tilde{\tilde{\tilde{Y}}},\tilde{\tilde{\tilde{G}}} \mid D, C_g, C_e)\right]}{\partial a_{ii}} = 0, \quad \text{for } i = 1, \ldots, p. \tag{2.45}$$

After we get the values of a_{ii}, we can transfer them back to D_{ii} using the following equation:

$$D_{ii} = exp(-exp(a_{ii})) \tag{2.46}$$

Parametrization for correlation matrix

Under complete data setting, the MLE for covariance matrix has a closed form, which is automatically positive (semi) definite. However, in the same setting, the MLE of the correlation matrix does not have a closed form (diagonal elements are constraint to be 1), and can only be obtained numerically. For any kind of search algorithm, one need to find a way to propose the next step of the search so that the correlation matrix comes out to be positive definite. Hence as part of estimating the correlation matrices (C_g and C_e), we explore one recently proposed parameterization which enforce the positive (semi) definiteness and the parameters are unrestricted. With the parameterization we can release the constrained problem to an unconstrained one.

Archakov & Hansen recently proposed a novel parameterization of the correlation matrix [30]. for a non-singular correlation matrix, C, the new parameterization of correlation matrix is :

$$\gamma(C) := vecl(logC) \tag{2.47}$$

where vecl(logC) denotes the vectorization operator of the lower off-diagonal elements of logC (the matrix logarithm of C). They showed mapping from C to γ is one-to-one hence the set of $n \times n$ non-singular correlation matrices is isomorphic with $\mathbb{R}^{n(n-1)/2}$ and they propose a fast algorithm for the computation of the inverse mapping. They showed the finite sample distribution of the vector $\gamma(\hat{C})$ is well approximated by a Gaussian distribution under standard regularity conditions with weakly correlated elements. The mapping, $\gamma(C)$ is invariant to the reordering of variables that define C, in the sense that a permutation of the variables that define C will merely result in permutation of the element of γ. This parameterization ensures positive definiteness without imposing additional restrictions and can be viewed as a multivariate generalization of Fisher's Z-transformation of a single correlation. One important proposition of this parameterization is the derivatives of the correlation matrix C with respect to the off-diagonal elements of the log-transformed correlation matrix G=logC have relatively simple expression and are:

$$\frac{dvecl[C]}{dvecl[G]} = E_l(I - AE_d^T(E_dAE_d^T)^{-1}E_d)A(E_l + E_u)^T \tag{2.48}$$

where $A = \frac{dvecC}{dvecG}$ and the matrices E_l, E_u and E_d are elimination matrices such that $veclM = E_l vecM$, $veclM^T = E_u vecM$ and $diagM = E_d vecM$ for any square matrix M of the same size as C and G. This proposition is really useful when we use this parameterization to estimate correlation matrices. Let $G_g = log(C_g)$ and $G_e = log(C_e)$. The target functions for elements in C_g and C_e are as follows:

$$\frac{\partial E_{\tilde{\tilde{\tilde{\tilde{G}}}}|\tilde{\tilde{\tilde{\tilde{Y}}}},D,C_g,C_e}\left[\log \ell(\tilde{\tilde{\tilde{\tilde{Y}}}}, \tilde{\tilde{\tilde{\tilde{G}}}} \mid D, C_g, C_e)\right]}{\partial \mathrm{vecl}(G_g)} = 0. \tag{2.49}$$

$$\frac{\partial E_{\tilde{\tilde{\tilde{\tilde{G}}}}|\tilde{\tilde{\tilde{\tilde{Y}}}},D,C_g,C_e}\left[\log \ell(\tilde{\tilde{\tilde{\tilde{Y}}}}, \tilde{\tilde{\tilde{\tilde{G}}}} \mid D, C_g, C_e)\right]}{\partial \mathrm{vecl}(G_e)} = 0. \tag{2.50}$$

And after we get the values of $vecl(G_e)$ and $vecl(G_g)$, we can get reconstructed estimations of C_g and C_e by applying the fast inverse mapping algorithm proposed by Archakov & Hansen [30]:

Algorithm: Inverse Mapping of a Vector vecl(G) to a Correlation Matrix C

1. Initialize the Matrix G:

1. Create a $p \times p$ zero matrix G.

2. Populate the lower triangular part (excluding diagonal) of G using $veclG$.

3. Add the transpose of G to itself to make it symmetric.

2. Initialize Variables:

1. Set $dist \leftarrow \sqrt{p}$.

2. Extract the diagonal elements of G into $diag\_vec$:

$$diag\_vec \leftarrow diag(G)$$

where $diag(G)$ extracts the diagonal elements of G.

3. Iterative Convergence Loop:

While $dist > \sqrt{p} \cdot tol\_value$, perform the following steps:

(a) Compute the matrix exponential of G:

$$exp(G)$$

(b) Extract the diagonal elements of $exp(G)$:

$$diag\_exp \leftarrow diag(exp(G))$$

(c) Compute the element-wise logarithm of $diag\_exp$:

$$diag\_delta \leftarrow \log(diag\_exp)$$

(d) Update the diagonal adjustment vector:

$$diag\_vec \leftarrow diag\_vec - diag\_delta$$

(e) Update G by replacing its diagonal elements with the adjusted diagonal values:

$$diag(G) \leftarrow diag\_vec$$

where $diag(G)$ sets the diagonal of G to $diag\_vec$.

(f) Compute the norm of $diag\_delta$ and set it as dist:

$$dist \leftarrow \|diag\_delta\|$$

4. Compute Output Matrix C:

1. Compute the matrix exponential of G:

$$C \leftarrow exp(G)$$

2. Set all diagonal elements of C to 1:

$$diag(C) \leftarrow 1$$

where $diag(C)$ sets the diagonal of C to 1.

5. Return C:

1. Return the resulting matrix C.

It is showed that the resulting C is a correlation matrix in [30].

2.3.3 EM Algorithm with Inner Newton-Raphson Iterations for M-step

We denote the set of parameters $a_{ii}, i = 1, \ldots, p$, $\text{vecl}(G_g)$, and $\text{vecl}(G_e)$ collectively as ξ. The EM algorithm alternates between updating ξ and transforming these updates back to the original parameters (D, C_g, C_e) for likelihood evaluations. The EM algorithm alternates between the following steps:

E-Step (Outer Iteration t): In the t-th EM iteration, compute the expected log-likelihood of the complete data, as defined in (2.41):

$$Q(\xi \mid \xi^{(t)}) = E_{\tilde{\tilde{\tilde{G}}}|\tilde{\tilde{\tilde{Y}}},D^{(t)},C_g^{(t)},C_e^{(t)}}\big[\log \ell(\tilde{\tilde{\tilde{Y}}}, \tilde{\tilde{\tilde{G}}} \mid D^{(t)}, C_g^{(t)}, C_e^{(t)})\big].$$

Here, $\xi^{(t)}$ is the current parameter estimate from the t-th EM iteration.

M-Step (Inner Iteration k): To maximize $Q(\xi \mid \xi^{(t)})$, we use the Newton-Raphson algorithm. For the inner iterations indexed by k:

1. Set an independent starting value for Newton-Raphson, denoted as $\xi^{(t,0)}$

2. Perform updates as follows:

$$\xi^{(t,k+1)} = \xi^{(t,k)} - \mathbf{H}^{-1}(\xi^{(t,k)})\nabla(\xi^{(t,k)}),$$

   where:

   - $\xi^{(t,k)}$ is the parameter estimate at the k-th NR iteration within the t-th EM iteration,
   - $\nabla(\xi^{(t,k)}) = \frac{\partial}{\partial \xi} Q(\xi \mid \xi^{(t,k)})$ is the gradient of $Q(\xi \mid \xi^{(t,k)})$,
   - $\mathbf{H}(\xi^{(t,k)}) = \frac{\partial^2}{\partial \xi^2} Q(\xi \mid \xi^{(t,k)})$ is the Hessian matrix of $Q(\xi \mid \xi^{(t,k)})$.

Gradient: The gradient $\nabla(\xi)$ is computed as:

$$\nabla(\xi) = \frac{\partial}{\partial \xi} E_{\tilde{\tilde{\tilde{G}}}|\tilde{\tilde{\tilde{Y}}},D,C_g,C_e}\big[\log \ell(\tilde{\tilde{\tilde{Y}}}, \tilde{\tilde{\tilde{G}}} \mid D, C_g, C_e)\big].$$

Hessian: The Hessian $\mathbf{H}(\xi)$ is computed as:

$$\mathbf{H}(\xi) = \frac{\partial^2}{\partial \xi^2} E_{\tilde{\tilde{\tilde{G}}}|\tilde{\tilde{\tilde{Y}}},D,C_g,C_e}\big[\log \ell(\tilde{\tilde{\tilde{Y}}}, \tilde{\tilde{\tilde{G}}} \mid D, C_g, C_e)\big].$$

Convergence of Newton-Raphson (k): The inner Newton-Raphson iterations terminate when the change in $Q(\xi \mid \xi^{(t,k)})$ satisfies:

$$|Q(\xi \mid \xi^{(t,k+1)}) - Q(\xi \mid \xi^{(t,k)})| < \text{tolerance}.$$

The result of the inner loop, $\xi^{(t+1)} = \xi^{(t,k_{\text{final}})}$, is used for the next EM iteration.

Choice of Starting Value $\xi^{(t,0)}$: The starting value $\xi^{(t,0)}$ for the Newton-Raphson iterations is derived as follows: Recall the sufficient statistics:

$$S_1 = E(T_1 \mid \tilde{\tilde{\tilde{Y}}}, D^{(t)}, C_g^{(t)}, C_e^{(t)}) = E\left(\sum_{\ell=1}^{n} \delta_\ell^{-1} g_\ell g_\ell^T \mid \tilde{\tilde{\tilde{Y}}}, D^{(t)}, C_g^{(t)}, C_e^{(t)}\right),$$

$$S_2 = E(T_2 \mid \tilde{\tilde{\tilde{Y}}}, D^{(t)}, C_g^{(t)}, C_e^{(t)}) = E\left(\sum_{\ell=1}^{n} e_\ell e_\ell^T \mid \tilde{\tilde{\tilde{Y}}}, D^{(t)}, C_g^{(t)}, C_e^{(t)}\right).$$

Under the assumption of the distributions of g_ℓ and e_ℓ as defined in (2.30), the expected values of S_1 and S_2 are:

$$E(S_1) = nD^{\frac{1}{2}} C_g D^{\frac{1}{2}}, \quad E(S_2) = n\tilde{D}^{\frac{1}{2}} C_e \tilde{D}^{\frac{1}{2}}.$$

Thus, the initial values of the variance components are chosen as follows for $1 \leqslant i, j \leqslant p$:

$$C_g^{(t,0)}[i,j] = \frac{S_1^{(t)}[i,j]}{\sqrt{S_1^{(t)}[i,i] S_1^{(t)}[j,j]}},$$

$$C_e^{(t,0)}[i,j] = \frac{S_2^{(t)}[i,j]}{\sqrt{S_2^{(t)}[i,i] S_2^{(t)}[j,j]}},$$

$$D^{(t,0)}[i,i] = \frac{S_1^{(t)}[i,i]}{S_1^{(t)}[i,i] + S_2^{(t)}[i,i]}.$$

Where $S_1^{(t)}$ and $S_2^{(t)}$ are computed in the E-step. These values are then transformed into the unconstrained parameter set $\xi^{(t,0)}$ using the previously defined parametrization rules.

Convergence of EM Algorithm (t): The EM algorithm stops when the change in the incomplete log-likelihood defined in (2.31) satisfies:

$$|\log \ell(\tilde{\tilde{\tilde{Y}}} \mid D^{(t+1)}, C_g^{(t+1)}, C_e^{(t+1)}) - \log \ell(\tilde{\tilde{\tilde{Y}}} \mid D^{(t)}, C_g^{(t)}, C_e^{(t)})| < \text{tolerance}.$$

Iterative Procedure:

1. **Initialization:** Start with initial parameter estimates $C_g^{(0)}, C_e^{(0)}, D^{(0)}$. These are transformed to $\xi^{(0)}$ using the parametrization defined earlier.

   Choice of $C_g^{(0)}$ and $C_e^{(0)}$: The matrices $C_g^{(0)}$ and $C_e^{(0)}$ are initialized as the sample correlation matrix calculated across the rows of $\tilde{\tilde{Y}}$.

   Choice of $D^{(0)}$: Each diagonal element $d_{ii}^{(0)}$ of $D^{(0)}$ is determined by fitting a linear mixed model (LMM) for each row of $\tilde{\tilde{Y}}$. The steps are as follows:

   (a) **Model Definition:** For the i-th row of $\tilde{\tilde{Y}}$, denoted $\tilde{\tilde{Y}}_i$, fit the LMM:

   $$\tilde{\tilde{Y}}_{i\ell} = \mu + g_{i\ell} + e_{i\ell}, \quad \ell = 1, \dots, n,$$

   where:

   - μ is the overall mean,
   - $g_{i\ell} \sim \mathcal{N}(0, \sigma_g^2)$ is the genetic effect (random effect),
   - $e_{i\ell} \sim \mathcal{N}(0, \sigma_e^2)$ is the residual effect.

   (b) **Variance Decomposition:** Estimate the variance components σ_g^2 and σ_e^2 by optimizing the profile log-likelihood over the heritability h^2, defined as:

   $$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}.$$

   (c) **Heritability Calculation:** Use the maximized h^2 from the profile log-likelihood as the estimate for heritability of the i-th trait.

   (d) **Diagonal Element:** Set $d_{ii}^{(0)} = h_i^2$, where h_i^2 is the estimated heritability for the i-th trait.

2. **Iterative Updates:** For each $t = 0, 1, 2, \ldots$, perform:

   (a) **E-Step:** Compute $Q(\xi \mid \xi^{(t)})$ using the current parameter estimates $\xi^{(t)}$.

   (b) **M-Step:** Perform Newton-Raphson (NR) iterations starting from $\xi^{(t,0)}$:

   $$\xi^{(t+1)} = \xi^{(t,k_{\text{final}})},$$

   where k_{final} is the last NR iteration that satisfies the M-step convergence criterion.

3. **Termination:** Stop the EM algorithm when the change in the incomplete log-likelihood satisfies the EM convergence criterion. We get $\xi^{(t_{final})}$ and transform them back to get our estimates of $\hat{D}$, $\hat{C}_g$ and $\hat{C}_e$.

2.4 Evaluating the Estimation of Variance Components

2.4.1 Procedures Used to Evaluate the Estimation of Variance Components

Recall that the variance components include the diagonal elements of the diagonal matrix D and the lower off-diagonal elements of the correlation matrices C_g and C_e, collectively denoted as Θ. To evaluate the performance of the EM algorithm incorporating the Newton-Raphson algorithm, as proposed in Section 2.3.3, we compute the Fisher information for this problem. Specifically, we analyze the scenario where the data are assumed to follow the distribution defined in (2.30). It is important to emphasize that this assumption is made solely for the purpose of evaluating the performance of the proposed procedure. We do not assume that real data necessarily follow this distribution; the assumption is used only to facilitate estimation and to assess the reliability of the parameter estimates. Recall we assume $y_1, \ldots, y_n$ are independently distributed as

$$y_\ell \sim \text{MVN}(0, \Sigma_\ell(\Theta)),$$

where

$$\Sigma_\ell(\Theta) = \delta_\ell D^{\frac{1}{2}} C_g D^{\frac{1}{2}} + \tilde{D}^{\frac{1}{2}} C_e \tilde{D}^{\frac{1}{2}}.$$

According to the chain rule of Fisher information and the formula for Fisher information for the multivariate normal distribution in [31], we have:

$$\mathcal{I}(\Theta) = \sum_{\ell=1}^{n} \mathcal{I}_{y_\ell}(\Theta), \tag{2.51}$$

where

$$\mathcal{I}_{y_\ell}(\Theta)_{m,n} = \frac{1}{2} \mathrm{Tr} \left(\Sigma_\ell(\Theta)^{-1} \frac{\partial \Sigma_\ell(\Theta)}{\partial \Theta_m} \Sigma_\ell(\Theta)^{-1} \frac{\partial \Sigma_\ell(\Theta)}{\partial \Theta_n} \right), \quad 1 \leqslant m, n \leqslant p^2. \tag{2.52}$$

The Fisher information $\mathcal{I}(\Theta)$ is a $p^2 \times p^2$ matrix and is for the entire sample of n individuals. Standard theory tells us that the maximum likelihood estimator (MLE), $\hat{\Theta}_{\text{MLE}}$, has an approximate distribution:

$$\hat{\Theta}_{\text{MLE}} \sim N(\Theta, \mathcal{I}(\Theta)^{-1}),$$

for large n, under regularity conditions sufficient for a central limit theorem.

We perform multiple simulation replicates, denoted by n_{reps}, where the parameter Θ is kept fixed across all replicates. Consequently, $\mathcal{I}(\Theta)$ also remains fixed. After performing the simulations, we obtain $\hat{\Theta}_1, \ldots, \hat{\Theta}_{n_{\text{reps}}}$. For each parameter, we use the `qqconf` R package, developed by Weine, McPeek and Abney [32], to create QQ-plots.

For example, consider the second element of Θ, denoted as Θ_2. The corresponding estimated values across n_{reps} replicates, $\hat{\Theta}_{.2} = (\hat{\Theta}_{12}, \ldots, \hat{\Theta}_{n_{\text{reps}}2})^T$, form an independent and identically distributed (iid) sample of size n_{reps} from $N(\Theta_2, [\mathcal{I}(\Theta)^{-1}]_{2,2})$. The QQ-plot provides a shaded simultaneous acceptance region to test whether the observed values (e.g., $\hat{\Theta}_{.2}$) follow the specified distribution (e.g., $N(\Theta_2, [\mathcal{I}(\Theta)^{-1}]_{2,2})$). The test is conducted using the method of Equal Local Level (ELL) [32]. If all the points lie within the shaded region, we conclude that the estimation for this parameter is valid.

For a single simulation setting, we generate p^2 such plots. However, with multiple simulation settings, the number of QQ-plots increases significantly. Instead of generating a QQ-plot for every parameter, we summarize the results using p-values. The `qqconf` package provides functions to calculate p-values for the two-sided ELL test. Specifically, we use the following R code to calculate the p-values:

```r
library(qqconf)

cal_p_value <- function(variance, estimated_result, mean, p) {

  result = rep(0, p^2)

  for (i in 1:p^2) {

    data = estimated_result[i, ]

    theor_mean = mean[i]

    theor_sd = sqrt(variance[i])

    zscores = (data - theor_mean) / theor_sd

    n = length(zscores)

    tmp1 = pnorm(zscores)

    tmp1 = sort(tmp1)

    tmp2 = pbeta(tmp1, c(1:n), c(n:1))

    tmp3 = min(min(tmp2), 1 - max(tmp2)) * 2

    lb = qbeta(tmp3 / 2, c(1:n), c(n:1))

    ub = qbeta(1 - tmp3 / 2, c(1:n), c(n:1))

    p = get_level_from_bounds_two_sided(lb, ub)

    result[i] = p

  }

  return(result)

}
```

Here:

- `variance` is a vector of variances computed from the Fisher information matrix, specifically $([\mathcal{I}(\Theta)^{-1}]_{1,1}, \ldots, [\mathcal{I}(\Theta)^{-1}]_{p^2,p^2})^T$,

- `mean` represents the true values of Θ used to simulate the data,

- `estimated_result` is a $p^2 \times n_{\text{reps}}$ matrix, where each row contains the estimated values for a parameter across n_{reps} replicates.

This function computes p-values to assess deviations of the observed data, specifically represented by the variable `estimated_result`, from theoretical distribution under the null hypothesis. The test is based on a QQ-plot approach using two-sided ELL. For each hypothesis, the function determines the largest significance level α such that the observed data lies within the acceptance region of the QQ-plot, constructed using two-sided ELL bounds. This acceptance region is defined based on the theoretical mean and variance for each parameter.

2.4.2 Simulation Results

We simulate data based on the model defined in (2.30) for n individuals and apply the EM algorithm incorporating the Newton-Raphson algorithm, as proposed in Section 2.3.3, to estimate the variance components. Furthermore, we compute the Fisher information using equations (2.51) and (2.52) under different simulation settings. Note that we have a genetic relationship matrix K, and the model in (2.30) requires the eigenvalues derived from the spectral decomposition of K. We simulate data based on three different structures of K:

1. $\frac{n}{2}$ independent sibling pairs,

2. n independent individuals,

3. 63 independent identical pedigrees.

2.4.2.1 Sibling Pairs Case

In this case, we assume there are 500 independent sibling pairs, resulting in a total of 1000 individuals. The K is defined by:

$$K_{sib} = \mathbf{I}_{\frac{n}{2}\times\frac{n}{2}} \otimes \mathbf{B}, \quad \text{where} \quad \mathbf{B} = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}.$$

We perform the spectral decomposition of K_{sib}, obtaining the eigenvalues denoted as V_{sib}. Each element of V_{sib} is used as δ_i in (2.30) to simulate the data.

Simulation setting 1:

We simulate data for 3 traits ($p = 3$), using two positive definite correlation matrices, C_g and C_e, along with a diagonal matrix D whose diagonal elements are constrained within the interval $(0, 1)$. And we generate 1000 replicates for this setting.

$$C_g = \begin{pmatrix} 1 & 0 & 0.5 \\ 0 & 1 & -0.5 \\ 0.5 & -0.5 & 1 \end{pmatrix}, \quad C_e = \begin{pmatrix} 1 & -0.1 & 0.3 \\ -0.1 & 1 & 0 \\ 0.3 & 0 & 1 \end{pmatrix}, \quad D = \begin{pmatrix} 0.3 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.4 \end{pmatrix}. \qquad (2.53)$$

In this case, we have 9 parameters: the first 3 correspond to the diagonal elements of D, the 4th to 6th are the lower off-diagonal elements of C_g, and the last 3 are the lower off-diagonal elements of C_e. We plot the QQ-plot for the estimation of the first parameter, as explained in Section 2.4.1:

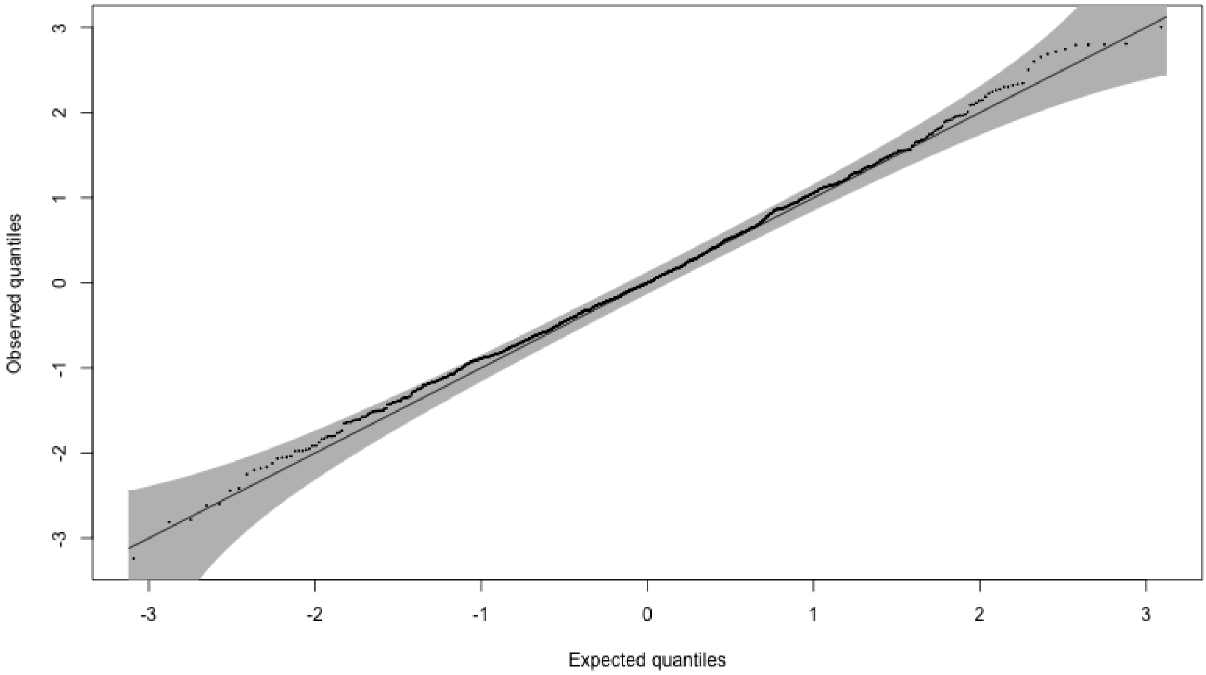


Figure 2.1: **QQ-plot for first parameter estimates in sibling-pairs case setting 1.** The plot contains 1000 points, each representing an estimate of the first parameter from one replicate. The ELL method [32] is used to test whether the estimates deviate from the expected normal distribution. The shaded region is the 95% confidence region computed using ELL.

Figure 2.1 shows that all the points lie within the shaded region as expected. Hence, we conclude that the estimates of the first parameter are valid. Instead of plotting the QQ-plots like Figure 2.1 for all the parameter, we can get the p-values of the corresponding QQ-plots as explained in Section 2.4.1:

Table 2.1: P-values of the QQ-plots for parameter estimation in sibling-pairs case setting 1

Parameters	**P-values**
Parameter 1	0.1764
Parameter 2	0.1985
Parameter 3	0.3291
Parameter 4	0.6937
Parameter 5	0.1054
Parameter 6	0.2360
Parameter 7	0.2683
Parameter 8	0.0532
Parameter 9	0.5380

The p-values assess deviations of the estimated parameters from the theoretical distribution under the null hypothesis.

From Table 2.1, we observe that there is no clear evidence suggesting that any of the estimated parameters deviate from their theoretical distribution.

Simulation setting 2:

We simulate data for 5 traits ($p = 5$), using two positive definite correlation matrices, C_g and C_e, along with a diagonal matrix D whose diagonal elements are constrained within the interval $(0, 1)$. And we generate 1000 replicates for this setting.

$$C_g = \begin{pmatrix} 1 & 0.5 & 0 & -0.5 & 0.1 \\ 0.5 & 1 & 0.2 & 0.3 & -0.3 \\ 0 & 0.2 & 1 & 0.4 & -0.1 \\ -0.5 & 0.3 & 0.4 & 1 & 0.1 \\ 0.1 & -0.3 & -0.1 & 0.1 & 1 \end{pmatrix},$$

$$C_e = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}, \tag{2.54}$$

$$D = \begin{pmatrix} 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.3 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.4 \end{pmatrix}.$$

In this case, we have 25 parameters: the first 5 correspond to the diagonal elements of D, the 6th to 15th are the lower off-diagonal elements of C_g, and the last 10 are the lower off-diagonal elements of C_e. The p-values of the corresponding QQ-plots:

Table 2.2: P-values of the QQ-plots for parameter estimation in sibling-pairs case setting 2

Parameters	**P-values**
Parameter 1	0.2349
Parameter 2	0.7848
Parameter 3	0.4587
Parameter 4	0.8329

Table 2.2 (continued)

Parameters	**P-values**
Parameter 5	0.6267
Parameter 6	0.0466
Parameter 7	0.6018
Parameter 8	0.9596
Parameter 9	0.4041
Parameter 10	0.5408
Parameter 11	0.1391
Parameter 12	0.3214
Parameter 13	0.6861
Parameter 14	0.1988
Parameter 15	0.3087
Parameter 16	0.8752
Parameter 17	0.7916
Parameter 18	0.5657
Parameter 19	0.9053
Parameter 20	0.4291
Parameter 21	0.2128
Parameter 22	0.8641
Parameter 23	0.8829
Parameter 24	0.5961
Parameter 25	0.6571

The p-values assess deviations of the estimated parameters from theoretical distribution under the null hypothesis.

There is only one p-value less than 0.05. Since we performed 25 tests, after accounting for multiple testing, we conclude that there is no clear evidence suggesting that any of the estimated parameters deviate from their theoretical distribution.

2.4.2.2 Independent Individuals Case

In this case, we assume there are 1000 independent individuals. The genetic relationship matrix K is defined as the identity matrix:

$$K = I_{n \times n},$$

where $n = 1000$. Since K is the identity matrix, its eigenvalues are all equal to 1. Thus, when simulating data using the model in (2.30), we set each δ_i to 1:

Simulation setting 1:

We simulate data for 3 traits ($p = 3$), using the same C_g, C_e and D (2.53) in sibling pairs case. And we generate 1000 replicates for this setting.

Table 2.3: P-values of the QQ-plots for parameter estimation in independent individuals case setting 1

Parameters	**P-values**
Parameter 1	0.0190
Parameter 2	0.1878
Parameter 3	0.8814
Parameter 4	0.7048
Parameter 5	0.1268
Parameter 6	0.7767
Parameter 7	0.2532
Parameter 8	0.4736
Parameter 9	0.3236

The p-values assess deviations of the estimated parameters from the theoretical distribution under the null hypothesis.

There is only one p-value less than 0.05. Since we performed 9 tests, after accounting for multiple testing, we conclude that there is no clear evidence suggesting that any of the estimated parameters deviate from their theoretical distribution.

Simulation setting 2:

We simulate data for 5 traits ($p = 5$), using the same C_g, C_e and D (2.54) in sibling pairs case. And we generate 1000 replicates for this setting.

Table 2.4: P-values of the QQ-plots for parameter estimation in independent individuals case setting 2

Parameters	**P-values**
Parameter 1	0.1789

Table 2.4 (continued)

Parameters	**P-values**
Parameter 2	0.0806
Parameter 3	0.3538
Parameter 4	0.9208
Parameter 5	0.2263
Parameter 6	0.3894
Parameter 7	0.7132
Parameter 8	0.3548
Parameter 9	0.0148
Parameter 10	0.8793
Parameter 11	0.5647
Parameter 12	0.1141
Parameter 13	0.9557
Parameter 14	0.8025
Parameter 15	0.2561
Parameter 16	0.7388
Parameter 17	0.8572
Parameter 18	0.3837
Parameter 19	0.2852
Parameter 20	0.4379
Parameter 21	0.4643
Parameter 22	0.7001
Parameter 23	0.0836
Parameter 24	0.2874

Table 2.4 (continued)

Parameters	**P-values**
Parameter 25	0.4408

The p-values assess deviations of the estimated parameters from the theoretical distribution under the null hypothesis.

There is only one p-value less than 0.05. Since we performed 25 tests, after accounting for multiple testing, we conclude that there is no clear evidence suggesting that any of the estimated parameters deviate from their theoretical distribution.

2.4.2.3 Independent Identical Pedigrees Case

In this case, we assume there are 63 independent and identical pedigrees, each consisting of 3 generations with 16 individuals per pedigree, resulting in a total of 1008 individuals. The pedigree structure used for data simulation is as follows:

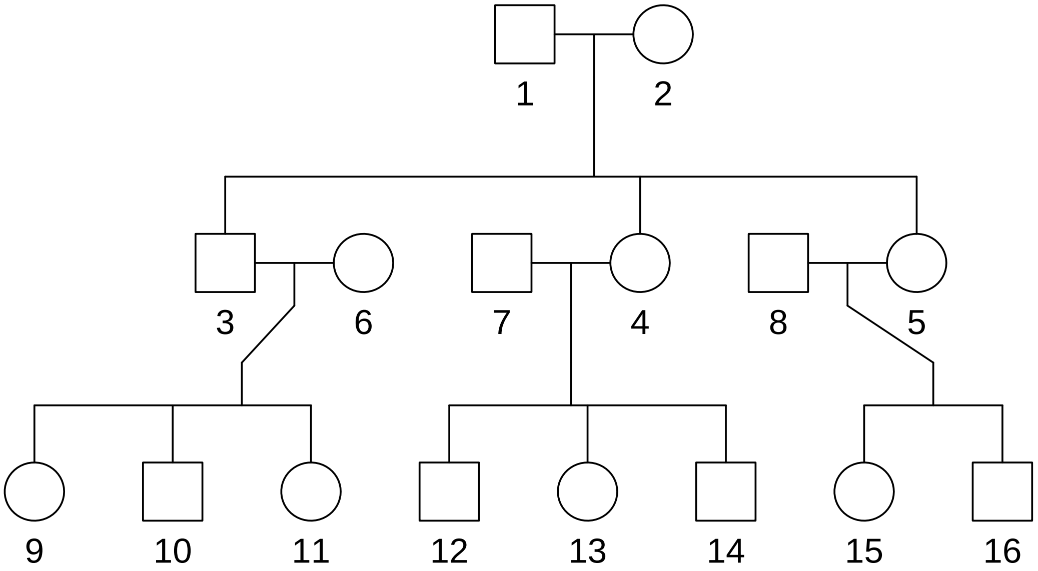


Figure 2.2: **Three-generation pedigree of 16 individuals used in the simulation studies.**

We obtain the kinship matrix for the pedigree shown in Figure 2.2, denoted as K_{ped}. Performing the eigen-decomposition on K_{ped} yields a vector of eigenvalues, denoted as V_{ped}. Since all 63 pedigrees are independent and identical, they share the same kinship matrix and, consequently, the same eigen-decomposition. In the simulation process, the eigenvalues V_{ped} are repeatedly assigned as variance components (δ_i) in (2.30) for each pedigree. Specifically:

- $\delta_1, \delta_2, \ldots, \delta_{16}$ is set to V_{ped},
- $\delta_{17}, \delta_{18}, \ldots, \delta_{32}$ is also set to V_{ped},
- This pattern continues for all 63 pedigrees.

Thus, the same set of eigenvalues, V_{ped}, is used for each of the 63 pedigrees, resulting in a total of $63 \times 16 = 1008$ variance components $(\delta_1, \delta_2, \ldots, \delta_{1008})$.

Simulation setting 1:

We simulate data for 3 traits ($p = 3$), using two positive definite correlation matrices, C_g and C_e, along with a diagonal matrix D whose diagonal elements are constrained within the interval $(0, 1)$. And we generate 1000 replicates for this setting.

$$C_g = \begin{pmatrix} 1.0 & -0.3 & 0.5 \\ -0.3 & 1.0 & 0.5 \\ 0.5 & 0.5 & 1.0 \end{pmatrix}, \quad C_e = \begin{pmatrix} 1 & 0 & 0.5 \\ 0 & 1 & -0.5 \\ 0.5 & -0.5 & 1 \end{pmatrix}, \quad D = \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 0.3 & 0 \\ 0 & 0 & 0.2 \end{pmatrix}. \tag{2.55}$$

Table 2.5: P-values of the QQ-plots for parameter estimation in independent identical pedigree case setting 1

Parameters	**P-values**
Parameter 1	0.7418
Parameter 2	0.0466
Parameter 3	0.4588
Parameter 4	0.6626
Parameter 5	0.3485
Parameter 6	0.1259
Parameter 7	0.7780
Parameter 8	0.1769
Parameter 9	0.5328

The p-values assess deviations of the estimated parameters from the theoretical distribution under the null hypothesis.

There is only one p-value less than 0.05. Since we performed 9 tests, after accounting for multiple testing, we conclude that there is no clear evidence suggesting that any of the estimated parameters deviate from their theoretical distribution.

Simulation setting 2:

We simulate data for 5 traits ($p = 5$), using the same C_e and D (2.54) in sibling pairs case and the positive definite correlation matrix C_g as follows:

$$C_g = \begin{pmatrix} 1 & 0.5 & 0 & -0.2 & 0.1 \\ 0.5 & 1 & 0.2 & 0.3 & -0.3 \\ 0 & 0.2 & 1 & 0.4 & -0.1 \\ -0.2 & 0.3 & 0.4 & 1 & 0.1 \\ 0.1 & -0.3 & -0.1 & 0.1 & 1 \end{pmatrix}, \tag{2.56}$$

And we generate 1000 replicates for this setting.

Table 2.6: P-values of the QQ-plots for parameter estimation in independent identical pedigree case setting 2

Parameters	**P-values**
Parameter 1	0.1093
Parameter 2	0.6356
Parameter 3	0.1756
Parameter 4	0.5070
Parameter 5	0.1019
Parameter 6	0.2333
Parameter 7	0.5035
Parameter 8	0.8354
Parameter 9	0.3671

Table 2.6 (continued)

Parameters	**P-values**
Parameter 10	0.0555
Parameter 11	0.0022
Parameter 12	0.1490
Parameter 13	0.5609
Parameter 14	0.5398
Parameter 15	0.4562
Parameter 16	0.2753
Parameter 17	0.5708
Parameter 18	0.5851
Parameter 19	0.3719
Parameter 20	0.5958
Parameter 21	0.5523
Parameter 22	0.2756
Parameter 23	0.5053
Parameter 24	0.5632
Parameter 25	0.8512

The p-values assess deviations of the estimated parameters from the theoretical distribution under the null hypothesis.

There is only one p-value less than 0.05. Since we performed 25 tests, after accounting for multiple testing, we conclude that there is no clear evidence suggesting that any of the estimated parameters deviate from their theoretical distribution.

2.4.2.4 Summary of Simulation Results

From the simulations conducted, it is evident that the procedure proposed in Section 2.3.3 effectively solves the log-likelihood maximization problem under various circumstances. The computational approach is feasible for a relatively small number of traits (approximately 5).

CHAPTER 3

QUASI-LIKELIHOOD MODEL FOR MULTIPLE BINARY TRAITS AND QUANTITATIVE TRAITS WITH SINGLE GENETIC VARIANT

We can easily extend the model to include quantitative traits. Assume there are n individuals and p traits of which b traits are binary and the rest are quantitative. For mean structure:

$$E(Y_{ij}|X,G) = \mu_{ij}, \quad g(\mu_{ij}) = (\beta X)_{ij} + (\gamma G^T)_{ij}, for\ 1 \leqslant i \leqslant b, \tag{3.1}$$

$$E(Y_{ij}|X,G) = \mu_{ij}, \quad \mu_{ij} = (\beta X)_{ij} + (\gamma G^T)_{ij}, for\ 1+b \leqslant i \leqslant p, \tag{3.2}$$

where

$$g(\mu_{ij}) = log\frac{\mu_{ij}}{1-\mu_{ij}}.$$

For the conditional variance structure, we have:

$$\Omega := Var(vec(Y)|X,G) = \Gamma^{1/2}\Sigma\Gamma^{1/2}. \tag{3.3}$$

Γ is an np-dimensional diagonal matrix, with s^{th} diagonal element, where $s = p(j-1)+i$, given by $\Gamma_{ss} = Var(Y_{ij}|X,G)$, which is equal to $\mu_{ij}(1-\mu_{ij})$ if $1 \leqslant i \leqslant b$ and σ_i^2 if $1+b \leqslant i \leqslant p$, where σ_i^2 represents the total residual variance of trait i. Σ is defined the same as the model with only binary traits (2.3):

$$\Sigma = K \otimes (D^{1/2}C_gD^{1/2}) + I_{n\times n} \otimes (\tilde{D}^{1/2}C_e\tilde{D}^{1/2}), \tag{3.4}$$

This model provides a unified framework for binary and quantitative traits, in the sense that if all traits are quantitative, our model becomes equivalent to a standard multi-trait LMM in [15] while each binary trait has the same binary-trait-specific model as in CARAT [11].

3.1 Parameter Estimation

3.1.1 Parameter Estimation for Coefficients

The coefficient estimation follows the same approach as proposed in Section 2.2. Recall that we bind the notations for the genotype vector and covariate matrix to define:

$$\tilde{X} = \begin{bmatrix} X \\ G^T \end{bmatrix}, \quad \tilde{\beta} = [\beta, \gamma].$$

The quasi-score function for $\text{vec}(\tilde{\beta})$ is given by:

$$U(\text{vec}(\tilde{\beta})) = M^T \Omega^{-1}(\text{vec}(Y) - \text{vec}(\mu)), \tag{3.5}$$

where

$$M = M(\text{vec}(\tilde{\beta})) = \frac{\partial \text{vec}(\mu)}{\partial \text{vec}(\tilde{\beta})} = Q(\tilde{X}^T \otimes I_{p\times p}).$$

Here, Q is an np-dimensional diagonal matrix whose s-th diagonal element, where $s = p(j-1)+i$, is given by:

- Γ_{ss} if $1 \leqslant i \leqslant b$ (i.e., for the binary traits),
- 1 if $b+1 \leqslant i \leqslant p$ (i.e., for the quantitative traits).

Setting $U(\tilde{\beta}) = 0$ and plug in the definition of Ω gives the generalized estimating equation for $\text{vec}(\tilde{\beta})$:

$$(\tilde{X} \otimes I_{p\times p}) J^{1/2} \Sigma^{-1} \Gamma^{-1/2} (\text{vec}(Y) - \text{vec}(\mu)) = 0. \tag{3.6}$$

Here, J is an np-dimensional diagonal matrix whose s-th diagonal element, where $s = p(j-1)+i$, is defined as:

- Γ_{ss} if $1 \leqslant i \leqslant b$ (i.e., for the binary traits),

- Γ_{ss}^{-1} if $b+1 \leqslant i \leqslant p$ (i.e., for the quantitative traits).

To solve for $vec(\tilde{\beta})$, a modified Newton-Raphson algorithm with Fisher scoring is used, which involves iteratively updating $vec(\beta)$ by:

$$\begin{aligned} vec(\tilde{\beta})_{(j+1)} = vec(\tilde{\beta})_{(j)} + \{J^T(vec(\tilde{\beta})_{(j)})\Omega^{-1}J(vec(\tilde{\beta})_{(j)})\}^{-1}\cdot \\ \{J^T(vec(\tilde{\beta})_{(j)})\Omega^{-1}[vec(Y) - vec(\mu(vec(\tilde{\beta})_{(j)}))]\}. \end{aligned} \tag{3.7}$$

Since the i^{th} row of Y represents the phenotypes of the n individuals for trait i, and the i^{th} row of $\tilde{\beta}$ represents the fixed effects of the covariates and genotype for trait i, we can estimate the first b rows (corresponding to the binary traits) of $\tilde{\beta}$ under the single binary trait model incorporating related individuals proposed in CARAT [11] independently. The remaining rows (corresponding to the quantitative traits) can be estimated independently under the single quantitative trait linear mixed model accounting for related individuals. These estimators are then stacked together to form the initial value for vec($\tilde{\beta}$). Starting from the initial values for vec($\tilde{\beta}$), we run this iterative procedure until convergence to obtain vec($\tilde{\beta}$), for fixed variance components (Θ). The technique developed in Section 2.2 facilitates the calculation process.

3.1.2 Parameter Estimation for Variance Components

We conducted simulations using the variance components parameter estimation procedure outlined in Section 2.3. Additionally, we performed simulations using an alternative estimation approach discussed below. This alternative method is faster while providing comparable results. Consequently, we opted to proceed with the alternative procedure for our simulation analysis.

Residual variance

The initial values of the total residual variance for each trait, σ_i^2 for $1 + b \leqslant i \leqslant p$, are estimated by the residual variance obtained from fitting the single quantitative trait linear mixed model accounting for related individuals independently. We plug in the initial values of the residual variance, correlation matrices, and D matrix to estimate the coefficients. After obtaining the coefficients estimator $\text{vec}(\widehat{\tilde{\beta}})$, we estimate the total residual variance as follows:

$$\hat{\sigma}_i^2 = \frac{\left[Y_{i\cdot} - \left(\hat{\tilde{\beta}}\tilde{X}\right)_{i\cdot}\right]^T \left[D_{ii}K + (1 - D_{ii})I_{n\times n}\right]^{-1} \left[Y_{i\cdot} - \left(\hat{\tilde{\beta}}\tilde{X}\right)_{i\cdot}\right]}{n},$$

where $Y_{i\cdot}$ denotes the i-th row of the Y matrix, and $\left(\hat{\tilde{\beta}}\tilde{X}\right)_{i\cdot}$ denotes the i-th row of $\hat{\tilde{\beta}}\tilde{X}$, for $i = b+1, \ldots, p$. Here, $\hat{\tilde{\beta}}$ is obtained as $\text{unvec}(\text{vec}(\widehat{\tilde{\beta}}), p, k)$.

Correlation matrices

We plug the initial values of $\text{vec}(\tilde{\beta})$ and σ_i^2 for $1 + b \leqslant i \leqslant p$ into Γ and denote the resulting value as $\hat{\Gamma}^{(0)}$. Similar to what is described in Section 2.3.2, we define:

$$\hat{\Gamma}^{(0)^{-1/2}}\text{vec}(Y) = \tilde{Y}^{(0)}_{\text{vec}}, \quad \hat{\Gamma}^{(0)^{-1/2}}\text{vec}(\mu) = \tilde{\mu}^{(0)}_{\text{vec}}.$$

We then let:

$$\tilde{\tilde{Y}}^{(0)} = \text{unvec}(\tilde{Y}^{(0)}_{\text{vec}} - \tilde{\mu}^{(0)}_{\text{vec}}, p, n),$$

which represents a $p \times n$ matrix created by reshaping the vector $\tilde{Y}^{(0)}_{\text{vec}} - \tilde{\mu}^{(0)}_{\text{vec}}$ of length pn. The initial values for the correlation matrices C_g and C_e are estimated using the sample correlation matrix calculated across the rows of $\tilde{\tilde{Y}}^{(0)}$. We assume $C_g = C_e$ in our analysis and do not perform further estimation for the correlation matrices.

Diagonal elements of D matrix (heritability for quantitative traits)

For $D_{11}, \ldots, D_{bb}$, i.e., the diagonal elements of the D matrix corresponding to the binary traits, we estimate them as the ξ parameter in Equation 5 of [11]. Specifically, we fit the single binary trait model incorporating related individuals, as proposed in CARAT [11], for each trait and obtain ξ. This parameter measures the proportion of trait i's residual variance that is due to the additive polygenic effect and is analogous to heritability for quantitative traits but defined on the logit scale. For the remaining diagonal elements of the D matrix corresponding to the quantitative traits, we fit a single quantitative trait linear mixed model accounting for related individuals independently for each trait and use the estimated heritability of each trait as the respective diagonal elements of D. We do not perform further estimation for the D matrix.

3.2 Retrospective Association Testing: Asymptotic Method

To detect association between traits and the SNP of interest, we test $H_0 : \gamma = 0$ against $H_1 : \gamma \neq 0$. We let $\hat{\mu}_0$, $\hat{\Sigma}_0$, $\hat{J}_0$ and $\hat{\Gamma}_0$ denotes the values of μ, Σ, J and Γ evaluated at $(\gamma, \beta, \Theta) = (0, \beta_0, \Theta_0)$, where $(\beta, \Theta) = (\beta_0, \Theta_0)$ represents the estimation of $vec(\beta)$ and variance components (lower off-diagonal elements of C_g and C_e, diagonal elements of D matrix and the residual variance σ_i^2 for quantitative trait) under the null. Evaluated at the null estimates, the coordinate of equation (3.6) corresponding to G becomes:

$$U_\gamma = (G^T \otimes I_{p\times p})\hat{J}_0^{1/2}\hat{\Sigma}_0\hat{\Gamma}_0^{-1/2}[Vec(Y) - Vec(\hat{\mu}_0)]. \tag{3.8}$$

Similar to CARAT [11], and using the improvements from CERAMIC [12], we build a quasi-likelihood model for G conditional on Y and X under the null hypothesis of no association, which is specified by the following assumptions:

$$E_0(G|X, Y) = X\alpha \text{ and } Var_0(G|X, Y) = \sigma_g^2 K, \tag{3.9}$$

where α is an unknown k-dimensional vector of coefficients, $\sigma_g^2 > 0$ is an unknown variance parameter and K is the same as the one in equation (2.3). The test statistic for retrospective quasi-likelihood score test of the null hypothesis $H_0 : \gamma = 0$ is:

$$T = U_\gamma^T \cdot [var(U_\gamma|X, Y)]^{-1} \cdot U_\gamma. \tag{3.10}$$

Let $\hat{J}_0^{1/2}\hat{\Sigma}_0\hat{\Gamma}_0^{-1/2}[Vec(Y) - Vec(\hat{\mu}_0)] = L$, and we form a $p \times n$ matrix that satisfy $vec(H) = L$, H is formed as follows, let first p elements of L to be the first column of H, second p elements of L to be the second column of H, so on so forth. So H is the matrix whose j^{th} column consists of elments $p(j-1)+1$ through pj of the vector L. Therefore by equation (2.6):

$$U_\gamma = (G^T \otimes I_{p\times p})vec(H) = vec(HG) = HG. \tag{3.11}$$

The last equal sign is because HG is a vector. Therefore

$$\begin{aligned} T &= (HG)^T \cdot [\text{var}(HG \mid X, Y)]^{-1} \cdot HG \\ &= G^T H^T \cdot [H\text{var}(G \mid X, Y)H^T]^{-1} \cdot HG \\ &= \frac{G^T H^T \cdot [HKH^T]^{-1} \cdot HG}{\hat{\sigma_g^2}}. \end{aligned} \tag{3.12}$$

Under regularity conditions, T has an asymptotic χ_p^2 distribution under null hypothesis. So significance of association is assessed by comparing the test statistic T to a χ_p^2 random variable. We take estimator $\hat{\sigma_g^2}$ to be the residual mean squared error from linear regression of genotype on covariate. Our test statistic is calculated based on parameter estimation under the null hypothesis. Consequently, for a given study, parameter estimation needs to be performed only once, ensuring the computational efficiency of our method.

3.3 Retrospective Association Testing: JASPER

3.3.1 Introduction to JASPER

Other than the method for assessment of significance we have showed in previous section, we also consider another method: JASPER (Joint Association analysis in Structured samples based on approximating a PERmutation distribution) [24] which can be applied to multi-traits test with multiple variants. JASPER has a lot appealing properties: (1) insensitive to misspecification of the phenotype model, (2) does not require knowledge of distribution of the test statistic under the null hypothesis, (3) allow population structure, related individuals, covariates, ascertainment, rare variants, and multiple traits (4) can properly control type I error and can provide substantial power gain over existing methods, (5) computationally efficient. These properties will help us achieve the goals of our study. JASPER extend the fast moment-matching approximation method to account for sample structure. It use a variance component to incorporate sample structure. JASPER relies on two components: (1) a transformation of the test statistic based on a null model proposed for the genetic markers, Model (3.9) in our case, where both the mean and variance structures are specified and incorporate the correlation present due to individuals with similar genetic backgrounds, represented by the K matrix, and (2) an approximation of the null distribution of the transformed test statistic based on its first three moments. The approximation used to estimate the p-value for the test statistic would rely on the rows of either the genotype matrix or the phenotype matrix being exchangeable, which in general is not true. The null model that views the genotype as random is considered and use it to transform the genotype model matrix so as to obtain exchangeable rows. To approximate the null distribution of the test statistic, we need to assess its values on the possible permutations of the exchangeable rows, or on a random sample of the permutation, but this is computationally intensive, so JASPER proposed to use a moment-matching procedure based on approximating this distribution with a Pearson type III distribution using exact analytical calculations of the first three moments of the test statistic. This approach eliminates the need to explicitly carry out the permutations themselves and dramatically decrease the computation cost.

3.3.2 Details of JASPER

JASPER is utilized for a broad class of genetic association test statistics that can be expressed in the following form:

$$T = \operatorname{tr}(S_G S_Y), \quad \text{with} \quad S_G = M_G \Delta M_G^T, \tag{3.13}$$

where S_Y is the phenotype kernel, a symmetric and positive definite $n \times n$ matrix that is a function of the phenotype matrix Y (of dimension $n \times p$, which is the transpose of the phenotype matrix in our setting) and the covariate matrix X (of dimension $n \times k$, which is the transpose of the covariate matrix in our setting). G is the $n \times g$ genotype matrix, indicates there are g genetic variants being test simultaneously. M_G is an $n \times g$ matrix, and Δ is a $g \times g$ symmetric, positive definite matrix. Consequently, S_G, the genotype kernel, is also a symmetric and positive definite $n \times n$ matrix. Let X_G be an $n \times \tilde{k}$ sub-matrix of X consisting of the confounding covariates, and define:

$$H_G = I - X_G(X_G^T X_G)^{-1} X_G, \tag{3.14}$$

as the projection matrix. JASPER assumes $M_G = MG$ and $MX_G = 0$, where M is an $n \times n$ matrix that is a function of X, is non-random given X, and is specifically taken to be $M = H_G$. Furthermore, under the null hypothesis, JASPER assumes that S_G and S_Y are conditionally independent given X.

When related individuals or population structure are present, the rows of M_G are not exchangeable. To address this, JASPER introduces a method to decorrelate the rows and columns of S_G, along with corresponding adjustments to S_Y, resulting in transformed matrices $\tilde{S_G}$ and $\tilde{S_Y}$. This ensures that the test statistic remains equivalent, as:

$$T = \text{tr}(S_G S_Y) = \text{tr}(\tilde{S_G}\tilde{S_Y}), \tag{3.15}$$

where $\tilde{S_G} = \tilde{M_G}\Delta\tilde{M_G}^T$. The empirical distribution of T can be obtained by permuting the rows and columns of $\tilde{S_G}$ (using the same permutation for both rows and columns). This is equivalent to permuting the rows of $\tilde{M_G}$. To assess the significance of T, one can compute its null distribution by considering all possible permutations of the rows of $\tilde{M_G}$ or by taking a random sample of permutations. However, this approach is computationally intensive. To overcome this, JASPER proposes a moment-matching procedure that approximates the null distribution of T using a Pearson Type III distribution. This approximation is achieved through exact analytical calculations of the first three moments of T.

3.3.2.1 Decorrelation of S_G

To decorrelate S_G, JASPER suggests the following steps:

1. Define:

   $$V_r = MKM^T,$$

   where V_r is an $n \times n$ matrix.

2. Perform the eigendecomposition of V_r:

   $$V_r = U\Lambda U^T,$$

   where U is the matrix of eigenvectors, and Λ is the diagonal matrix of eigenvalues.

3. Determine the rank of V_r, denoted by $\tilde{n}$, typically given by:

$$\tilde{n} = n - \tilde{k},$$

   where $\tilde{k}$ is the number of confounding covariates.

4. Construct the following matrices:

   - $\Lambda_{1/2}$: an $n \times \tilde{n}$ matrix containing the $\tilde{n}$ nonzero columns of $\Lambda^{1/2}$.
   - $\Lambda_{-1/2}$: an $n \times \tilde{n}$ matrix containing the $\tilde{n}$ nonzero columns of $(\Lambda^{-})^{1/2}$, where Λ^{-} is the Moore-Penrose generalized inverse of Λ.

5. Define the decorrelated matrices:

$$\begin{aligned} \tilde{M_G} &= \Lambda_{-1/2}^T U^T M_G, \\ \tilde{S_Y} &= \Lambda_{1/2}^T U^T S_Y U \Lambda_{1/2}, \\ \tilde{S_G} &= \tilde{M_G} \Delta \tilde{M_G}^T. \end{aligned}$$

Here, $\tilde{S_Y}$ and $\tilde{S_G}$ represent the decorrelated phenotype and genotype kernels with dimension $\tilde{n} < n$, respectively.

3.3.3 Application of JASPER on BCMAP Single Genetic Variant Case

The test statistic we proposed in (3.12) belongs to the JASPER test statistics class. We can write it as

$$\begin{aligned} T_{\text{JASPER}} &= G^T H^T \cdot [HKH^T]^{-1} \cdot HG \\ &= \text{tr}(G^T H^T \cdot [HKH^T]^{-1} \cdot HG) \\ &= \text{tr}(H^T \cdot [HKH^T]^{-1} \cdot HGG^T). \end{aligned} \tag{3.16}$$

Thus, we naturally have $H^T \cdot [HKH^T]^{-1} \cdot H = S_Y$ as required for JASPER. We assume $X_G = \mathbf{1}_n$, that is X_G is only a intercept column. Since we have only one genetic variant being tested, Δ is a scalar and we set $\Delta = \frac{1}{2\hat{f}(1-\hat{f})}$, where $\hat{f}$ is the estimated allele frequency of the genetic variant. Therefore, $S_G = (I - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)G\frac{1}{2\hat{f}(1-\hat{f})}G^T(I - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)^T$. Once the necessary test statistics are obtained, we can follow the decorrelation procedures outlined in Section 3.3.2.1 to get $\tilde{S_Y}$ and $\tilde{S_G}$ and then apply JASPER test.

3.4 Computational Complexity

The primary computational challenge in BCMAP is the eigen-decomposition of the $n \times n$ matrix K, which is required to account for related individuals and population structure. This step has a computational complexity of $O(n^3)$ but needs to be performed only once per study. Notably, most methods that incorporate related individuals and population structure also require this eigen-decomposition. When the number of traits and covariates is relatively small, the computational complexity for parameter estimation and test statistic calculation is approximately $O(n^2)$ per SNP.

3.5 Simulation Studies

We conducted a series of simulations to evaluate the performance of our method, BCMAP, in achieving correct type I error control and demonstrating higher power compared to existing methods across various scenarios. For the association tests, we utilized both the asymptotic method introduced in Section 3.2, referred to as BCMAP-Asymptotic, and the JASPER method introduced in Section 3.3, referred to as BCMAP-JASPER. To ensure a fair comparison, we selected methods that can incorporate related individuals and covariate information. Additionally, we aimed to assess whether modeling binary traits separately would result in improvements.

For comparisons, we included GEMMA [15], a method that tests associations between multiple quantitative traits and a single genetic variant, while accounting for related individuals and covariate information. GEMMA [15] provides three tests: the Wald test, the Likelihood Ratio Test (LRT), and the score test. Our simulations found that the LRT from GEMMA does not perform well under our simulation settings. Specifically, it occasionally produces extremely small p-values and, at other times, large p-values (equal to 1). As a result, we decided not to include the LRT in our comparisons. We also compared BCMAP with a univariate approach, which involves performing a univariate association test for each trait using Wald test from a linear mixed model that accounts for related individuals and covariate information. The smallest p-value from all univariate tests is then selected, and Bonferroni correction is applied by multiplying the smallest p-value by the number of traits, with the result capped at 1. We refer to this method as Bonf-minP.

3.5.1 Simulation Settings

Simulation Setting for Two Sub-population

One of the advantages of BCMAP is its ability to incorporate related individuals and population structure. To evaluate this, we simulated data for two sub-populations with related individuals. A total of 62 three-generation pedigrees, each consisting of 16 individuals as shown in Figure 2.2, were simulated, resulting in a sample size of 992 individuals.

Genotypes for the founders were generated first, and the genotypes for non-founders within each pedigree were simulated using a gene-dropping approach. For genotypes in the two sub-populations, we applied the Balding-Nichols model with $F = 0.01$, where ancestral allele frequencies for SNPs were drawn independently and uniformly between 0.2 and 0.8. Of these pedigrees, 31 were assigned to sub-population 1, and the remaining 31 to sub-population 2. The founders of each pedigree were assumed to be randomly sampled from their respective sub-populations.

Covariate and Trait Models

Two observed covariates are simulated: one is a continuous covariate drawn from a standard Gaussian distribution, and the other is drawn from a Gaussian distribution with mean 0 and variance 4. An additional covariate is used to simulate phenotypes but is assumed to be unobserved. This covariate corresponds to the major causal variant, M, which is a vector of length n and is simulated as a genetic variant, as described in the previous subsection and it serves as a source of model misspecification. In all simulation scenarios, covariate values are assumed to be independent across individuals and are re-simulated for each replicate. We first simulate a random variable μ and then use this random variable to simulate the phenotype Y.

We consider two models to simulate phenotypes, given the genotype and covariate information:

Logistic Model for Binary Traits

$$\mu = \beta_{p\times k}X_{k\times n} + \gamma_{p\times g}G_{n\times g}^{T} + \delta_{p\times 1}M_{n\times 1}^{T} + \alpha + \epsilon, \tag{3.17}$$

for $i = 1, \ldots, p$ and $j = 1, \ldots, n$, we have:
for binary traits:

$$Y_{ij} \sim \text{Bernoulli}(p_{ij}), \tag{3.18}$$

and

$$\text{logit}(p_{ij}) = \mu_{ij}. \tag{3.19}$$

For quantitative traits:

$$Y_{ij} = \mu_{ij}. \tag{3.20}$$

Here, X is a $3 \times n$ covariate matrix, including the intercept and the two observed covariates described above. Y is the phenotype matrix, G is the genotype matrix with g genetic variants, in single genetic variant case $g = 1$ and G will be a vector, and M is the unobserved major causal variant. If we are simulating under the null, we will set $\gamma_{p\times g} = 0$. The random effects are modeled as:

$$\alpha \sim N(0, K \otimes (W_1^{1/2} C W_1^{1/2})), \tag{3.21}$$

$$\epsilon \sim N(0, I_{n\times n} \otimes (W_2^{1/2} C W_2^{1/2})), \tag{3.22}$$

where K is estimated from 10^5 SNPs, C is a correlation matrix, and W_1 and W_2 are diagonal matrices, with each diagonal element corresponding to the variance of the additive genetic effect and the environmental effect for each trait, respectively. For each binary trait i, $\frac{(W_1)_{ii}}{(W_1)_{ii}+(W_2)_{ii}}$ represents the heritability analogue on the logit scale. For each quantitative trait i, $\frac{(W_1)_{ii}}{(W_1)_{ii}+(W_2)_{ii}}$ represents the heritability.

Liability Threshold Model for Binary Traits

$$\mu = \beta_{p\times k} X_{k\times n} + \gamma_{p\times g} G^T_{n\times g} + \delta_{p\times 1} M^T_{n\times 1} + \alpha + \epsilon, \tag{3.23}$$

for $i = 1, \ldots, p$ and $j = 1, \ldots, n$, we have:
for binary traits:

$$Y_{ij} = \begin{cases} 1 & \text{if } \mu_{ij} \geqslant 0, \\ 0 & \text{if } \mu_{ij} < 0, \end{cases} \tag{3.24}$$

and for quantitative traits:

$$Y_{ij} = \mu_{ij}. \tag{3.25}$$

The definitions of the parameters in the liability threshold model are the same as those in the logistic model. For each binary trait i, $\frac{(W_1)_{ii}}{(W_1)_{ii}+(W_2)_{ii}}$ represents the heritability analogue on the liability threshold model. For each quantitative trait i, $\frac{(W_1)_{ii}}{(W_1)_{ii}+(W_2)_{ii}}$ represents the heritability.

Genetic Relationship Matrix

The genetic relationship matrix estimate K in (3.21) is estimated from 10^5 SNPs and is calculated by:

$$K_{ij} = \frac{1}{L}\sum_{l=1}^{L}\frac{(G_{il}-2\hat{p}_l)(G_{jl}-2\hat{p}_l)}{2\hat{p}_l(1-\hat{p}_l)}, \quad \text{where } L = 10^5 \text{ and } \hat{p}_l = \frac{1}{2n}\sum_{i=1}^{n} G_{il}. \tag{3.26}$$

Type I Error and Power Simulation Setting

Type I error simulations For type I error simulations, we simulated 10 sets of 10^5 variants for each setting and calculated the 10 corresponding GRM estimates based on Equation (3.26). For each set of 10^5 variants, phenotypes were re-simulated 100 times, and 100 variants were randomly selected to be tested against the simulated phenotypes. This process resulted in a total of 10^5 replicates.

Power simulations For power simulations, 10^5 variants were simulated only once for each setting and the corresponding GRM estimate is calculated based on Equation (3.26). From these variants, one variant was selected as causal at a time and tested for association with the traits and phenotypes are re-simulated each time. This process was repeated 1000 times, resulting in 1000 replicates for evaluating power.

3.5.2 Simulation Results: Logistic Model for Binary Traits

In this section, we examine the simulation results based on the logistic model for binary traits defined in Section 3.5.1.

3.5.2.1 2 Binary Traits, 1 Quantitative Trait Case

We simulated two binary traits and one quantitative trait. The first two traits are binary, while the last trait is quantitative.

Setting 1

The correlation matrix we used to simulate data is

$$C = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}.$$

The remaining parameters are chosen such that:

1. The simulated data result in a prevalence of approximately 40% for the two binary traits.

2. The sample correlation matrix for the three traits is approximately:

$$\begin{bmatrix} 1 & 0.6 & 0.3 \\ 0.6 & 1 & 0.3 \\ 0.3 & 0.3 & 1 \end{bmatrix}.$$

3. The heritability analogue on the logit scale, as defined earlier, is approximately 90% for the binary traits, and the heritability for the quantitative trait is approximately 50%.

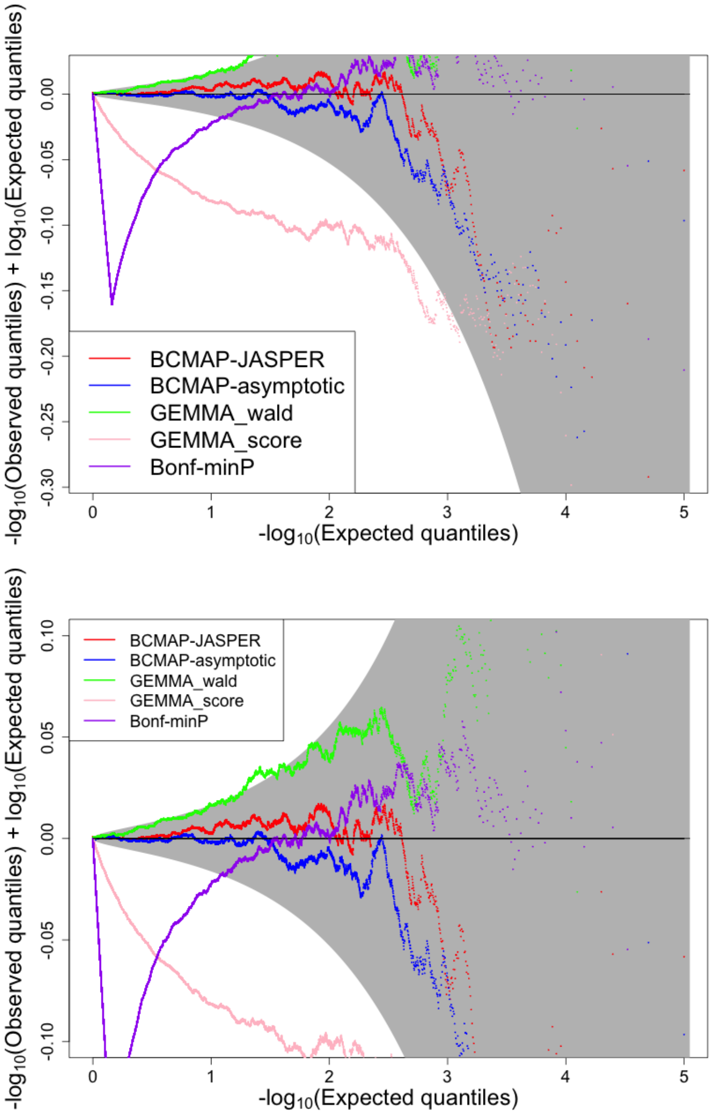
4. For each binary trait, the Bernoulli variance explains, on average, 30% of the total variability in the binary case-control status.

5. On the logit scale, considering the variability explained by the covariates, the major

causal variant, and the additive polygenic effect (α_{ij}), the proportion of variability explained by the covariates is approximately 80%.

Type I error result We test the null hypothesis that there is no genetic association between a single genetic variant and the three traits we simulated. Since we simulated 10^5 replicates, we obtained 10^5 p-values. Under the null hypothesis, the p-values are expected to follow a uniform distribution. Figure 3.1 depicts the (differenced) QQ-plots of the 10^5 p-values against the standard uniform distribution: we take the $-\log_{10}$-scaled p-values and plot the difference between the observed quantiles and the theoretical quantiles (quantiles of Uniform$(0, 1)$). The QQ-plot is generated using the `qqconf` R package [32], which provides a shaded simultaneous acceptance region to assess whether the p-values follow a uniform distribution using ELL method as described in Section 2.4.1. If all the p-values lie within the shaded region, we conclude that the type I error is well controlled. We observe that both our asymptotic method and the JASPER method control the type I error well. The score test from GEMMA and Bonf-minP provide conservative p-values, whereas the Wald test from GEMMA produces inflated p-values.

Figure 3.1: **(Differenced) QQ-plots for p-values: single genetic variant, logistic model for binary traits, two binary traits and one quantitative trait (setting 1):** Top: original scale; bottom: zoomed in. The shaded region is the 99% confidence region by ELL



Setting 2

The correlation matrix we used to simulate data is

$$C = \begin{bmatrix} 1 & -0.3 & 0.5 \\ -0.3 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}.$$

The remaining parameters are chosen such that:

1. The simulated data result in a prevalence of approximately 25% for the two binary traits.

2. The sample correlation matrix for the three traits is approximately:

$$\begin{bmatrix} 1 & 0.6 & 0.5 \\ 0.6 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}.$$

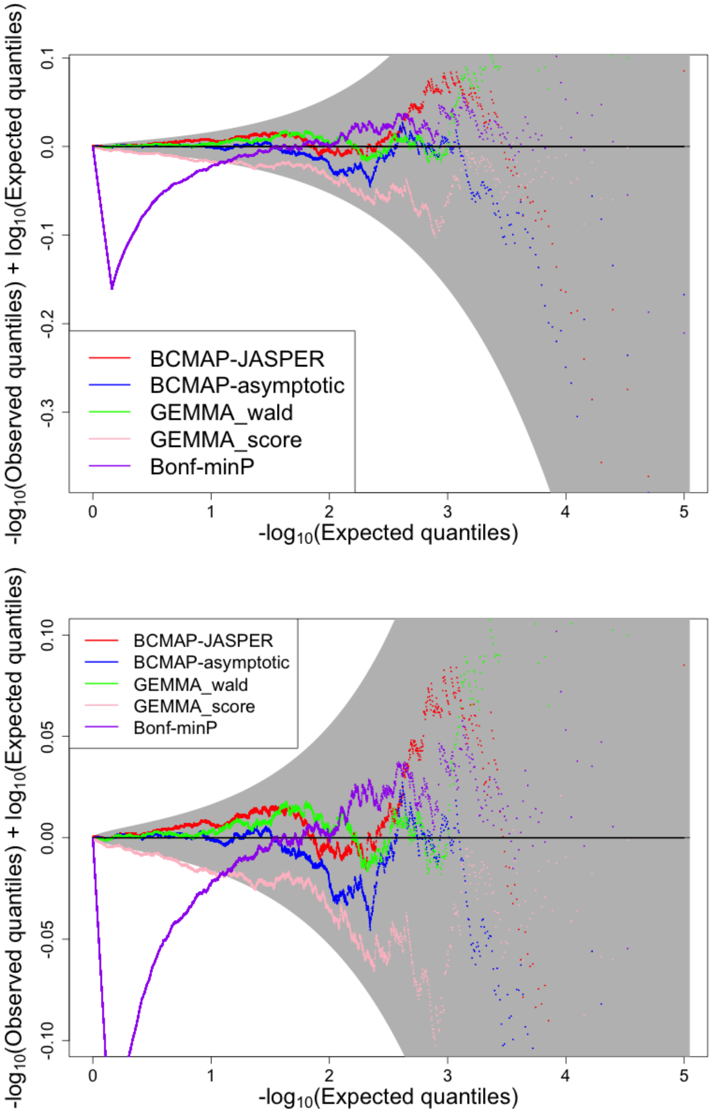
3. The heritability analogue on the logit scale, as defined earlier, is approximately 50% for the first binary trait and 40% for the second binary trait, and the heritability for the quantitative trait is approximately 50%.

4. For each binary trait, the Bernoulli variance explains, on average, 30% of the total variability in the binary case-control status.

5. On the logit scale, considering the variability explained by the covariates, the major causal variant, and the additive polygenic effect (α_{ij}), the proportion of variability explained by the covariates is approximately 90%.

Type I error result We simulate the data under the null hypothesis that there is no association between the 3 traits and the causal SNP. From Figure 3.2 we observe under this simulation setting our asymptotic method, the JASPER method and Wald test from GEMMA control the type I error well. The score test from GEMMA and Bonf-minP still provide conservative p-values. The improvement of the GEMMA Wald test might be attributed to the decrease in the heritability analogue on the logit scale for binary traits. If this value is too large, it may become challenging for GEMMA to perform accurate analysis.

Figure 3.2: **(Differenced) QQ-plots for p-values: single genetic variant, logistic model for binary traits, two binary traits and one quantitative trait (setting 2):** Top: original scale; bottom: zoomed in. The shaded region is the 99% confidence region by ELL



Power analysis result Since under this simulation setting, GEMMA Wald test has correct type I error control, we could perform power analysis to compare if our method has higher power. The parameter γ is chosen such that, on the logit scale, considering the variability explained by the covariates, the major causal variant, the additive polygenic effect (α_{ij}), and the causal SNP, the proportion of variability explained by the causal SNP is approximately 1%. For the quantitative trait, the causal SNP explains approximately 1% of the total variance. Wang, Meigs, and Dupuis [1] simulated two random variables based on a linear mixed model with a positive correlation between the variables. One of the variables was then transformed into a binary trait using a threshold model. Their simulation results suggested that when the two untransformed traits have opposite directions of association with a causal SNP, their joint modeling approach for one binary trait and one quantitative trait is more powerful than conducting univariate tests for each trait. This observation motivates our interest in investigating whether our methods yield similar outcomes.

Notably, we simulate the random variable μ first and use it to simulate the traits. Specifically:

- $\mu_{1\cdot}$ corresponds to the first binary trait. We refer to this as the untransformed first binary trait, following [1].
- $\mu_{2\cdot}$ corresponds to the second binary trait, referred to as the untransformed second binary trait.
- $\mu_{3\cdot}$ represents the quantitative trait.

The correlation used to simulate the data between the untransformed first binary trait and the quantitative trait is positive, and the correlation between the untransformed second binary trait and the quantitative trait is also positive.

We simulated data with the following scenarios:

1. $\text{sign}(\gamma) = c(1,1,1)^T$: The causal SNP has the same direction of effect for the untransformed binary traits and the quantitative trait. We denote this scenario as "Same Direction."

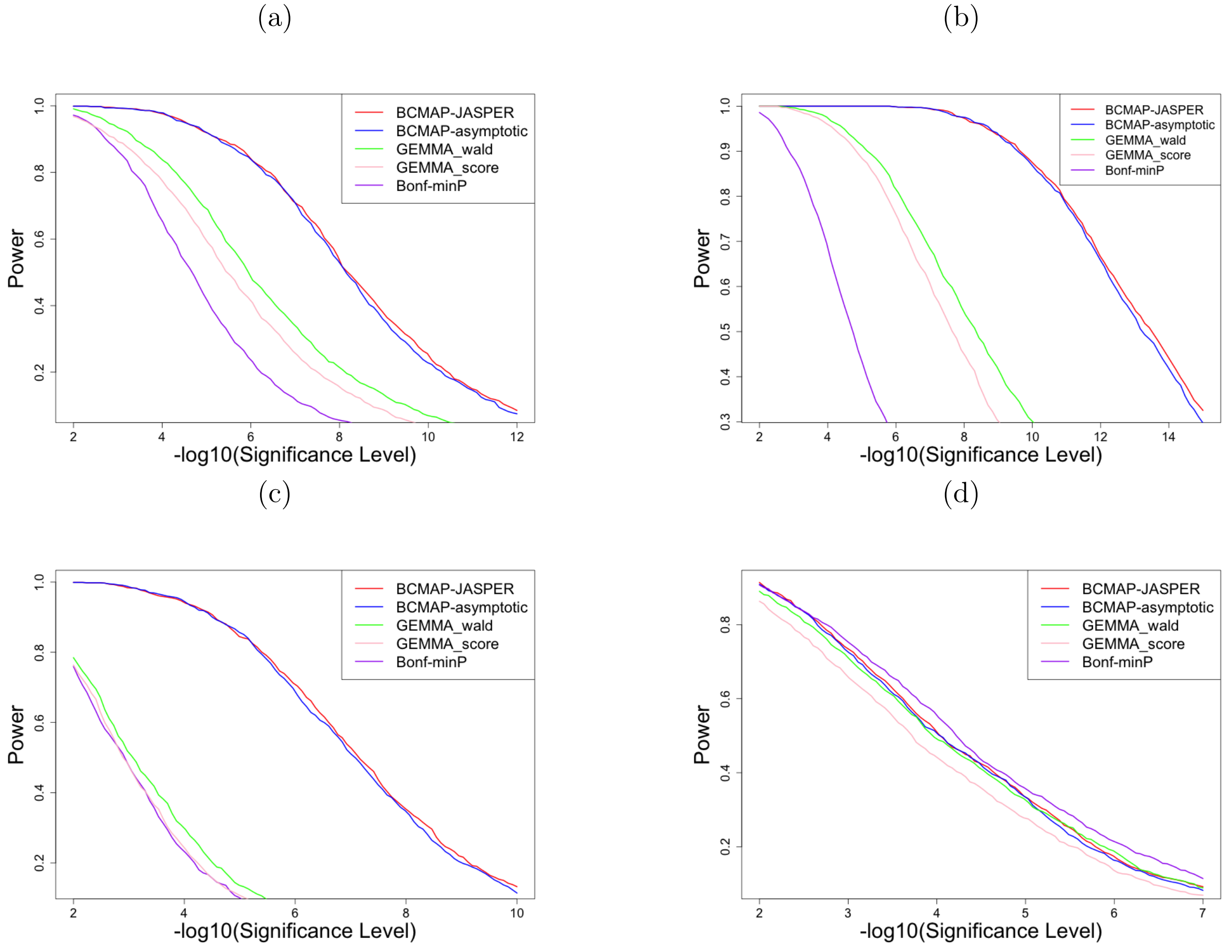
2. $\text{sign}(\gamma) = c(1,1,-1)^T$: The causal SNP has opposite directions of effect for the untransformed binary traits and the quantitative trait. We denote this as "Opposite Direction."

3. $\text{sign}(\gamma) = c(1,1,0)^T$: The causal SNP affects only the binary traits. We denote this as "Only Binary Traits."

4. $\text{sign}(\gamma) = c(0,0,1)^T$: The causal SNP affects only the quantitative trait. We denote this as "Only Quantitative Trait."

We plot the power with respect to the $-\log_{10}$ transformation of different significance levels. A higher curve in the plot indicates that the method has greater power. Figure 3.3 shows the power curves for the four scenarios. We observe that the asymptotic method and the JASPER method from BCMAP produce comparable results. Except for the "Only Quantitative Trait" case, BCMAP demonstrates higher power compared to GEMMA and Bonf-minP. For the "Opposite Direction" case, the gap between BCMAP and the other methods is more pronounced than in the "Same Direction" case, which aligns with the findings from [1]. When only the quantitative trait is associated with causal SNP, all the methods produce comparable results. And when binary trait is associated with causal SNP, BCMAP has higehr power.

Figure 3.3: **Power curves: single genetic variant, logistic model for binary traits, two binary traits and one quantitative trait (setting 2):**



Power curves for different simulation scenarios: (a). Same Direction. (b). Opposite Direction. (c). Only Binary Traits. (d). Only Quantitative Trait.

3.5.2.2 1 Binary Trait, 2 Quantitative Traits Case

We simulated one binary trait and two quantitative traits. The first trait is binary, while the last two traits are quantitative.

Setting 1

The correlation matrix we used to simulate data is

$$C = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}.$$

The remaining parameters are chosen such that:

1. The simulated data result in a prevalence of approximately 50% for the binary trait.
2. The sample correlation matrix for the three traits is approximately:

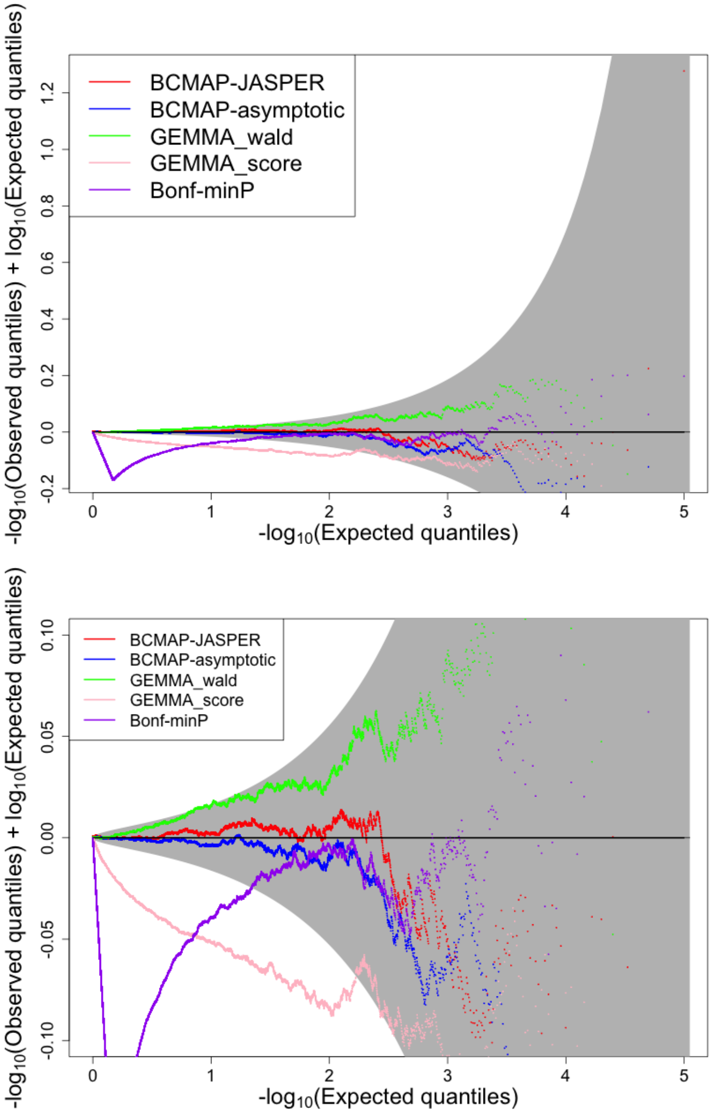
$$\begin{bmatrix} 1 & -0.4 & 0.3 \\ -0.4 & 1 & -0.3 \\ 0.3 & -0.3 & 1 \end{bmatrix}.$$

3. The heritability analogue on the logit scale, as defined earlier, is approximately 70% for the binary trait. The heritability for the first quantitative trait is approximately 40%, and for the second quantitative trait, it is approximately 50%.
4. For the binary trait, the Bernoulli variance explains, on average, 30% of the total variability in the binary case-control status.
5. On the logit scale, considering the variability explained by the covariates, the major causal variant, and the additive polygenic effect (α_{ij}), the proportion of variability explained by the covariates is approximately 80%.

Type I error result We simulate the data under the null hypothesis that there is no association between the 3 traits and the causal SNP. From Figure 3.4 we observe under this simulation setting BCMAP control the type I error well. The score test from GEMMA and

Bonf-minP still provide conservative p-values. The GEMMA Wald test provides slightly inflated p-values, possibly due to the high heritability analogue (70%) on the logit scale for the binary trait.

Figure 3.4: **(Differenced) QQ-plots for p-values: single genetic variant, logistic model for binary traits, one binary trait and two quantitative traits (setting 1):** Top: original scale; bottom: zoomed in. The shaded region is the 99% confidence region by ELL



Setting 2

The correlation matrix we used to simulate data is

$$C = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}.$$

The remaining parameters are chosen such that:

1. The simulated data result in a prevalence of approximately 25% for the binary trait.

2. The sample correlation matrix for the three traits is approximately:

$$\begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.4 \\ 0.5 & 0.4 & 1 \end{bmatrix}.$$

3. The heritability analogue on the logit scale, as defined earlier, is approximately 40% for the binary trait. The heritability for the 2 quantitative traits is approximately 50%.

4. For the binary trait, the Bernoulli variance explains, on average, 30% of the total variability in the binary case-control status.

5. On the logit scale, considering the variability explained by the covariates, the major causal variant, and the additive polygenic effect (α_{ij}), the proportion of variability explained by the covariates is approximately 90%.

Power analysis result Since the heritability analogue on the logit scale for the binary trait is not very large, the GEMMA Wald test might have correct type I error control. We directly conducted a power analysis for this setting. The parameter γ is chosen such that, on the logit scale, considering the variability explained by the covariates, the major causal variant, the additive polygenic effect (α_{ij}), and the causal SNP, the proportion of variability explained by the causal SNP is approximately 1%. For the quantitative traits, the causal SNP explains approximately 0.5% of the total variance. Similar to before,

- $\mu_{1\cdot}$ corresponds to the binary trait. We refer to this as the untransformed binary trait.
- $\mu_{2\cdot}$ represents the first quantitative trait.
- $\mu_{3\cdot}$ represents the second quantitative trait.

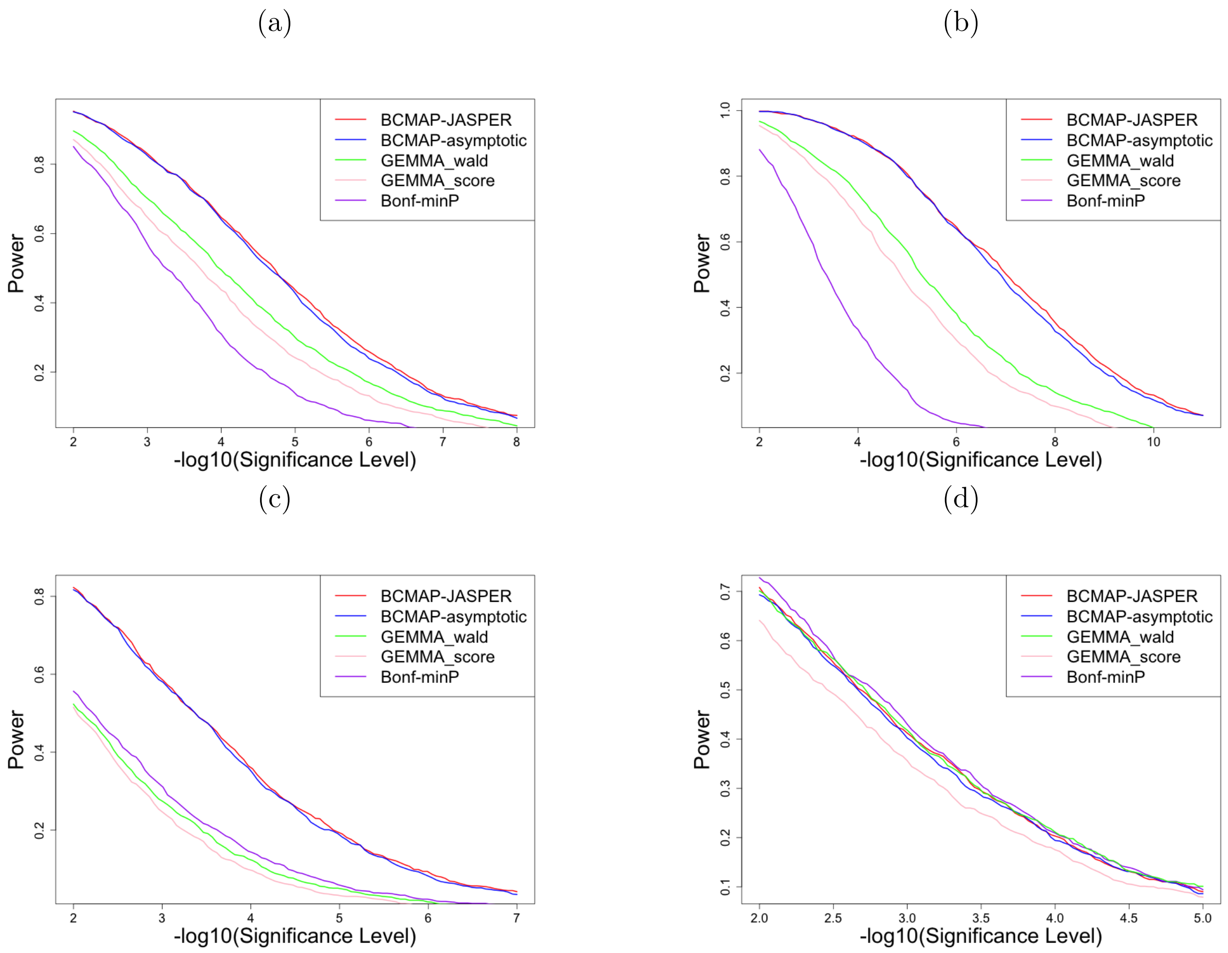
The correlations used to simulate the data between the untransformed binary trait and the quantitative traits are positive.

We simulated data with the following scenarios:

1. $\text{sign}(\gamma) = c(1, 1, 1)^T$: The causal SNP has the same direction of effect for the untransformed binary trait and the quantitative traits. We denote this scenario as "Same Direction."
2. $\text{sign}(\gamma) = c(1, -1, -1)^T$: The causal SNP has opposite directions of effect for the untransformed binary trait and the quantitative traits. We denote this as "Opposite Direction."
3. $\text{sign}(\gamma) = c(1, 0, 0)^T$: The causal SNP affects only the binary trait. We denote this as "Only Binary Trait."
4. $\text{sign}(\gamma) = c(0, 1, 1)^T$: The causal SNP affects only the quantitative traits. We denote this as "Only Quantitative Traits."

Figure 3.5 shows similar results as what we have in two binary traits and one quantitative trait case.

Figure 3.5: **Power curves: single genetic variant, logistic model for binary traits, one binary trait and two quantitative traits (setting 2):**



(a) (b) (c) (d)

Power curves for different simulation scenarios: (a). Same Direction. (b). Opposite Direction. (c). Only Binary Trait. (d). Only Quantitative Traits.

3.5.3 Simulation Results: Liability Threshold Model for Binary Traits

In this section, we examine the simulation results based on the liability threshold model for binary traits defined in Section 3.5.1. Note that our quasi-likelihood model for binary traits is based on the logistic model, so the model misspecification problem in this setting is more severe.

3.5.3.1 2 Binary Traits, 1 Quantitative Trait Case

Setting 1

The correlation matrix we used to simulate data is

$$C = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}.$$

The remaining parameters are chosen such that:

1. The simulated data result in a prevalence of approximately 40% for the two binary traits.

2. The sample correlation matrix for the three traits is approximately:

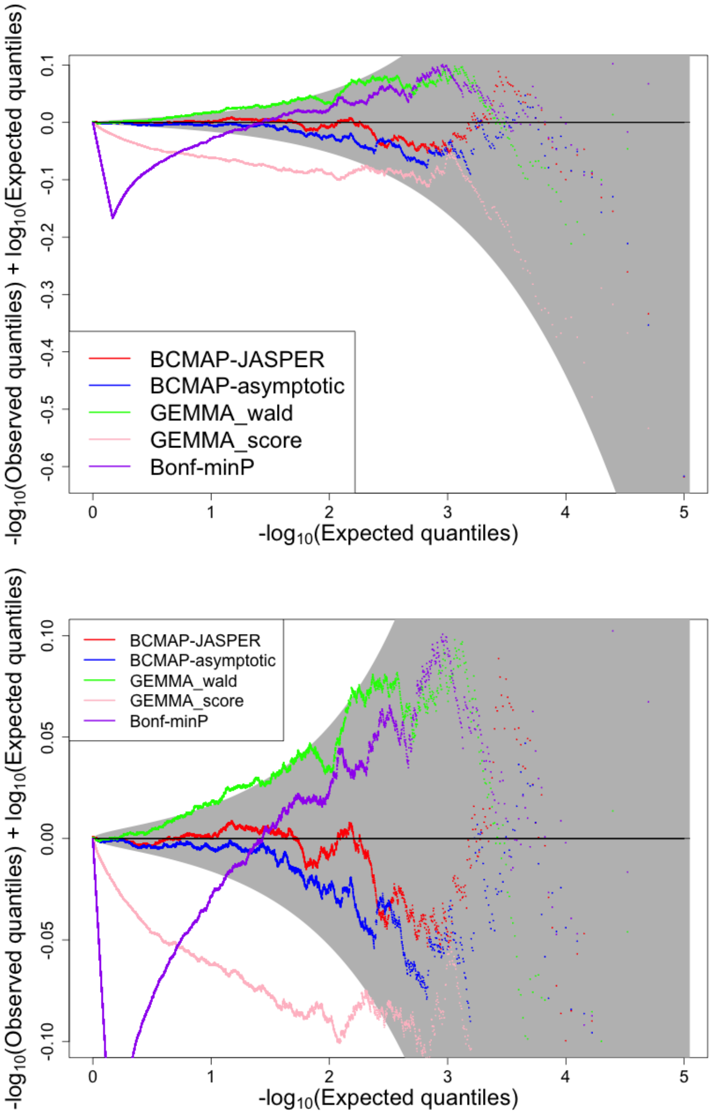
$$\begin{bmatrix} 1 & 0.7 & 0.4 \\ 0.7 & 1 & 0.3 \\ 0.4 & 0.3 & 1 \end{bmatrix}.$$

3. The heritability analogue on the liability threshold model, is approximately 90% for the binary traits, and the heritability for the quantitative trait is approximately 50%.

4. On the liability threshold model, considering the variability explained by the covariates, the major causal variant, the additive polygenic effect (α_{ij}) and the environment effect (ϵ_{ij}), the proportion of variability explained by the covariates is approximately 75%.

Type I error result We simulate the data under the null hypothesis that there is no association between the 3 traits and the causal SNP. From Figure 3.6 we observe the result is similar to Figure 3.1. The GEMMA Wald test provides inflated p-values, possibly due to the high heritability analogue on the liability threshold model for the binary traits.

Figure 3.6: **(Differenced) QQ-plots for p-values: single genetic variant, liability threshold model for binary traits, two binary traits and one quantitative trait (setting 1):** Top: original scale; bottom: zoomed in. The shaded region is the 99% confidence region by ELL



Setting 2

The correlation matrix we used to simulate data is

$$C = \begin{bmatrix} 1 & 0.3 & 0.5 \\ 0.3 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}.$$

The remaining parameters are chosen such that:

1. The simulated data result in a prevalence of approximately 25% for the two binary traits.

2. The sample correlation matrix for the three traits is approximately:

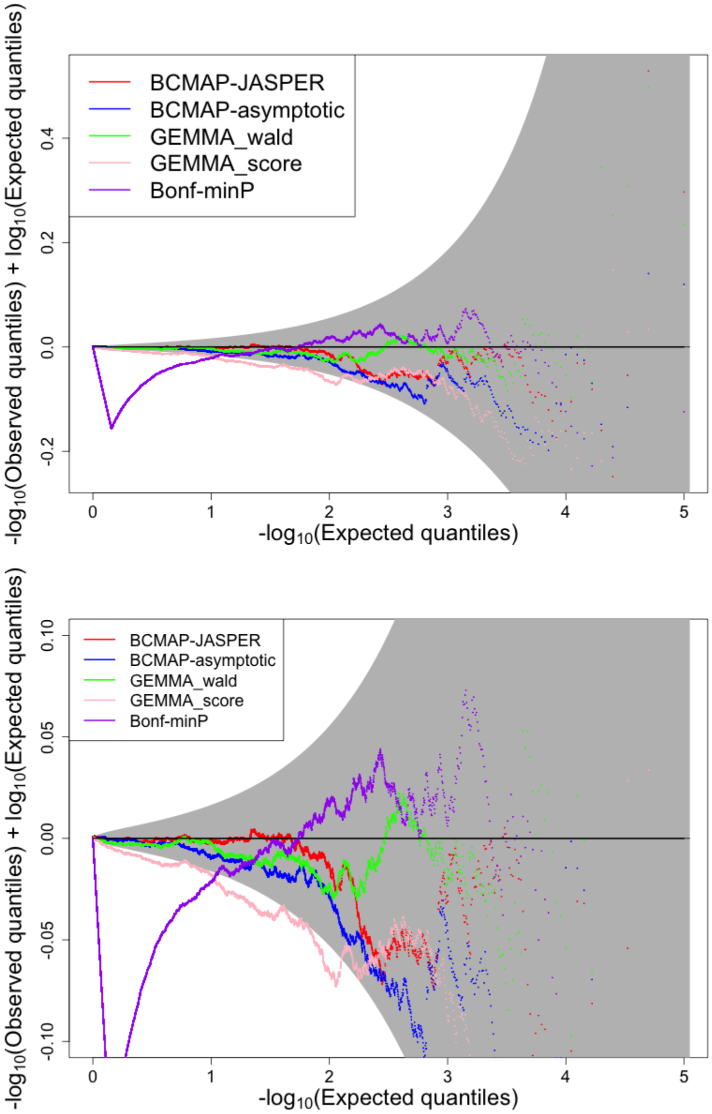
$$\begin{bmatrix} 1 & 0.3 & 0.6 \\ 0.3 & 1 & 0.3 \\ 0.6 & 0.3 & 1 \end{bmatrix}.$$

3. The heritability analogue on the liability threshold model, is approximately 14% for the binary traits, and the heritability for the quantitative trait is approximately 50%.

4. On the liability threshold model, considering the variability explained by the covariates, the major causal variant, the additive polygenic effect (α_{ij}) and the environment effect (ϵ_{ij}), the proportion of variability explained by the covariates is approximately 90%.

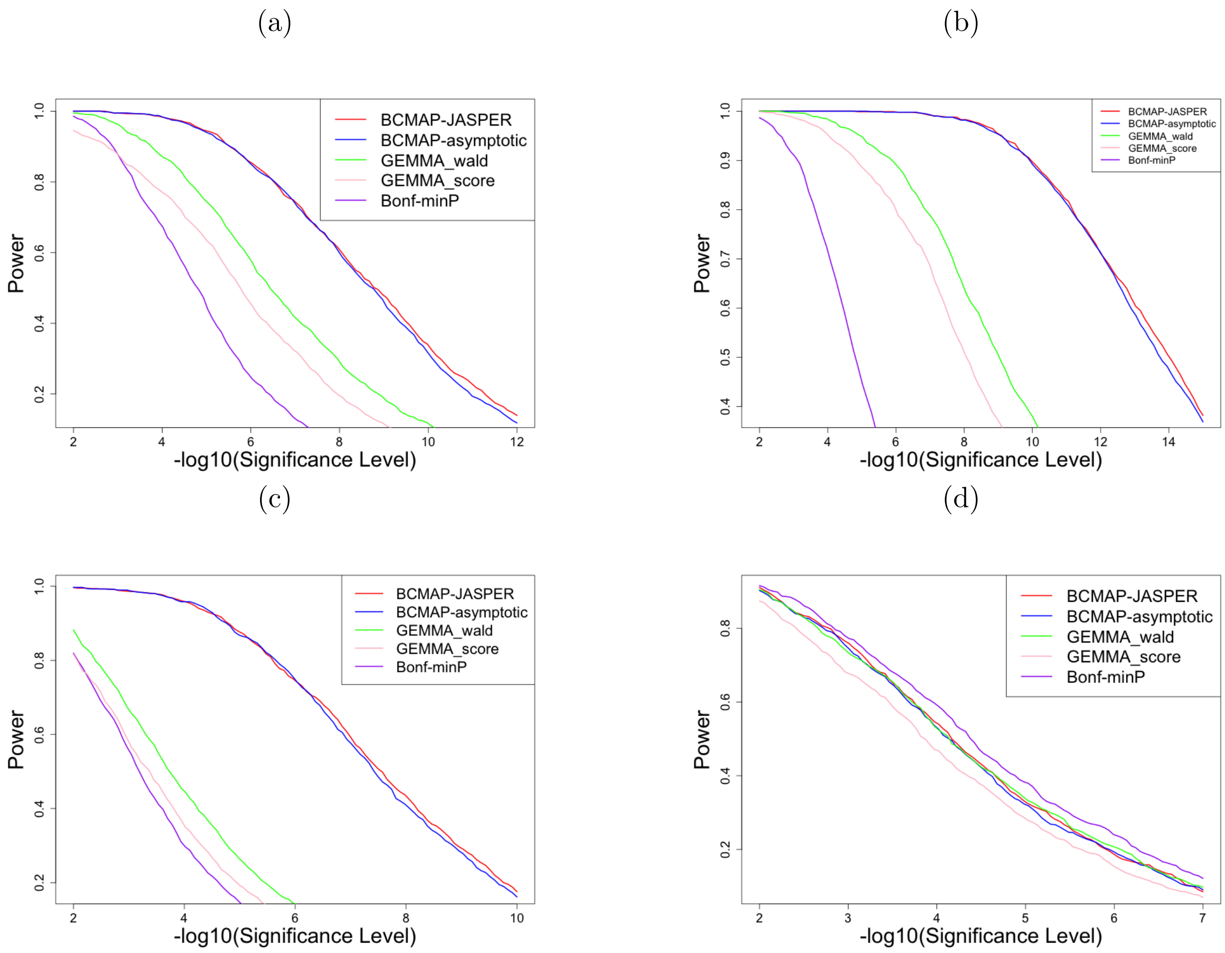
Type I error result We simulate the data under the null hypothesis that there is no association between the 3 traits and the causal SNP. From Figure 3.7 we observe the result is similar to Figure 3.2. The improvement of the GEMMA Wald test might be attributed to the decrease in the heritability analogue on the liability threshold model for binary traits.

Figure 3.7: **(Differenced) QQ-plots for p-values: single genetic variant, liability threshold model for binary traits, two binary traits and one quantitative trait (setting 2):** Top: original scale; bottom: zoomed in. The shaded region is the 99% confidence region by ELL



Power analysis result We conducted a similar power analysis as in the case of the logistic model for binary traits. Figure 3.8 shows results similar to those in Figure 3.3.

Figure 3.8: **Power curves: single genetic variant, liability threshold model for binary traits, two binary traits and one quantitative trait (setting 2):**


(a)

(b)

(c)

(d)

Power curves for different simulation scenarios: (a). Same Direction. (b). Opposite Direction. (c). Only Binary Traits. (d). Only Quantitative Trait.

3.5.3.2 1 Binary Trait, 2 Quantitative Traits Case

The correlation matrix we used to simulate data is

$$C = \begin{bmatrix} 1 & 0.3 & 0.5 \\ 0.3 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}.$$

The remaining parameters are chosen such that:

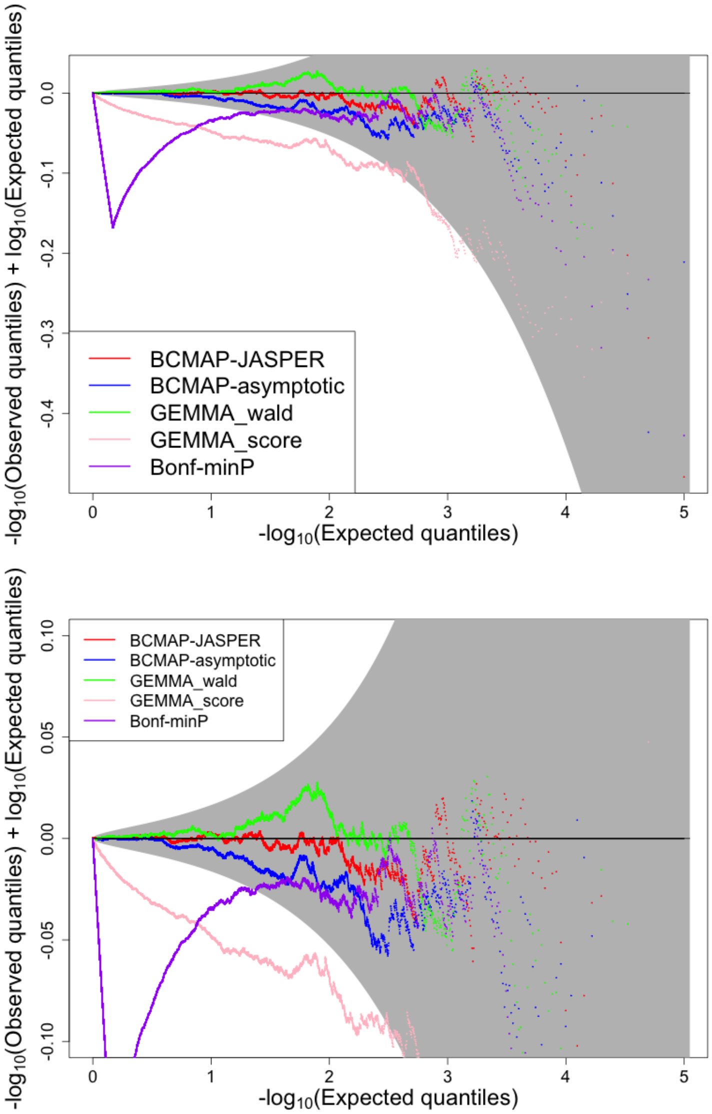
1. The simulated data result in a prevalence of approximately 25% for the two binary traits.
2. The sample correlation matrix for the three traits is approximately:

$$\begin{bmatrix} 1 & 0.5 & 0.3 \\ 0.5 & 1 & 0.3 \\ 0.3 & 0.3 & 1 \end{bmatrix}.$$

3. The heritability analogue on the liability threshold model, is approximately 14% for the binary trait, and the heritability for the 2 quantitative traits is approximately 50%.
4. On the liability threshold model, considering the variability explained by the covariates, the major causal variant, the additive polygenic effect (α_{ij}) and the environment effect (ϵ_{ij}), the proportion of variability explained by the covariates is approximately 90%.

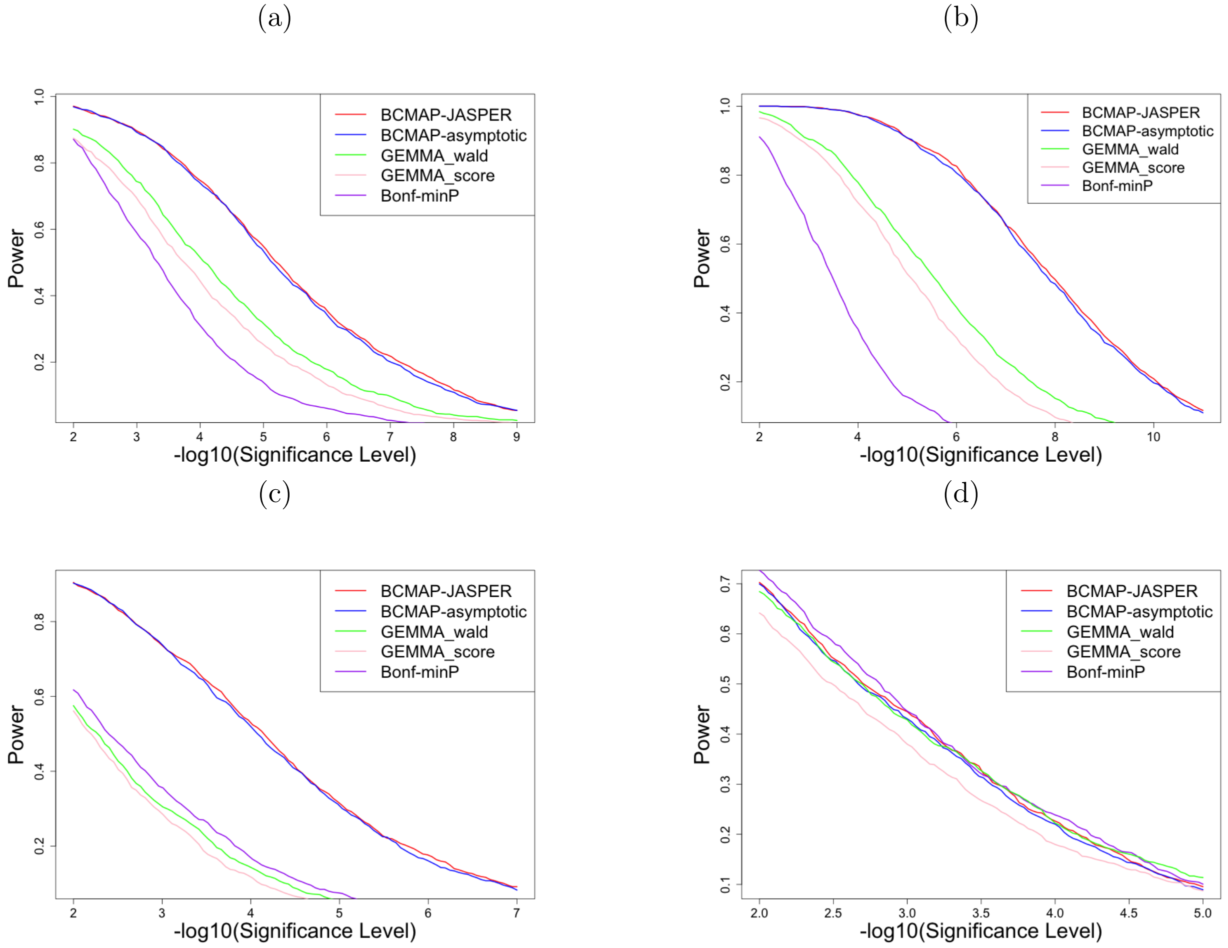
Type I error result We simulate the data under the null hypothesis that there is no association between the 3 traits and the causal SNP. Figure 3.9 shows that BCMAP and the GEMMA Wald test provide correct type I error control, while the GEMMA score test and Bonf-minP produce conservative p-values under the null. The heritability analogue on the liability threshold model for binary traits is low in this setting, which might explain why the GEMMA Wald test performs well.

Figure 3.9: **(Differenced) QQ-plots for p-values: single genetic variant, liability threshold model for binary traits, one binary trait and two quantitative traits:** Top: original scale; bottom: zoomed in. The shaded region is the 99% confidence region by ELL



Power analysis result Under this setting, we conducted a similar power analysis as in the case of the logistic model for binary traits. Figure 3.10 shows results similar to those in previous simulations.

Figure 3.10: **Power curves: single genetic variant, liability threshold model for binary traits, one binary trait and two quantitative traits:**


(a)

(b)

(c)

(d)

Power curves for different simulation scenarios: (a). Same Direction. (b). Opposite Direction. (c). Only Binary Trait. (d). Only Quantitative Traits.

3.5.4 Simulation Results: Problem of Ascertainment

In this section, we examine the simulation results under the scenario of ascertainment. Ascertainment arises when certain individuals in the target population have a higher or lower likelihood of being included in the sample compared to others. This is particularly relevant for binary traits, where cases are often oversampled to increase statistical power. Consequently, ascertainment is a common issue in studies involving binary traits. GEMMA Wald test occasionally exhibits incorrect type I error control, Bonf-minP and GEMMA score test end to yield conservative p-values under the null hypothesis. Therefore, these methods are omitted from the evaluation.

We continue to investigate the scenario involving related individuals and population structure. Specifically, we simulate a large number of individuals across two subpopulations using a similar approach as described in 3.5.1. "Simulation Setting for Two Sub-population" part. The simulation involves two binary traits and one quantitative trait, generated based on the models discussed in 3.5.1 "Covariate and Trait Models" part. The parameters are chosen to ensure that the prevalences of the first and second binary traits are approximately 5%. Additionally, the sample correlation between the two binary traits is set to a reasonable amount. To introduce ascertainment, individuals are selected based on the first binary trait. From subpopulation 1, we randomly retain 250 cases and 250 controls, and similarly, from subpopulation 2, we randomly retain 250 cases and 250 controls. This results in a total of 500 cases and 500 controls for the first binary trait, ensuring a balanced case-control ratio. The simulation settings are designed to ensure a reasonable mix of cases and controls for the second binary trait as well. After we get the ascertainment individuals, we simulate genotypes and calculate the GRM estimate as described in Section 3.5.1. This is a phenotype-based ascertainment, which not only influences the sample composition but also introduces model misspecification, as the ascertainment process is not explicitly accounted for in the simulation models.

3.5.4.1 Data Simulated Based on Logistic Model for Binary Traits Before Ascertainment

The correlation matrix we used to simulate data is

$$C = \begin{bmatrix} 1 & -0.3 & 0.5 \\ -0.3 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}.$$

The remaining parameters are chosen such that:

1. The simulated data before ascertainment result in a prevalence of approximately 5% for the two binary traits.
2. The sample correlation matrix for the three traits is approximately:

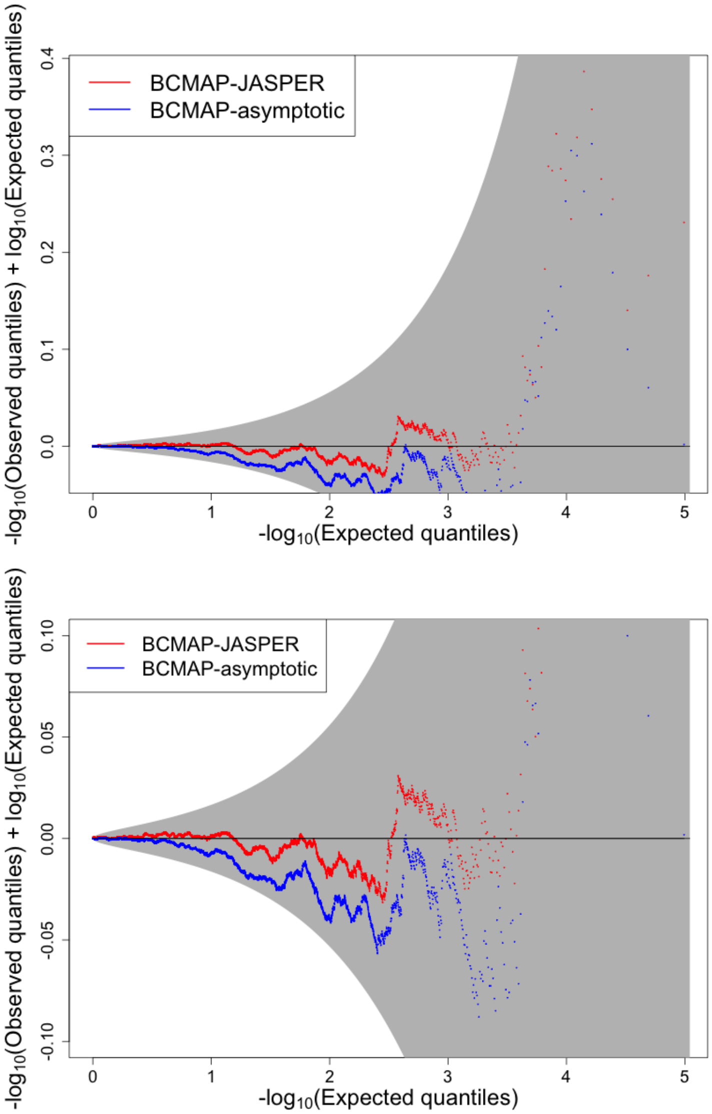
$$\begin{bmatrix} 1 & 0.6 & 0.7 \\ 0.6 & 1 & 0.5 \\ 0.7 & 0.5 & 1 \end{bmatrix}.$$

   so the sample correlation between the two binary traits is about 0.6.

Type I error results

We simulate the data under the null hypothesis that there is no association between the three traits and the causal SNP. Subsequently, we perform ascertainment, followed by the association test to obtain the p-values. Figure 3.11 shows BCMAP controls type I error correctly.

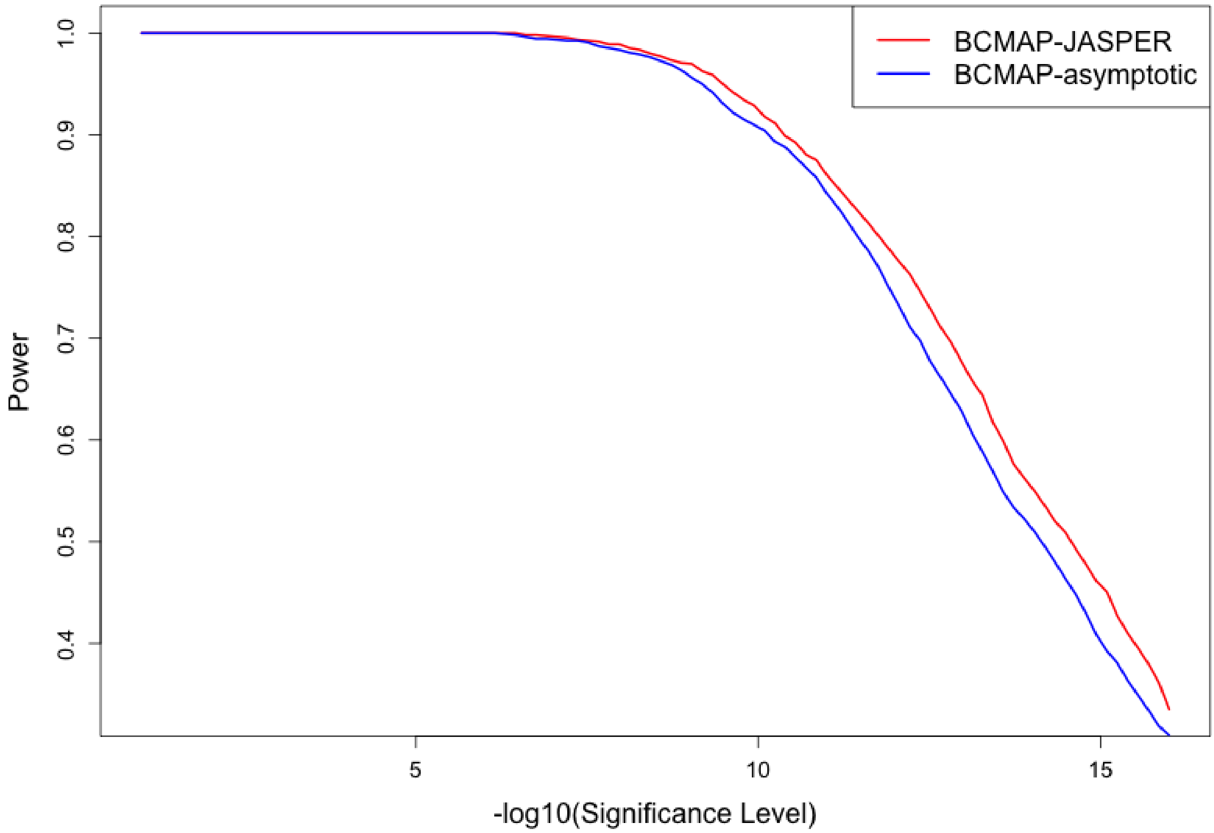
Figure 3.11: **(Differenced) QQ-plots for p-values: ascertainment, single genetic variant, logistic model for binary traits, two binary traits and one quantitative trait:** Top: original scale; bottom: zoomed in. The shaded region is the 99% confidence region by ELL



Power analysis results

The power simulation under ascertainment is computationally intensive and requires significant time. Since previous simulations indicate that BCMAP achieves the highest power under the "Opposite Direction" setting, we conducted the simulation only for this scenario. The variance explained by the causal SNP is set to be 7% in this case. Figure 3.12 demonstrates that BCMAP can detect signals even under ascertainment.

Figure 3.12: **Power curves: ascertainment, single genetic variant, logistic model for binary traits, two binary traits and one quantitative trait, "Opposite Direction"**



3.5.4.2 Data Simulated Based on Liability Threshold Model for Binary Traits Before Ascertainment

The correlation matrix we used to simulate data is

$$C = \begin{bmatrix} 1 & -0.3 & 0.5 \\ -0.3 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}.$$

The remaining parameters are chosen such that:

1. The simulated data before ascertainment result in a prevalence of approximately 5% for the two binary traits.

2. The sample correlation matrix for the three traits is approximately:

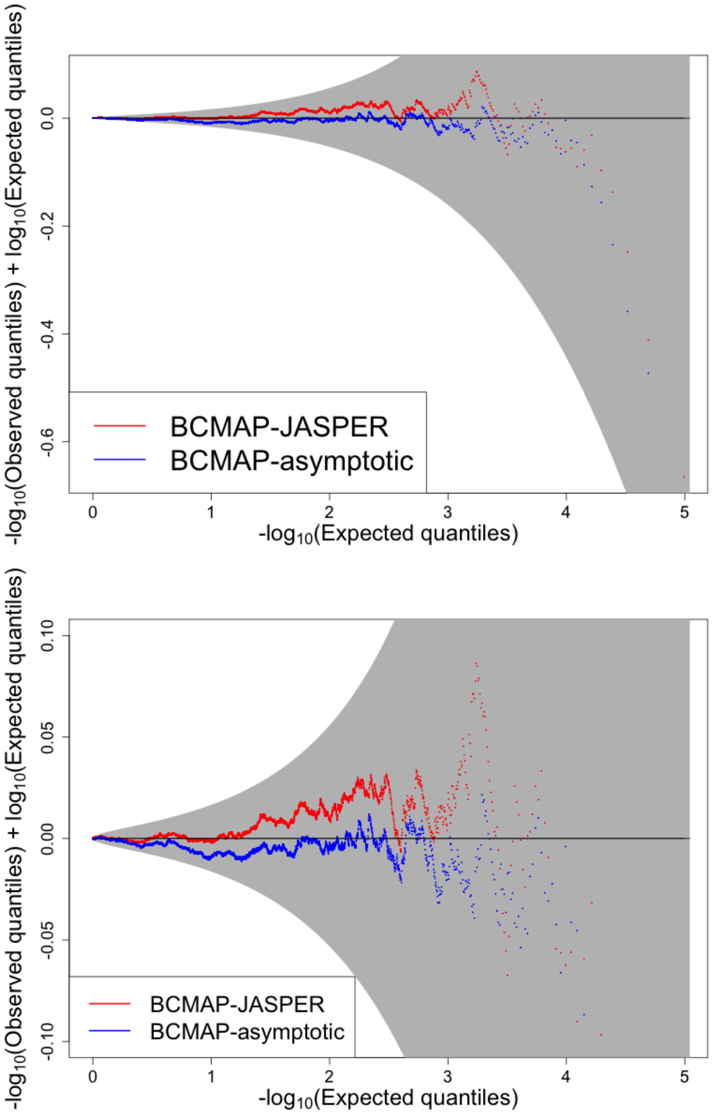
$$\begin{bmatrix} 1 & 0.4 & 0.7 \\ 0.4 & 1 & 0.3 \\ 0.7 & 0.3 & 1 \end{bmatrix}.$$

   so the sample correlation between the two binary traits is about 0.4.

Type I error results

We simulate the data under the null hypothesis that there is no association between the three traits and the causal SNP. Subsequently, we perform ascertainment, followed by the association test to obtain the p-values. Figure 3.13 shows BCMAP controls type I error correctly.

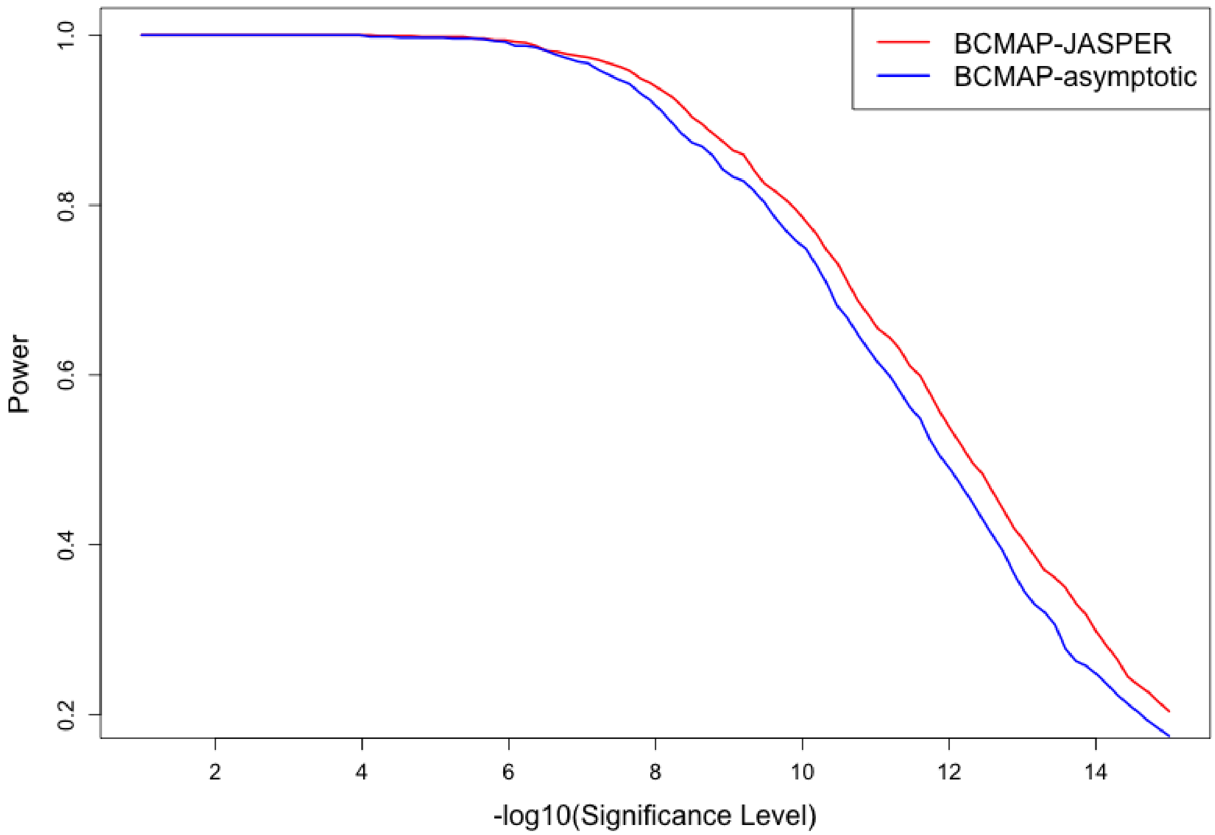
Figure 3.13: **(Differenced) QQ-plots for p-values: ascertainment, single genetic variant, liability threshold model for binary traits, two binary traits and one quantitative trait:** Top: original scale; bottom: zoomed in. The shaded region is the 99% confidence region by ELL



Power analysis results

We simulate data under the "Opposite Direction" setting. Figure 3.14 demonstrates that BCMAP can detect signals even under ascertainment.

Figure 3.14: **Power curves: ascertainment, single genetic variant, liability threshold model for binary traits, two binary traits and one quantitative trait, "Opposite Direction"**



3.5.5 Discussion of Simulation Results

From the simulations we conducted, we observe that BCMAP is robust to model misspecification and ascertainment. When binary traits are associated with the causal SNPs, BCMAP provides higher power compared to GEMMA and Bonf-minP. Conversely, when only quantitative traits are associated with the causal SNPs, BCMAP demonstrates comparable power to GEMMA and Bonf-minP. Therefore, in scenarios where a binary trait of interest is present, applying BCMAP may result in greater power.

CHAPTER 4

QUASI-LIKELIHOOD MODEL FOR MULTIPLE BINARY TRAITS AND QUANTITATIVE TRAITS WITH MULTIPLE GENETIC VARIANTS

The model can be easily extended to include multiple genetic variants. This extension is designed to perform genetic association testing for multiple traits with multiple genetic variants simultaneously. Assume there are n individuals and p traits of which b traits are binary and the rest are quantitative. And there are g genetic variants being tested, so G is a $n \times g$ matrix.

For mean structure:

$$E(Y_{ij}|X,G) = \mu_{ij}, \quad g(\mu_{ij}) = (\beta X)_{ij}, for\ 1 \leqslant i \leqslant b, \tag{4.1}$$

$$E(Y_{ij}|X,G) = \mu_{ij}, \quad \mu_{ij} = (\beta X)_{ij}, for\ 1+b \leqslant i \leqslant p, \tag{4.2}$$

where

$$g(\mu_{ij}) = log\frac{\mu_{ij}}{1-\mu_{ij}}.$$

For the conditional variance structure, we have:

$$\Omega := Var(vec(Y)|X,G) = \Gamma^{1/2}\Sigma\Gamma^{1/2}. \tag{4.3}$$

Γ is an np-dimensional diagonal matrix, with s^{th} diagonal element, where $s = p(j-1)+i$, given by $\Gamma_{ss} = Var(Y_{ij}|X,G)$, which is equal to $\mu_{ij}(1-\mu_{ij})$ if $1 \leqslant i \leqslant b$ and σ_i^2 if $1+b \leqslant i \leqslant p$, where σ_i^2 represents the total residual variance of trait i. Instead of model genotypes as fixed effects, we model them as random effects in Σ so that we have more degrees of freedom.

$$\Sigma = \tau^2[(GW^2G^T) \otimes (D^{1/2}C_gD^{1/2})] + K \otimes (D^{1/2}C_gD^{1/2}) + I_{n\times n} \otimes (\check{D}^{1/2}C_e\check{D}^{1/2}) \quad (4.4)$$

τ^2 is the parameter of interest that controls the ratio of proportion of residual variance due to the genetic variant effects vs. proportion of residual variance due to iid noise, W is an optional $g \times g$ weight matrix for the variants and $\check{D} = I - (\tau^2 + 1)D$, where we impose the additional constraint $0 \leqslant (\tau^2 + 1)d_i \leqslant 1$ for $1 \leqslant i \leqslant p$. We are interested in testing the null hypothesis that $\tau^2 = 0$. Under the null, the model of multiple genetic variants reduces to the same null model as the single variant one, so the estimation of nuisance parameters (coefficient and variance components) under the null proceeds in exactly the same way as before.

4.1 Retrospective Association Testing: JASPER

When testing multiple traits with multiple genetic variants, asymptotic approximations may fail in scenarios involving high-dimensional outcomes [33, 34] or limited sample sizes [35, 36]. JASPER [24] is a more robust method so we propose to utilize JASPER for our association test. Recall JAPSER relies on the transformation of the test statistic based on a null model proposed for the genetic markers. We build a quasi-likelihood model for G conditional on Y and X under the null hypothesis of no association, which is specified by the following assumptions:

$$E_0(G|X,Y) = X^TB \text{ and } Var_0(G|X,Y) = F \otimes K, \quad (4.5)$$

where B is an unknown coefficients matrix, K is the genetic covariance among the individuals and F is a positive semi-definite matrix that represents the covariance among the genetic variants and we do not need to make any assumptions of F when applying JASPER. The retrospective test statistic for testing the null hypothesis $H_0 : \tau^2 = 0$ is $T = tr((S_G)(H^TAH))$

where H is the phenotype information matrix we have in single genetic variant case (3.12), A is chosen so that in the special case of all quantitative traits, we get what is in the JASPER [24].

4.1.1 Linear Mixed Models in JASPER

To determine A, we first look at the linear mixed model (LMM) for multiple quantitative traits in JASPER [24], expressed as:

$$Y = X\beta + G\gamma + \alpha + \epsilon, \tag{4.6}$$

where Y is the $n \times p$ phenotype matrix, X is the $n \times k$ covariate matrix, and G is the $n \times g$ genotype matrix. The term α represents additive polygenic random effects, where:

$$\text{vec}(\alpha) \sim N(0, V_a \otimes K), \tag{4.7}$$

with V_a being an unknown $p \times p$ positive definite matrix representing the covariance among traits due to additive polygenic effects, and K being the genetic relationship matrix. The random effects for the tested variants are represented by γ, an $g \times p$ matrix. Although the full distribution of γ is not specified, it is assumed that:

$$E[\text{vec}(\gamma)] = 0, \quad \text{and} \quad \text{Var}[\text{vec}(\gamma)] = \tau^2 V_g \otimes W, \tag{4.8}$$

where V_g is a $p \times p$ covariance matrix that is either pre-specified or set equal to V_a or V_e. W is a pre-specified $g \times g$ positive definite "weight matrix", and τ^2 is an unknown scalar. For this model, the conditional expectation and variance of Y given X and G are:

$$E[Y|X,G] = \mu, \quad \text{and} \quad \text{Var}[\text{vec}(Y)|X,G] = \Omega, \tag{4.9}$$

where $\mu = X\beta$ and:

$$\Omega = \tau^2 V_g \otimes (GWG^T) + V_a \otimes K + V_e \otimes I_{n\times n}. \tag{4.10}$$

To test the null hypothesis $H_0 : \tau^2 = 0$, which corresponds to a joint test of association between the p traits and the g SNPs, they propose to use

$$S_Y = \widetilde{Y}\widehat{V_g}\widetilde{Y}^T, \tag{4.11}$$

where $\widetilde{Y}$ is an $n \times p$ matrix given by:

$$\text{vec}(\widetilde{Y}) = \hat{\Omega}_0^{-1}\text{vec}(Y - \hat{\mu}), \tag{4.12}$$

with:

$$\text{vec}(\hat{\mu}) = \widetilde{X}(\widetilde{X}^T\hat{\Omega}_0^{-1}\widetilde{X})^{-1}\widetilde{X}^T\hat{\Omega}_0^{-1}\text{vec}(Y), \tag{4.13}$$

and:

$$\widetilde{X} = I_{p\times p} \otimes X, \quad \hat{\Omega}_0 = \widehat{V_a} \otimes K + \widehat{V_e} \otimes I_{n\times n}. \tag{4.14}$$

Here, $\widehat{V_a}$ and $\widehat{V_e}$ are the estimates of V_a and V_e under the null hypothesis. If V_g is pre-specified, then $\widehat{V_g} = V_g$. Otherwise, if V_g is set equal to V_a or V_e, then:

$$\widehat{V_g} = \widehat{V_a}, \quad \text{or} \quad \widehat{V_g} = \widehat{V_e}. \tag{4.15}$$

Therefore, one can use $S_Y = \widetilde{Y}\widehat{V_a}\widetilde{Y}^T$ or $S_Y = \widetilde{Y}\widehat{V_e}\widetilde{Y}^T$ to form the test statistic.

4.1.2 Application of JASPER on BCMAP Multiple Genetic Variants Case

Genotype Kernel

We still assume that X_G, as defined in Section 3.3.2, represents the intercept. Additionally, we assume that Δ is a diagonal matrix, where each diagonal element is defined as: $\Delta_{ii} = \frac{1}{2\hat{f}_i(1-\hat{f}_i)}$, where $\hat{f}_i$ denotes the estimated allele frequency for genetic variant i. Therefore, $S_G = (I - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)G\Delta G^T(I - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)^T$.

Phenotype Kernel

Recall for phenotype kernel we have $S_Y = H^T AH$, A is chosen so that in the special case of all quantitative traits, we get what is in the JASPER [24]. We have two choices, setting A to be $\widehat{V_a}$, the estimated additive polygenic traits covariance or setting A to be $\widehat{V_e}$, the estimated environment traits covariance. For simplicity we assume $C_g = C_e$.

Let $\widehat{D_0}$ and $\widehat{C_0}$ be the estimations of D and C under the null model. $\widehat{J_0}$ is the matrix defined in (3.6), evaluated under the null. $\widehat{\Sigma_0}$, $\widehat{\Gamma_0}$, and $\widehat{\mu_0}$ are the parameters evaluated under the null. Define:

$$\text{vec}(\widetilde{Y}) = \widehat{J_0}^{1/2}\widehat{\Sigma_0}\widehat{\Gamma_0}^{-1/2}[\text{vec}(Y) - \text{vec}(\widehat{\mu_0})], \tag{4.16}$$

$$\text{vec}(\widetilde{\widetilde{Y}}) = \widehat{\Sigma_0}\widehat{\Gamma_0}^{-1/2}[\text{vec}(Y) - \text{vec}(\widehat{\mu_0})]. \tag{4.17}$$

$$A = R^{1/2}\widehat{D_0}^{1/2}\widehat{C_0}\widehat{D_0}^{1/2}R^{1/2}, \tag{4.18}$$

where R is a $p \times p$ diagonal matrix with the i^{th} diagonal element given by:

$$R_{ii} = \frac{(\widetilde{\widetilde{Y}}^T K\widetilde{\widetilde{Y}})_{ii}}{(\widetilde{Y}^T K\widetilde{Y})_{ii}}. \tag{4.19}$$

If all the traits are quantitative, then:

$$\widehat{\Gamma_0} = \widetilde{\Gamma_0} \otimes I, \quad \widehat{J_0} = \widehat{\Gamma_0} \tag{4.20}$$

and:

$$\text{vec}(\widetilde{\widetilde{Y}}) = \widehat{\Gamma_0}^{1/2}\text{vec}(\widetilde{Y}) = (\widetilde{\Gamma_0}^{1/2} \otimes I)\text{vec}(\widetilde{Y}) = \text{vec}(I\widetilde{Y}\widetilde{\Gamma_0}^{1/2}). \tag{4.21}$$

Thus:

$$\widetilde{\widetilde{Y}} = \widetilde{Y}\widetilde{\Gamma_0}^{1/2}. \tag{4.22}$$

For the diagonal elements:

$$(\widetilde{\widetilde{Y}}^T K\widetilde{\widetilde{Y}})_{ii} = \sigma_i^2(\widetilde{Y}^T K\widetilde{Y})_{ii}. \tag{4.23}$$

This implies:

$$R_{ii} = \sigma_i^2, \quad R = \widetilde{\Gamma_0}, \tag{4.24}$$

so:

$$A = V_a \quad \text{in JASPER. Which is the trait covariance due to additive polygenetic effect.} \tag{4.25}$$

Likewise, if we define

$$A = R^{1/2}\widehat{\tilde{D}_0}^{1/2}\widehat{C_0}\widehat{\tilde{D}_0}^{1/2}R^{1/2}, \tag{4.26}$$

We have when all the traits are quantitative

$$A = V_e \quad \text{in JASPER. Which is the trait covariance due to environmental effect.} \tag{4.27}$$

4.2 Computational Complexity

Similar to the single genetic variant case, the primary computational challenge in BCMAP for multiple genetic variants is the eigen-decomposition of the $n \times n$ matrix K, which is necessary to account for related individuals and population structure. This step has a computational complexity of $O(n^3)$, but it only needs to be performed once per study. Notably, most methods that incorporate related individuals and population structure also require this eigen-decomposition. When the number of traits and covariates is relatively small, the computational complexity for parameter estimation and test statistic calculation is approximately $O(gn^2)$ for a set of g SNPs tested simultaneously.

4.3 Simulation Study

We conducted a series of simulations to evaluate the performance of our method, BCMAP for multiple genetic variants, to verify that it achieves correct type I error control and demonstrates strong power. MultiSKAT [25] is a method capable of incorporating related individuals, population structure, and covariate information to test multi-trait multi-variant associations for quantitative traits. However, based on our simulation results, when covariates are included in the model, the `MultiSKAT` R program produces unreliable p-values under the null hypothesis. Consequently, we do not compare this method with our approach. Instead, we compare our method to an alternative approach where all binary traits are modeled as quantitative traits under our model setting. We denote this approach as BCMAP-quan. Both method using JASPER for association tests. We have two choices for the A matrix in the phenotype kernel. We denote it as V_a when, for all traits being quantitative, $A = V_a$. Similarly, we denote it as V_e when, for all traits being quantitative, $A = V_e$. In the appendix of [22], the authors suggested that when multiple candidate phenotype kernels are available, one can consider a linear combination of these kernels with predefined weights. Following this approach, we also consider taking the average of the phenotype kernels with V_a and V_e plugged in, which we refer to as the `avg` method. Finally, we also consider taking the smaller p-value obtained with V_a and V_e plugged in, and applying Bonferroni correction to that p-value by multiplying it by 2 (capped at 1). We denote this approach as `Bonf`.

4.3.1 Simulation Settings

The simulation model is the same as what we have in single genetic variant case 3.5.1, except we have $g = 50$, that is, we test associations with 50 genetic variants simultaneously now. We still simulate data based on the two sub-populatoin setting described in 3.5.1.

Type I Error and Power Simulation Setting

Type I error simulations For type I error simulations, we simulated 10 sets of 10^5 variants for each setting and calculated the 10 corresponding GRM estimates based on Equation (3.26). For each set of 10^5 variants, they are randomly split into 2000 non-overlapping marker panels, each contains 50 marks. Phenotypes were re-simulated 100 times, and 100 marker panels were randomly selected to be tested against the simulated phenotypes. This process resulted in a total of 10^5 replicates.

Power simulations For power simulations, 10^5 variants were simulated only once for each setting, and the corresponding GRM estimate was calculated based on Equation (3.26). These variants were randomly divided into 2000 non-overlapping marker panels, each containing 50 markers. From these panels, one panel was selected at a time to test for association with the traits. Since it is unlikely that all markers in a panel are causal, half of the markers in the selected panel were randomly designated as causal. Among the causal markers, half were assigned positive effects, while the other half were assigned negative effects. Phenotypes were re-simulated for each iteration. This process was repeated 1000 times, resulting in 1000 replicates for evaluating power.

4.3.2 Simulation Results

We conducted simulations using both the logistic model for binary traits and the liability threshold model for binary traits. Both simulation settings produced qualitatively similar results. Therefore, in this thesis, we present only the results from simulations based on the logistic model for binary traits.

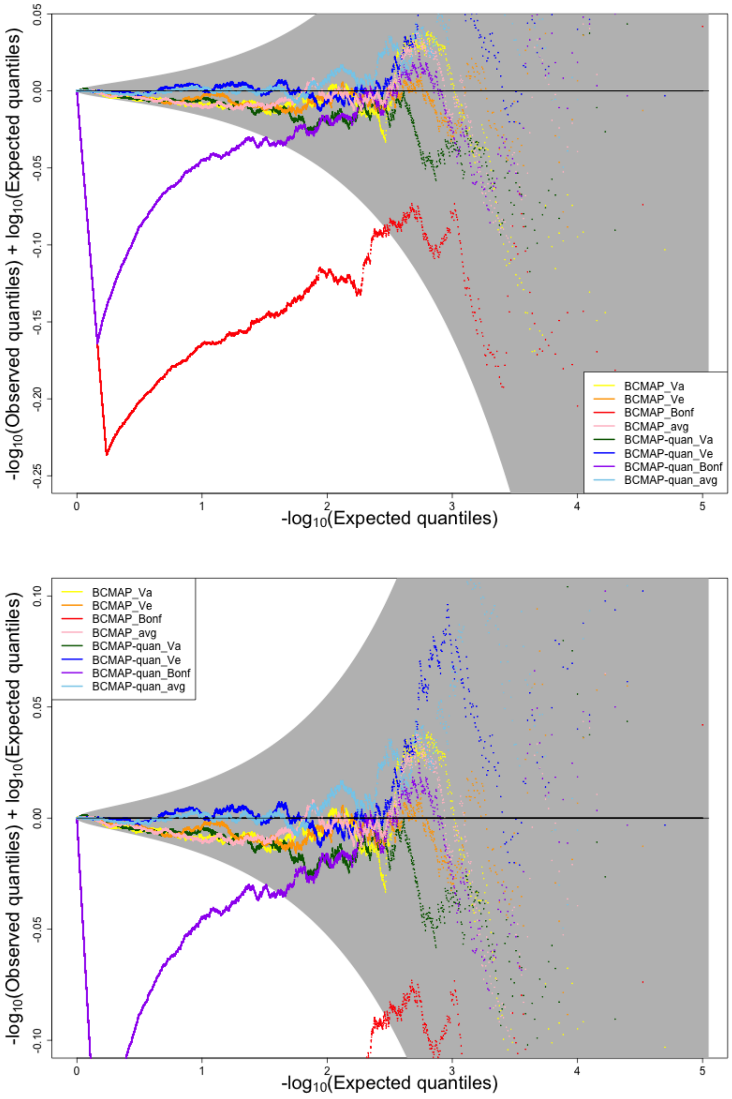
4.3.2.1 2 Binary Traits, 1 Quantitative Trait Case

We simulated 2 binary traits and 1 quantitative trait based on the logistic model for binary traits. Under the null hypothesis, we use the same set of parameters as in Setting 2 of the single genetic variant case for two binary traits and one quantitative trait, with the logistic model applied for the binary traits, as described in 3.5.2.1.

Type I error results

We simulate the data under the null hypothesis that there is no association between the three traits and the causal SNPs. Figure 4.1 demonstrates that both BCMAP and BCMAP-quan, with different phenotype kernels (excluding the Bonf method), correctly control the type I error. This result highlights the robustness of JASPER. However, the Bonf method provides conservative p-values under the null hypothesis.

Figure 4.1: **(Differenced) QQ-plots for p-values: multiple genetic variants, logistic model for binary traits, two binary traits and one quantitative trait:** Top: original scale; bottom: zoomed in. The shaded region is the 99% confidence region by ELL



Power analysis results

The parameter γ is chosen such that, on the logit scale, considering the variability explained by the covariates, the major causal variant, the additive polygenic effect (α_{ij}), and the causal SNPs, the proportion of variability explained by all the causal SNPs is approximately 3%. For the quantitative trait, all the causal SNPs explain approximately 3% of the total variance.

We are interested in three scenarios:

1. **All Traits**: All traits are associated with the causal SNPs.

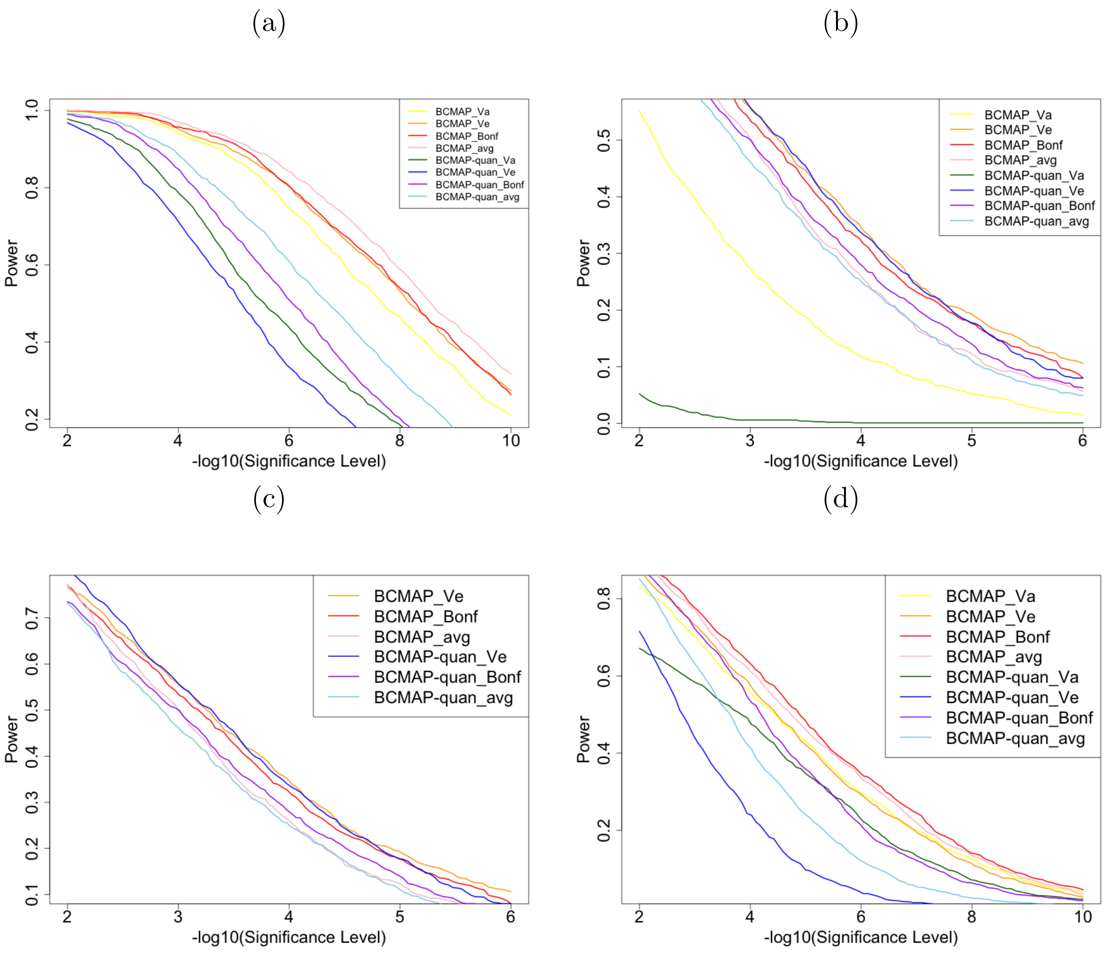
2. **Only Binary Traits**: Only the binary traits are associated with the causal SNPs.

3. **Random Two Traits**: At each trial, two traits are randomly selected to be associated with the causal SNPs.

For the case where only the binary traits are associated with the causal SNP, BCMAP with V_a plugged in (BCMAP_Va) and BCMAP-quan with V_a (BCMAP-quan_Va) plugged in provide lower power than others. Therefore, we also plot the power curves without these two methods to make a better comparison.

Figure 4.2 shows that under different settings, BCMAP with the Bonf method consistently provides either the largest or second-largest power compared to all other candidates. Therefore, when no prior information about the phenotype kernel is available, and binary traits are of interest, we recommend applying BCMAP with V_a and V_e plugged in, selecting the smaller p-value, and applying Bonferroni correction by multiplying it by 2 (capped at 1).

Figure 4.2: **Power curves: multiple genetic variants, logistic model for binary traits, two binary traits and one quantitative trait:**


(a)

(b)

(c)

(d)

Power curves for different simulation scenarios: (a). All Traits. (b). Only Binary Traits. (c). Only Binary Traits (exclude BCMAP_Va and BCMAP-quan_Va). (d). Random Two Traits.

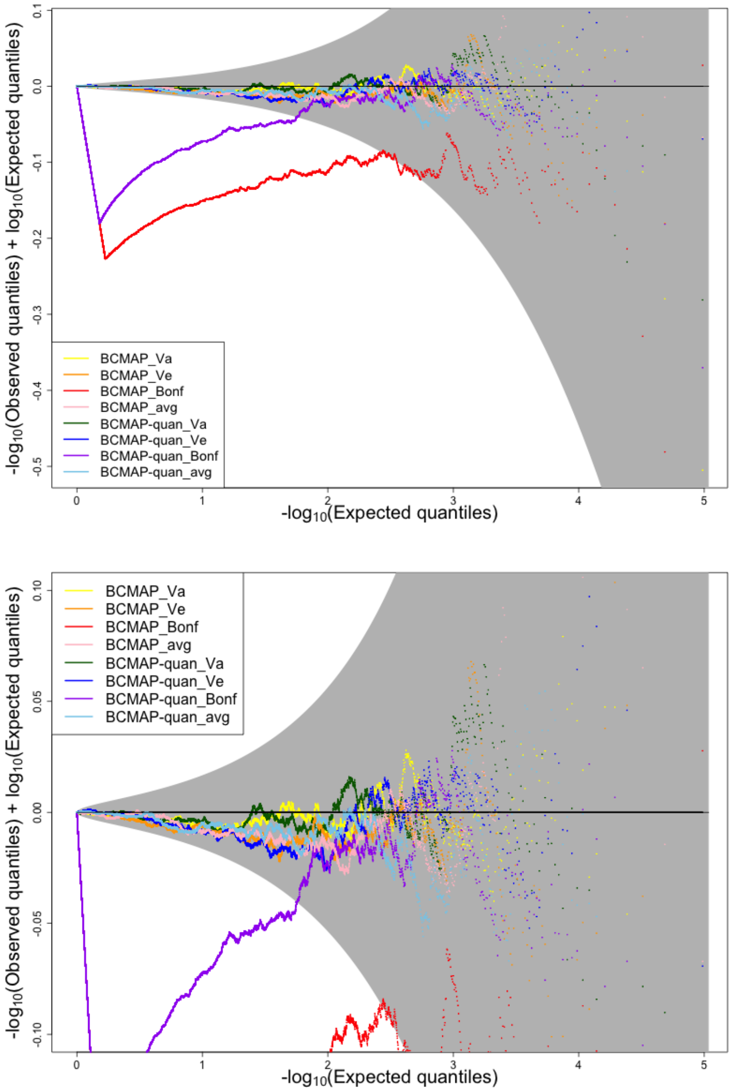
4.3.2.2 1 Binary Trait, 2 Quantitative Traits Case

We simulated 1 binary trait and 2 quantitative traits based on the logistic model for binary traits. Under the null hypothesis, we use the same set of parameters as in Setting 2 of the single genetic variant case for one binary trait and two quantitative trait, with the logistic model applied for the binary trait, as described in 3.5.2.2.

Type I error results

We simulate the data under the null hypothesis that there is no association between the three traits and the causal SNPs. Figure 4.3 demonstrates similar results as Figure 4.1

Figure 4.3: **(Differenced) QQ-plots for p-values: multiple genetic variants, logistic model for binary traits, one binary trait and two quantitative traits:** Top: original scale; bottom: zoomed in. The shaded region is the 99% confidence region by ELL



Power analysis results

We are interested in four scenarios:

1. **All Traits**: All traits are associated with the causal SNPs.

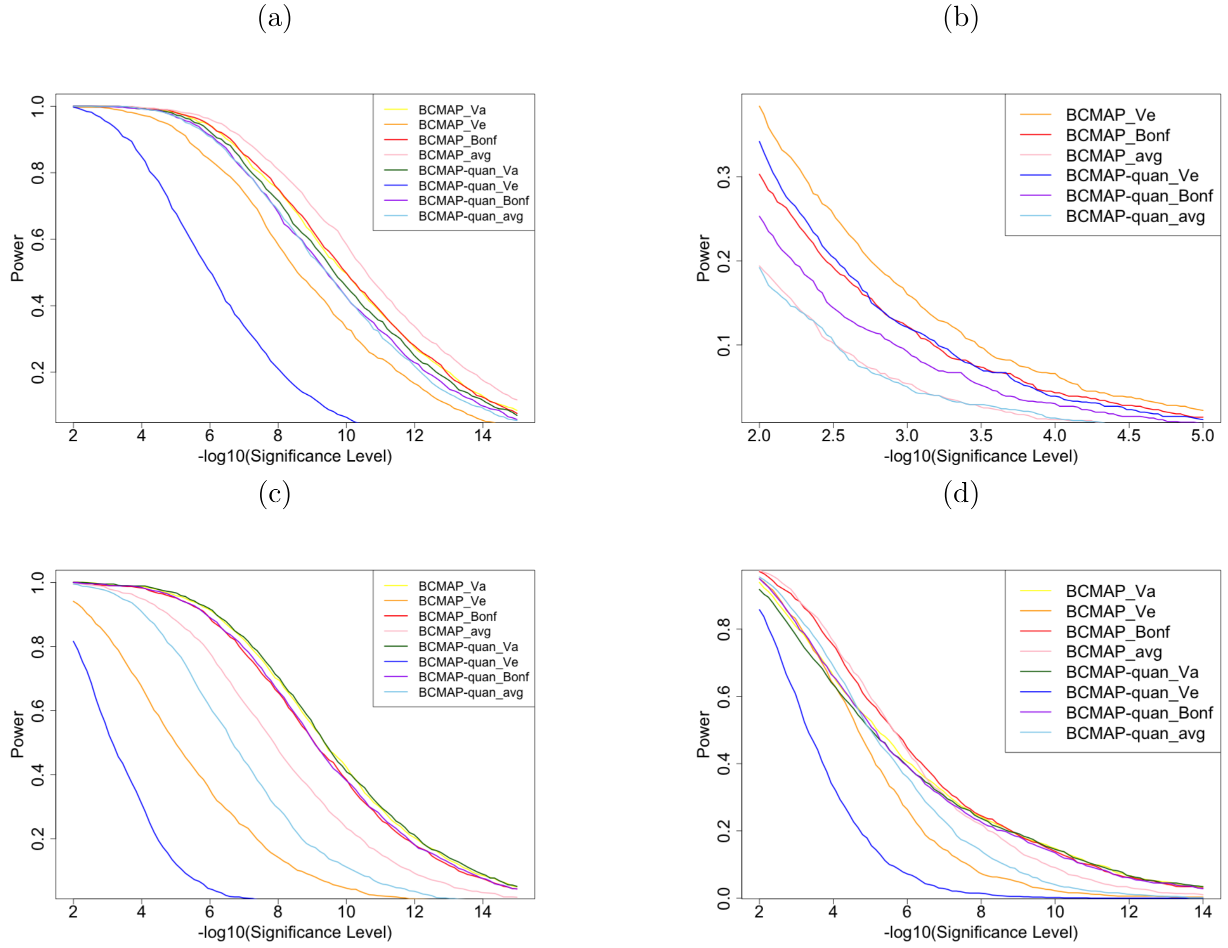
2. **Only Binary Trait**: Only the binary trait is associated with the causal SNPs.

3. **Only Quantitative Traits**: Only the quantitative traits are associated with the causal SNPs.

4. **Random Two Traits**: At each trial, two traits are randomly selected to be associated with the causal SNPs.

For the case where only the binary traits are associated with the causal SNPs, BCMAP with V_a plugged in (BCMAP_Va) and BCMAP-quan with V_a (BCMAP-quan_Va) plugged in provide extremely small power so we exclude these two in the plot. Figure 4.4 also demonstrates that, under different settings, BCMAP with the Bonf method consistently provides either the largest or second-largest power compared to all other candidates. This further supports our recommendation to use BCMAP with the Bonf method.

Figure 4.4: **Power curves: multiple genetic variants, logistic model for binary traits, one binary trait and two quantitative traits:**



Power curves for different simulation scenarios: (a). All Traits. (b). Only Binary Trait (exclude BCMAP_Va and BCMAP-quan_Va). (c). Only Quantitative Traits. (d). Random Two Traits.

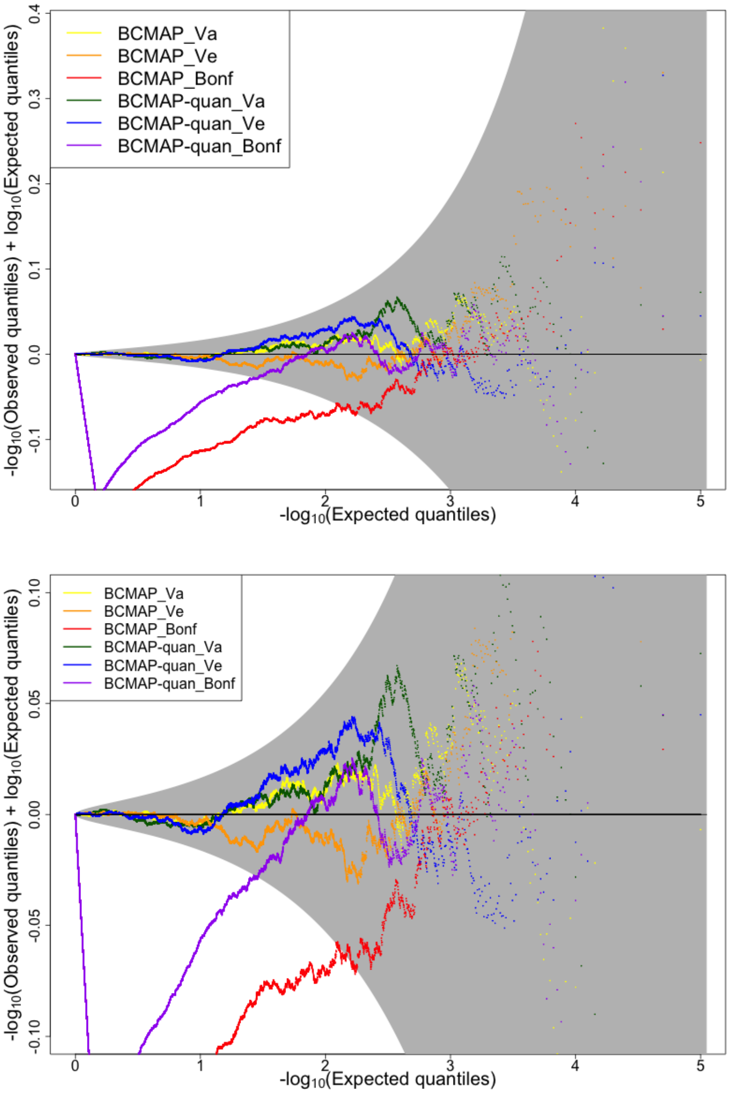
4.3.2.3 Ascertainment with 2 Binary Traits and 1 Quantitative Trait

We simulate the data with ascertainment as discussed in Section 3.5.4. Under the null hypothesis, we use the same set of parameters specified in that section.

Type I error results

We simulate the data under the null hypothesis that there is no association between the three traits and the causal SNPs. Since `avg` method is not optimal in previous simulations, we do not include that in our analysis. Figure 4.5 demonstrates similar results as Figure 4.1.

Figure 4.5: **(Differenced) QQ-plots for p-values: ascertainment, multiple genetic variants, logistic model for binary traits, two binary traits and one quantitative trait:** Top: original scale; bottom: zoomed in. The shaded region is the 99% confidence region by ELL



Power analysis results

The parameter γ is chosen such that, on the logit scale, considering the variability explained by the covariates, the major causal variant, the additive polygenic effect (α_{ij}), and the causal SNPs, the proportion of variability explained by all the causal SNPs is approximately 7%. For the quantitative trait, all the causal SNPs explain approximately 7% of the total variance.

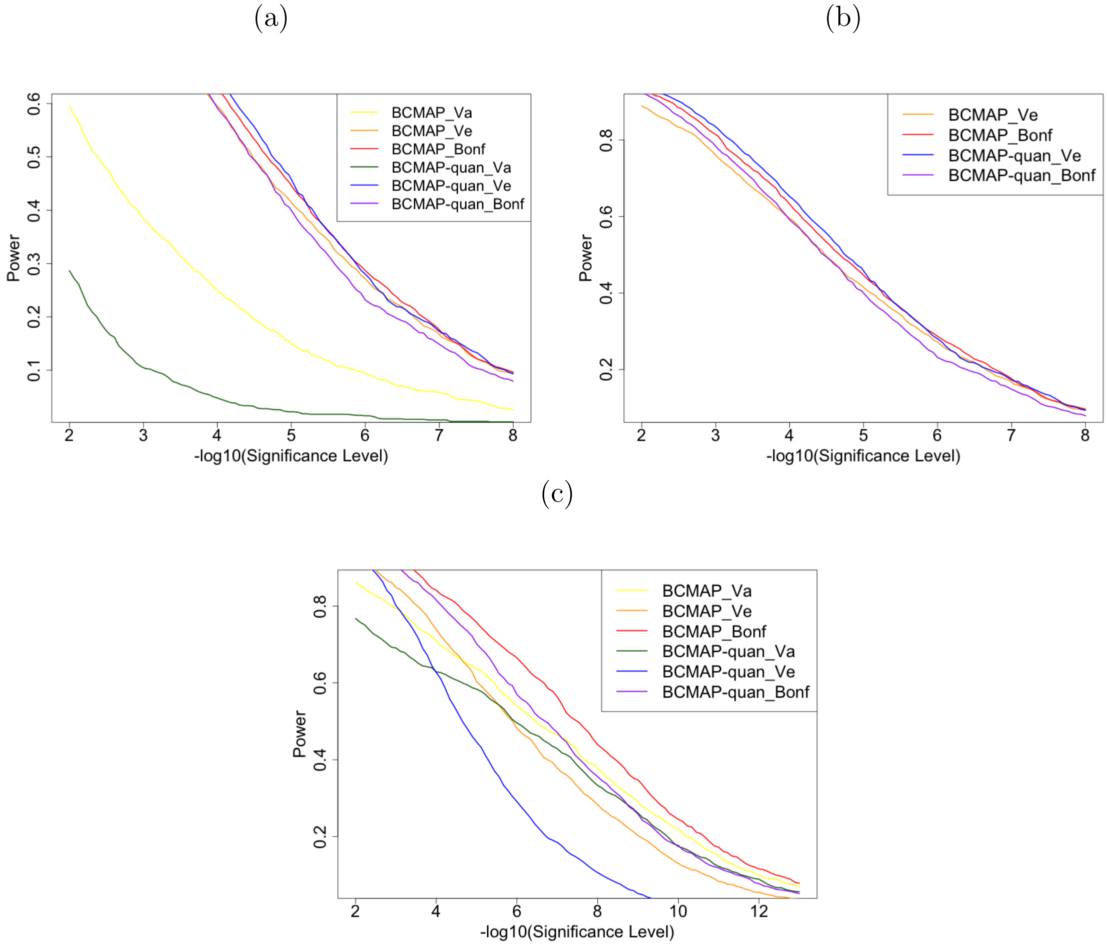
We are interested in two scenarios:

1. **Only Binary Traits**: Only the binary traits are associated with the causal SNPs.
2. **Random Two Traits**: At each trial, two traits are randomly selected to be associated with the causal SNPs.

For the case where only the binary traits are associated with the causal SNP, BCMAP with V_a plugged in (BCMAP_Va) and BCMAP-quan with V_a (BCMAP-quan_Va) plugged in provide lower power than others. Therefore, we also plot the power curves without these two methods to make a better comparison.

Figure 4.6 further highlights the consistent performance of BCMAP with the Bonf method, as it provides either the largest or second-largest power across different settings. This reinforces its utility as a robust approach in scenarios where phenotype kernel information is limited.

Figure 4.6: **Power curves: ascertainment, multiple genetic variants, logistic model for binary traits, two binary traits and one quantitative trait:**



Power curves for different simulation scenarios: (a). Only Binary Traits. (b). Only Binary Traits (exclude BCMAP_Va and BCMAP-quan_Va). (c). Random Two Traits.

4.3.3 Discussion of Simulation Results

From the simulation results, we observe that the BCMAP multiple genetic variants version is robust to model misspecification and ascertainment. There are multiple choices for phenotype kernels when conducting the test. In cases where no prior information about the phenotype kernel is available and binary traits are of interest, we recommend applying BCMAP with V_a and V_e plugged in, selecting the smaller p-value, and applying Bonferroni correction by multiplying it by 2 (capped at 1) to achieve greater power.

CHAPTER 5

DATA ANALYSIS

We apply BCMAP to analyze diabetes and body mass index (BMI) from the Framingham Heart Study (FHS) [37]. FHS is a a long-term observational study that spans multiple generations and includes both unrelated and related individuals. Our use of the FHS data was approved by the institutional review board of the Biological Sciences Division of the University of Chicago. Our analysis focuses on the Offspring Cohort. Participants in this cohort underwent measurements up to nine times, roughly every four years. Diabetes and BMI are the phenotypes of interest, while age and sex are the covariates included in the analysis. The phenotypes and covariates are determined as follows: For each exam, if an individual has a blood glucose (BG) level $\geqslant$ 200mg/dl, a fasting plasma glucose (FPG) level $\geqslant$ 126mg/dl, or is under treatment for diabetes, the individual is considered to have diabetes at that exam. We identify the earliest exam at which diabetes is detected and use the corresponding age and BMI as covariates. If age or BMI are unavailable for that exam, we select the nearest available values. However, if the nearest values are more than 10 years apart from the exam, the individual is excluded from the analysis. For individuals who never have diabetes across all exams (including those with some but not all missing exams), we use the age and BMI recorded at the individual's last attended exam as covariates. If this information is unavailable, we choose the nearest available values. Again, if the nearest values are more than 10 years apart from the last attended exam, the individual is excluded from the analysis.

Among the study participants with available Affymetrix 500K genotype data, we exclude individuals who meet either of the following criteria: (1) completeness (the proportion of markers with successful genotype calls) $\leqslant 96\%$, or (2) empirical self-kinship coefficient $\widehat{\Phi_{ii}} > 1.05$. Additionally, we exclude from our analysis SNPs that meet any of the following criteria: (1) call rate $\leqslant 96\%$, (2) Mendelian error rate $> 2\%$, or (3) minor allele frequency (MAF)

$< 1\%$. We impute any missing genotypes using IMPUTE2 [38]. We also restrict our analysis to SNPs located on autosomes. These quality control steps result in a final dataset of 380,364 SNPs and 3,372 individuals with genotype, phenotype and covariate data. The GRM estimate K is calculated using all the SNPs with an MAF greater than 5% by the equation (3.26) (L is no longer 10^5 in this case). The sample correlaton between BMI and diabetes is around 0.3.

5.1 Single Genetic Variant

We conducted a genome-wide association analysis between the SNPs and the combination of diabetes (binary) and BMI (quantitative) by applying the BCMAP single genetic variant version discussed in Section 3, including age and sex as covariates. We use both asymptotic and JASPER methods to conduct the association tests. We select all the SNPs with p-values less than 10^{-4} for both the asymptotic method and the JASPER method.

On chromosome 1, we identified a SNP near the gene **KCNJ10**, which resides in a region on chromosome 1q previously linked to type 2 diabetes in Pima Indians and six other populations [39]. Additionally, we found SNPs near the genes **LOC105378617**, **TMCO1**, and **TMCO1-AS1**, which have not been previously proposed to be related to diabetes or BMI. Notably, SNPs near **LOC105378617** have p-values smaller than the GWAS significance threshold (5×10^{-8}), suggesting an opportunity to study this gene further in the context of diabetes and BMI.

On chromosome 2, we identified SNPs near the gene **FMNL2**. Genome-wide association studies (GWAS) have reported SNPs in **FMNL2** associated with body height, bone density, and traits linked to diabetes, such as IgG glycosylation [40].

On chromosome 3, we identified a SNP near the gene **SEC22A**, where GWAS have reported associations between SNPs in this gene and body height [41]. We also found SNPs near the gene **ADCY5**, which has been linked to type 2 diabetes, body height, birth weight, fasting glucose, waist circumference adjusted for BMI, and BMI itself in previous GWAS studies [42].

On chromosome 5, we identified SNPs near the gene **ANXA6**. Currently, no studies associate **ANXA6** with diabetes or BMI, but further investigation may reveal potential links.

On chromosome 6, SNPs near the gene **RIPOR2** were identified. GWAS have associated SNPs in this gene with hip circumference adjusted for BMI, bone density, and body height [43].

On chromosome 7, we identified SNPs near the gene **STEAP2-AS1**. GWAS have reported associations between SNPs in **STEAP2-AS1** and type 2 diabetes, fasting glucose, and bone density [44].

On chromosome 8, we found SNPs near **LOC107986933**, but there is limited research on this gene.

On chromosome 10, we identified many SNPs near the gene **TCF7L2**. Studies have shown that this gene is associated with BMI and diabetes [45, 46].

On chromosome 20, we identified a SNP near the gene **ANGPT4**. No findings have associated **ANGPT4** with diabetes or BMI, warranting further investigation.

Finally, we also found SNPs that are not close to any known gene, and no findings have associated these SNPs with diabetes or BMI, indicating a need for additional studies.

Table 5.1: GWAS of BMI and Diabetes for FHS Offspring Cohort

SNP	**Chr**	**Position (GRCh37)**	**Nearest Gene**	**Asymptotic**	**JASPER**
rs6425866	1	30641646	*LOC105378617*	8.7×10^{-9}	3.0×10^{-8}
rs6704040	1	30644674	*LOC105378617*	8.7×10^{-9}	3.0×10^{-8}
rs17503555	1	182537599	*NA*	2.9×10^{-6}	3.3×10^{-6}

Table 5.1 (continued)

SNP	**Chr**	**Position (GRCh37)**	**Nearest Gene**	**Asymptotic**	**JASPER**
rs17568993	1	182537494	*NA*	4.6×10^{-6}	5.3×10^{-6}
rs7518099	1	165736880	*TMCO1, TMCO1-AS1*	3.0×10^{-5}	5.3×10^{-5}
rs4657476	1	165732661	*TMCO1*	3.0×10^{-5}	5.2×10^{-5}
rs17375748	1	160010151	*KCNJ10*	4.3×10^{-5}	5.9×10^{-5}
rs6733002	2	153200867	*FMNL2*	8.7×10^{-7}	2.6×10^{-6}
rs6741728	2	153268746	*FMNL2*	6.5×10^{-6}	1.8×10^{-5}
rs9823302	3	178140217	*NA*	2.0×10^{-6}	6.4×10^{-6}
rs7643790	3	122926556	*SEC22A*	2.2×10^{-5}	2.7×10^{-5}
rs9850375	3	123023615	*ADCY5*	7.0×10^{-5}	7.8×10^{-5}
rs4958895	5	150487195	*ANXA6*	2.1×10^{-5}	2.6×10^{-5}
rs4958893	5	150486991	*ANXA6*	3.0×10^{-5}	3.5×10^{-5}
rs673782	6	24964002	*RIPOR2*	1.8×10^{-5}	1.1×10^{-5}
rs365630	6	24967115	*RIPOR2*	1.8×10^{-5}	1.1×10^{-5}
rs11970548	6	14823968	*NA*	2.7×10^{-5}	4.0×10^{-5}
rs4711791	6	44598555	*NA*	2.8×10^{-5}	3.5×10^{-5}
rs432006	6	24972903	*RIPOR2*	4.3×10^{-5}	2.7×10^{-5}
rs10952976	7	89544391	*STEAP2-AS1*	9.2×10^{-6}	5.7×10^{-6}
rs6972809	7	89544278	*STEAP2-AS1*	1.3×10^{-5}	8.3×10^{-6}
rs11489497	7	89533047	*STEAP2-AS1*	4.3×10^{-5}	3.4×10^{-5}
rs17161595	7	9388999	*NA*	5.9×10^{-5}	7.5×10^{-5}
rs7832518	8	25649116	*LOC107986933*	9.0×10^{-6}	1.9×10^{-5}

Table 5.1 (continued)

SNP	**Chr**	**Position (GRCh37)**	**Nearest Gene**	**Asymptotic**	**JASPER**
rs1593403	8	25646700	*LOC107986933*	2.1×10^{-5}	4.0×10^{-5}
rs12243326	10	114788815	*TCF7L2*	1.7×10^{-7}	1.2×10^{-7}
rs4506565	10	114756041	*TCF7L2*	7.5×10^{-7}	6.4×10^{-7}
rs4132670	10	114767771	*TCF7L2*	9.9×10^{-7}	7.8×10^{-7}
rs10823687	10	72890441	*NA*	1.3×10^{-6}	1.5×10^{-6}
rs7901695	10	114754088	*TCF7L2*	1.4×10^{-6}	1.2×10^{-6}
rs7090550	10	72913623	*NA*	7.4×10^{-6}	8.3×10^{-6}
rs10885409	10	114808072	*TCF7L2*	8.1×10^{-5}	7.2×10^{-5}
rs11196205	10	114807047	*TCF7L2*	8.3×10^{-5}	7.3×10^{-5}
rs11196208	10	114811316	*TCF7L2*	8.6×10^{-5}	7.7×10^{-5}
rs6589086	11	109638891	*NA*	9.9×10^{-6}	1.6×10^{-5}
rs6486090	11	13128303	*NA*	8.9×10^{-5}	7.3×10^{-5}
rs4638447	13	64533688	*NA*	1.9×10^{-5}	1.7×10^{-5}
rs12461941	19	37984650	*NA*	3.1×10^{-5}	2.2×10^{-5}
rs910389	20	894722	*ANGPT4, LOC105372492*	7.5×10^{-6}	2.1×10^{-5}

5.2 Multiple Genetic Variants

To apply the BCMAP multiple genetic variants version, we utilize the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [47]. We identify the genes in Type II diabetes mellitus (KEGG ID: H00409), Type I diabetes mellitus (KEGG ID: H00408), and Genetic obesity (KEGG ID: H02106), as well as the genes in the pathways related to these diseases in KEGG. We only focus on the genes that are on autosome and total of 2507 genes are studied. We aim to test whether the cis-SNPs of these genes are associated with diabetes and BMI simultaneously. We find the GRCh37 positions of these genes. Then, we determine the starting and ending positions by extending 500 kb upstream and downstream of each gene (subtracting 500 kb from the starting position and adding 500 kb to the ending position of the gene), and the SNPs available within this region are designated as the cis-SNPs for each gene. We map the positions of SNPs in the Framingham Heart Study (FHS) using rs numbers to GRCh37 positions. Among the genes analyzed, the number of cis-SNPs per gene ranges from 14 to 433, with a median of 130 cis-SNPs.

We use V_a and V_e as the A matrix in the phenotype kernel as we did in simulations (Section 4.3). We did not detect any set of cis-SNPs that is significant. The smallest p-value with V_a plugged in is 2.7×10^{-4} for the gene *TLR4*, and the smallest p-value with V_e plugged in is 3.0×10^{-4} for the gene *CPA2*. The previously detected *TCF7L2* in single genetic analysis is included in the analysis, and its p-value is 0.38 with V_a plugged in and 0.06 with V_e plugged in. Using cis-SNPs is more common when studying gene expressions and may not be suitable for the type of analysis we conducted. Further study may be needed for this analysis.

Table 5.2: Number of Genes for each Pathway/Disease

Pathway/Disease ID	**Number of Genes**
H02106	23
H00408	22
H00409	13
hsa04930	45
hsa04110	149
hsa04115	72
hsa04350	105
hsa04911	83
hsa04972	99
hsa04330	61
hsa03320	72
hsa04310	166
hsa04141	156
hsa04714	196
hsa04923	54
hsa04935	119
hsa04940	42
hsa04659	104
hsa04658	89
hsa04672	45
hsa04060	285
hsa04630	159
hsa04151	348

CHAPTER 6

CONCLUSIONS

We developed BCMAP (Binary and Continuous Multi-trait Association test with Population structure), a novel modeling framework for multi-trait mapping of a combination of binary and quantitative phenotypes, based on a mixed-effects quasi-likelihood framework. BCMAP accommodates covariates, population structure, and relatedness, capturing the dichotomous nature of binary traits, and is suitable for testing both single and multiple genetic variants. Our test employs a retrospective approach and incorporates the recently proposed fast, powerful, and robust genetic association test method, JASPER.

We solve the challenging parameter estimation problem by employing useful parameterizations and utilizing the EM algorithm with Newton-Raphson updates. Additionally, we developed a method to evaluate this estimation procedure. Simulations for the single genetic variant version have shown that BCMAP is robust to model misspecification and ascertainment. When binary traits are associated with the causal SNP, BCMAP gains more power compared to existing methods. For multiple genetic variants tested with multiple traits simultaneously, the choice of phenotype kernel influences the results. We proposed several phenotype kernels to address this. Simulations demonstrated that the BCMAP multiple genetic variants version is robust to model misspecification and ascertainment. Additionally, when binary traits are associated with causal SNPs, modeling binary traits separately provides more power. We applied BCMAP to the Framingham Heart Study to analyze diabetes and BMI. For single genetic variant association tests, we identified several SNPs near genes known to be associated with diabetes, height, weight, or BMI. We also identified SNPs without prior knowledge, which could lead to further interest in studying these SNPs and nearby genes for BMI or diabetes. For multiple genetic variants association tests, we analyzed the cis-SNPs of genes known to be associated with Type I diabetes, Type II diabetes, and genetic obesity, as well as genes within pathways known to be associated with these diseases. We did not detect any set of cis-SNPs significantly associated with diabetes and BMI. Further study may be needed for this analysis.

REFERENCES

1. S. Wang, J. B. Meigs, and J. Dupuis. Joint association analysis of a binary and a quantitative trait in family samples. *European Journal of Human Genetics*, 25(1):130–136, 2017.

2. Matthew Stephens. A unified framework for association analysis with multiple related phenotypes. *PLoS One*, 8(7):e65245, 2013.

3. Kyoko Watanabe, Sven Stringer, Oleksandr Frei, and et al. A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics*, 51:1339–1348, 2019.

4. Andrew D. Grotzinger, Timothy T. Mallard, Wale A. Akingbuwa, and et al. Genetic architecture of 11 major psychiatric disorders at biobehavioral, functional genomic and molecular genetic levels of analysis. *Nature Genetics*, 54:548–559, 2022.

5. Xiaoqian Han, Puya Gharahkhani, Anne R. Hamel, and et al. Large-scale multitrait genome-wide association analyses identify hundreds of glaucoma risk loci. *Nature Genetics*, 55:1116–1125, 2023.

6. Ana L. Arruda, Alice Hartley, Georgia Katsoula, George D. Smith, Andrew P. Morris, and Eleftheria Zeggini. Genetic underpinning of the comorbidity between type 2 diabetes and osteoarthritis. *American Journal of Human Genetics*, 110(8):1304–1318, August 2023.

7. Yingjie Han, Jonghun Byun, Chao Zhu, and et al. Multitrait genome-wide analyses identify new susceptibility loci and candidate drugs to primary sclerosing cholangitis. *Nature Communications*, 14:1069, 2023.

8. Patrick Turley, Raymond K Walters, Omid Maghzian, Aysu Okbay, James J Lee, Mark A Fontana, Tuan Anh Nguyen-Viet, Robbee Wedow, Markus Zacher, Nicholas A Furlotte, 23andMe Research Team, Social Science Genetic Association Consortium, Patrik Magnusson, Sven Oskarsson, Magnus Johannesson, Peter M Visscher, David Laibson, David Cesarini, Benjamin M Neale, and Daniel J Benjamin. Multi-trait analysis of genome-wide association summary statistics using mtag. *Nature Genetics*, 50(2):229–237, 2018. Erratum in: Nat Genet. 2019 Jul;51(7):1190. doi: 10.1038/s41588-019-0444-5. Erratum in: Nat Genet. 2019 Aug;51(8):1295. doi: 10.1038/s41588-019-0469-9.

9. Jamie E Craig, Xiaohu Han, Ayham Qassim, and et al. Multitrait analysis of glaucoma identifies new risk loci and enables polygenic prediction of disease susceptibility and progression. *Nature Genetics*, 52:160–166, 2020.

10. Chen Xu, Santhi K Ganesh, and Xiaojing Zhou. mtpgs: Leverage multiple correlated traits for accurate polygenic score construction. *American Journal of Human Genetics*, 110(10):1673–1689, 2023.

11. D. Jiang, S. Zhong, and M. S. McPeek. Retrospective binary-trait association test elucidates genetic architecture of crohn disease. *American Journal of Human Genetics*, 98(2):243–255, 2016.

12. S. Zhong, D. Jiang, and M. S. McPeek. Ceramic: Case-control association testing in samples with related individuals, based on retrospective mixed model analysis with adjustment for covariates. *PLoS Genetics*, 12(10), 2016.

13. H. Chen, C. Wang, M. P. Conomos, A. M. Stilp, Z. Li, T. Sofer, A. A. Szpiro, W. Chen, J. M. Brehm, J. C. Celedón, S. Redline, G. J. Papanicolaou, T. A. Thornton, C. C. Laurie, K. Rice, and X. Lin. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *American Journal of Human Genetics*, 98(4):653–666, 2016.

14. D. Jiang, J. Mbatchou, and M. S. McPeek. Retrospective association analysis of binary traits: Overcoming some limitations of the additive polygenic model. *Human Heredity*, 80(4):187–195, 2015.

15. X. Zhou and M. Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, 11(4):407–409, 2014.

16. L. Luo, J. Shen, H. Zhang, et al. Multi-trait analysis of rare-variant association summary statistics using mtar. *Nature Communications*, 11:2850, 2020.

17. Il-Youp Kwak and Wei Pan. Gene- and pathway-based association tests for multiple traits with gwas summary statistics. *Bioinformatics*, 33(1):64–71, January 2017.

18. Anna Cichonska, Juho Rousu, Pekka Marttinen, Antti J. Kangas, Pasi Soininen, Terho Lehtimäki, Olli T. Raitakari, Marjo-Riitta Järvelin, Veikko Salomaa, Mika Ala-Korpela, Samuli Ripatti, and Matti Pirinen. metacca: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics*, 32(13):1981–1989, July 2016.

19. Sophie Van der Sluis, Conor V. Dolan, Jingmei Li, Yanbo Song, Pak Sham, Danielle Posthuma, and Menghui X. Li. MGAS: a powerful tool for multivariate gene-based genome-wide association analysis. *Bioinformatics*, 31(7):1007–1015, April 2015.

20. Kelly A. Broadaway, Don J. Cutler, Raphael Duncan, J. Lester Moore, Erin B. Ware, Min A. Jhun, Lawrence F. Bielak, Wei Zhao, Jennifer A. Smith, Patricia A. Peyser, Sharon L. R. Kardia, Debashis Ghosh, and Michael P. Epstein. A statistical approach for testing cross-phenotype effects of rare variants. *American Journal of Human Genetics*, 98(3):525–540, March 2016.

21. Biao Wu and James S. Pankow. Sequence kernel association test of multiple continuous phenotypes. *Genetic Epidemiology*, 40(2):91–100, February 2016.

22. Xianqi Zhan, Na Zhao, Anna Plantinga, Timothy A. Thornton, Karen N. Conneely, Michael P. Epstein, and Michael C. Wu. Powerful genetic association analysis for common or rare variants with high-dimensional structured traits. *Genetics*, 206(4):1779–1790, August 2017.

23. Yifan Wen and Qing Lu. An optimal kernel-based multivariate u-statistic to test for associations with multiple phenotypes. *Biostatistics*, 23(3):705–720, July 2022.

24. Joelle Mbatchou and Mary Sara McPeek. Jasper: Fast, powerful, multitrait association testing in structured samples gives insight on pleiotropy in gene expression. *American Journal of Human Genetics*, 111(8):1750–1769, August 2024. Epub 2024 Jul 17.

25. D. Dutta, L. Scott, M. Boehnke, and S. Lee. Multi-skat: General framework to test for rare-variant association with multiple phenotypes. *Genetic Epidemiology*, 43(1):4–23, Feb 2019. Epub 2018 Oct 8.

26. Wikipedia contributors. Vectorization (mathematics).

27. Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44:821–824, 2012.

28. Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M. Kadie, Richard I. Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature Methods*, 8:833–835, 2011.

29. Matti Pirinen, Peter Donnelly, and Chris C. A. Spencer. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *The Annals of Applied Statistics*, 7(1):369–390, 2013.

30. I. Archakov and P. R. Hansen. A new parametrization of correlation matrices. *Econometrica*, 89(4):1699–1715, 2021.

31. Wikipedia contributors. Fisher information — wikipedia, the free encyclopedia.

32. E Weine, MS McPeek, and M Abney. Application of equal local levels to improve q-q plot testing bands with r package qqconf. *Journal of Statistical Software*, 106(10):1–10, 2023.

33. J. Chen, W. Chen, N. Zhao, M. C. Wu, and D. J. Schaid. Small sample kernel association tests for human genetic and microbiome association studies. *Genet Epidemiol*, 40(1):5–19, 2016.

34. X. Zhan, N. Zhao, A. Plantinga, T. A. Thornton, K. N. Conneely, M. P. Epstein, and M. C. Wu. Powerful genetic association analysis for common or rare variants with high-dimensional structured traits. *Genetics*, 206(4):1779–1790, 2017.

35. Michael C. Wu, Seunggeun Lee, Tian Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 89(1):82–93, July 2011.

36. Seunggeun Lee, Michael J. Emond, Michael J. Bamshad, Kathleen C. Barnes, Mark J. Rieder, Deborah A. Nickerson, NHLBI GO Exome Sequencing Project—ESP Lung Project Team, David C. Christiani, Mark M. Wurfel, and Xihong Lin. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics*, 91(2):224–237, August 2012.

37. Manning Feinleib, William B. Kannel, Robert J. Garrison, Patrick M. McNamara, and William P. Castelli. The framingham offspring study. design and preliminary data. *Preventive Medicine*, 4(4):518–525, December 1975.

38. Bryan N. Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6):e1000529, June 2009.

39. Vidya S. Farook, Robert L. Hanson, Johanna K. Wolford, Clifton Bogardus, and Michal Prochazka. Molecular analysis of kcnj10 on 1q as a candidate gene for type 2 diabetes in pima indians. *Diabetes*, 51(11):3342–3346, November 2002.

40. European Bioinformatics Institute GWAS Catalog. Fmnl2 gene - gwas catalog. `https://www.ebi.ac.uk/gwas/genes/FMNL2`, 2024. Accessed: 2024-12-05.

41. European Bioinformatics Institute GWAS Catalog. Sec22a gene - gwas catalog. `https://www.ebi.ac.uk/gwas/genes/SEC22A`, 2024. Accessed: 2024-12-05.

42. European Bioinformatics Institute GWAS Catalog. Adcy5 gene - gwas catalog. `https://www.ebi.ac.uk/gwas/genes/ADCY5`, 2024. Accessed: 2024-12-05.

43. European Bioinformatics Institute GWAS Catalog. Ripor2 gene - gwas catalog. `https://www.ebi.ac.uk/gwas/genes/RIPOR2`, 2024. Accessed: 2024-12-05.

44. European Bioinformatics Institute GWAS Catalog. Steap2-as1 gene - gwas catalog. `https://www.ebi.ac.uk/gwas/genes/STEAP2-AS1`, 2024. Accessed: 2024-12-05.

45. Lindsay Fernández-Rhodes, Ann G. Howard, Monda Graff, et al. Complex patterns of direct and indirect association between the transcription factor-7 like 2 gene, body mass index and type 2 diabetes diagnosis in adulthood in the hispanic community health study/study of latinos. *BMC Obesity*, 5:26, 2018.

46. European Bioinformatics Institute GWAS Catalog. Tcf7l2 gene - gwas catalog. `https://www.ebi.ac.uk/gwas/genes/TCF7L2`, 2024. Accessed: 2024-12-05.

47. Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.