ABSTRACT

TING, BRYAN W. Likelihood-Based Computational Approaches for Association Studies in Statistical Genetics. (Under the direction of Dr. Fred Wright and Dr. Yihui Zhou).

Decreasing sequencing costs and increasing computational power in recent years have provided an increasing availability of genomic data. Such increased availability of data invites greater demand for methodological sophistication in modeling, testing, and inference. Such methodological improvements can produce additional computational challenges and/or suffer from the lack of clear statistical frameworks.

We look to meet the demand for improvements with novel methods for genomics and precision medicine, that can also be applied in other settings. High priority concerns include computation time and statistical efficiency.

We present HAP-SAMPLE2, a generalization of the bioinformatics software HAP-SAMPLE, a tool for case/control data-resampling. We preserve the spirit and core functionality of HAP-SAMPLE, and extend it to incorporate functionality for admixed populations. The updated HAP-SAMPLE would flexibly take user inputs for assumptions such as disease prevalence and allele effect size to simulate admixed individuals under case/control or quantitative phenotype designs, for either common or rare variants.

We introduce a Two-Stage Composite Likelihood approach for multivariate probit estimation for responses with binary components. This approach is designed to be fast, lending itself to settings such as Genome Wide Association Studies (GWAS) for detecting Single Nucleotide Polymorphism (SNP) associations with mean changes in phenotype.

We extend this approach by incorporating consideration for SNP-associated heteroskedasticity (i.e. variance changes in phenotype) for where the multivariate responses can include both binary and continuous components. Such modeling is naturally more computationally intensive, and perhaps more appropriate for better-characterizing a narrower subset of SNPs. Simultaneous testing for mean and variance associations could serve as a subsequent step after first identifying candidate regions from the faster Two-Stage Composite Likelihood approach for mean associations.

Likelihood-Based Computational Approaches for Association Studies in Statistical Genetics

by
Bryan W. Ting

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Bioinformatics

Raleigh, North Carolina

2020

APPROVED BY:

Dr. Fred Wright Co-Chair of Advisory Committee	Dr. Yihui Zhou Co-Chair of Advisory Committee
Dr. Alison Motsinger-Reif External Member	Dr. Xinxia Peng
Dr. David Reif	

DEDICATION

To my family and friends, present and future.

BIOGRAPHY

Prior to the Bioinformatics PhD program at North Carolina State University, Bryan completed his undergrad at the New York University Stern School of Business. He graduated with a Bachelor of Science in 2011, having majored in Finance & Statistics, along with a minor in Anthropology. After graduation, he signed-on to the Advisory Analytics team at BlackRock in New York City. In 2015, he transitioned to North Carolina State University, where he received a Master of Statistics in 2017. Bryan also spent the summer of 2019 in Austin, as a Researcher in Investments Research at Dimensional Fund Advisors.

ACKNOWLEDGEMENTS

I wish to thank my advisors—Dr. Fred Wright and Dr. Yihui Zhou—for their direction and guidance over the years, which have passed all too quickly. I would also like to thank my other committee members, Dr. Alison Motsinger-Reif, Dr. Xinxia Peng, Dr. David Reif, and my graduate school representative Dr. Ronald Sederoff, for their feedback and advice throughout this experience. I had the privilege and pleasure of having some of them as my course instructors, as well, where I was able to benefit from their mentorship and knowledge in a different context.

Dr. Spencer Muse was one of the first people I met here at North Carolina State University—and in Raleigh. I thank him for his invaluable guidance and support during my time here, from the beginning. I want to thank Dana Ripperton and Jenni Wilson for their help in making this a smooth experience. I would also like to thank Kevin Dudley and Chris Smith for their advice with systems and computing at the BRC.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1

Introduction

To describe the research progression of statistical genetics in recent years as clockwork would likely be an understatement. Researchers in statistical genetics continue to produce impressive findings, both novel and confirmatory. Genomics companies continue to develop methods, expand product offerings, and extend market-reach, while nascent start-ups emerge in anticipation of new niches and opportunities.

Large-scale data-sets and resources, both proprietary and public, provide a myriad of avenues for researchers to explore research pathways. Publicly available data-sets include 1000 Genomes, which contains 84.4 million autosomal variants from 26 human populations (Auton, A. and Abecasis, G. and Altshuler, D. et al., 2015). Tools such as the UCSC's genome browser provide drill-down capabilities for user-specified regions of interest (Kent et al., 2002). Software packages in R and Python provide a plethora of options for statistical and machine learning, many of which are tailored toward bioinformatics and statistical genetics.

Despite these advancements, classical issues such as statistical power and multiple-testing remain key concerns in better utilizing genomic data-sets and reviewing study designs (Yang and Wang, 2012; Wright et al., 2007). It has been widely-noted that in the past decade and a half, sequencing costs have abided by—and even surpassed—Moore's Law (Muir et al., 2016). With the cost per genome declining each year, perhaps at an exponential pace, data has become increasingly available at cost-effective rates. Genomic data-sets are growing in two dimensions, with more observations in one dimension and more variants in the other.

Larger data-sets and more powerful computational hardware enable more robust analyses and greater ambition in study designs. Yet, contemporaneous advances in computational power may be insufficient for researchers to fully utilize the data available. Increasingly paramount are issues of scale, storage, and speed for methods in genomics (November, 2018; Muir et al., 2016). Concerns with replicability—in the aftermath of a replication "crisis" in academia—have seen an increasing researcher mind-share (Fanelli, 2018; Fidler and Wilcox,

2018).

1.1 Likelihood

The likelihood function is a "natural starting point" for many topics statistical in nature—and/or otherwise involving data—serving as a foundation in many areas of statistics (Boos and Stefanski, 2013; Hastie et al., 2017). In a typical likelihood set-up, we consider the likelihood of a data-set as the product of the likelihood of its observations. In turn, the likelihood of each observation is equal to a probability density or mass at that observation. Letting L represent our likelihood function:

$$L(\boldsymbol{\theta}; \boldsymbol{Z}) = \prod_{i}^{N} L(\boldsymbol{\theta}; \boldsymbol{z_i}) = \prod_{i}^{N} f(\boldsymbol{z_i}; \boldsymbol{\theta})$$

Where $\boldsymbol{\theta}$ is a vector of parameters, $\boldsymbol{z_i}$ is the i'th observation (typically a row) from data-matrix $\boldsymbol{Z}$, with N total observations (Hastie et al., 2017). $f(\boldsymbol{z_i}; \boldsymbol{\theta})$ is the density or mass at that observation. Often, the likelihood is expressed via the log-likelihood (Hastie et al., 2017), which can be analytically and computationally more favorable:

$$\ln L(\boldsymbol{\theta}; \boldsymbol{Z}) = \ln \prod_{i}^{N} L(\boldsymbol{\theta}; \boldsymbol{z_i}) = \ln \prod_{i}^{N} f(\boldsymbol{z_i}; \boldsymbol{\theta}) = \sum_{i}^{N} \ln L(\boldsymbol{\theta}; \boldsymbol{z_i}) = \sum_{i}^{N} \ln f(\boldsymbol{z_i}; \boldsymbol{\theta})$$

1.1.1 Maximum Likelihood

Likelihood analyses provide many well-known results in inference, estimation (both point and interval), and hypothesis testing (Boos and Stefanski, 2013). An important tool from likelihood theory—and perhaps one of the most important in statistics in general—is maximum likelihood (White, 1982). In maximum likelihood estimation, we "[assume] that the most reasonable values" for our parameters, $\boldsymbol{\theta}$, are those which yield the highest value for the likelihood of the data-set (Hastie et al., 2017). We treat the likelihood (usually the log-likelihood) as an objective function upon which to perform optimization by finding values for

the parameters:

$$max_{\boldsymbol{\theta}} \ln L(\boldsymbol{\theta}; \boldsymbol{Z})$$

The solution to which is our maximum likelihood estimate, $\hat{\boldsymbol{\theta}}_{\boldsymbol{mle}}$.

Under certain conditions, and if the model is correctly specified, maximum likelihood estimation has many favorable properties. Two of which are statistical consistency and asymptotic efficiency. The maximum likelihood estimator (MLE) is consistent, in that the MLE for a given $\boldsymbol{\theta}$ converges in probability to the true value for $\boldsymbol{\theta}$, $\boldsymbol{\theta}_{\boldsymbol{true}}$, as sample size grows toward infinity (Griliches et al., 1983):

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{mle}} \xrightarrow{\text{p}} \boldsymbol{\theta}_{\boldsymbol{true}}$$

With stronger conditions than for convergence in probability, $\hat{\boldsymbol{\theta}}_{\boldsymbol{mle}}$ converges almost surely to the true value for $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{mle}} \xrightarrow{\text{a.s.}} \boldsymbol{\theta}_{\boldsymbol{true}}$$

Although not necessarily the case for finite samples, the MLE is asymptotically efficient. The MLE converges upon the normal distribution and attains the Cramér-Rao lower bound for variance as the sample size grows toward infinity (Griliches et al., 1983; Casella and Berger, 2017):

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\boldsymbol{mle}} - \boldsymbol{\theta}_{\boldsymbol{true}}) \xrightarrow{\text{d}} \phi\left(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\theta})^{-\mathbf{1}}\right)$$

That is, no other consistent estimator has lower asymptotic variance than that of the MLE. $\boldsymbol{I}(\boldsymbol{\theta})$ refers to the Fisher Information Matrix, where a typical element $I(\boldsymbol{\theta})_{jk}$ is the expected value of the outer product of the log-likelihood gradients corresponding to elements j and k of $\boldsymbol{\theta}$:

$$I(\boldsymbol{\theta})_{jk} = J(\boldsymbol{\theta})_{jk} = E\left[\left(\frac{d}{d\theta_j} \ln f(\boldsymbol{Z}; \boldsymbol{\theta})\right)\left(\frac{d}{d\theta_k} \ln f(\boldsymbol{Z}; \boldsymbol{\theta})\right)\right]$$

Furthermore, under certain regularity conditions, and if $\ln f(\boldsymbol{Z}; \boldsymbol{\theta})$ is twice-differentiable, the Fisher Information Matrix can be stated as the negative expected value of the second derivatives (Casella and Berger, 2017; Cattelan and Sartori, 2016):

$$I(\boldsymbol{\theta})_{jk} = H(\boldsymbol{\theta})_{jk} = -E\left[\frac{d^2}{d\theta_j d\theta_k} \ln f(\boldsymbol{Z}; \boldsymbol{\theta})\right]$$

1.1.2 Some Considerations for Maximum Likelihood Estimation

If the likelihood is misspecified and/or if other conditions are otherwise not met, then the expected properties of maximum likelihood may not hold. In particular, $\boldsymbol{J}(\boldsymbol{\theta})$ from above may not necessarily equal $\boldsymbol{H}(\boldsymbol{\theta})$ and the inverse Fisher Information Matrix $\boldsymbol{I}(\boldsymbol{\theta})^{-1}$ may no longer be appropriate as the asymptotic covariance matrix. Under more general conditions, however, the Godambe Information Matrix, $\boldsymbol{G}(\boldsymbol{\theta})$, can be used instead (Cattelan and Sartori, 2016). That is, $\boldsymbol{G}(\boldsymbol{\theta})^{-1}$ can serve in place of $\boldsymbol{I}(\boldsymbol{\theta})^{-1}$ as the asymptotic covariance matrix, where:

$$\boldsymbol{G}(\boldsymbol{\theta})^{-1} = \boldsymbol{H}(\boldsymbol{\theta})^{-1}\boldsymbol{J}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})^{-1}$$

Many popular regression methods, such as the logistic regression, are typically estimated via maximum likelihood (Hastie et al., 2017; Menard, 2002). Despite the favorable properties of maximum likelihood estimation, it can come with substantial challenges. Closed form solutions are available for certain distributions, e.g. if z is normally distributed; however, this is not always the case (Hastie et al., 2017). More cumbersome likelihoods can naturally present some substantial computational/numerical challenges (Hastie et al., 2017; Varin et al., 2011). Naturally, this would invite methods to serve as accompaniments and/or alternatives to maximum likelihood estimation. However, it is not guaranteed that such methods retain MLE-like behavior, thus potentially calling for further supplementary methodologies (Varin et al., 2011; Cattelan and Sartori, 2016; Hardin, 2002).

1.2 Genotype-Phenotype Association Studies

The association between genotype and phenotype has long-captured researcher interest, with wide-ranging applications from paleontology to the agricultural sciences (Rashid et al., 2014; Zuidhof et al., 2014). Modern-day research in human genotype-phenotype associations play massive roles in various fields, in particular psychiatry and medicine (Visscher et al., 2017). Several companies provide and/or have provided Direct-to-Consumer (DTC) genomics services—up to and including phenotypic inferences—albeit in recent years such companies

have been tempered by privacy and legal concerns, at least in the United States (Yim Seon-Hee, 2014). The data and analytical capabilities of these DTC genomics companies are hardly frivolous, as evidenced by GlaxoSmithKline's $300 million investment in 23andMe (Hirschler, 2018).

As genotyping technologies have matured, Genome Wide Association Study/Studies (GWAS) has/have become the archetype for genotype-phenotype association studies (Chanock et al., 2007). GWAS is typically performed one locus at a time, to identify candidate loci associated with a phenotype of interest (Bush and Moore, 2012). These candidate loci can then be used to make predictions for individuals using their genotypes, and/or to find regions to target for disease prevention and intervention. Multi-locus approaches have been attempted, but tend to become less and less tractable as the number of loci under simultaneous consideration grows larger. GWAS, in the usual "agnostic" formulation (Chanock et al., 2007), does not ascribe any ex-ante view on the viability of a particular variant as a candidate.

1.2.1 GWAS Implementation

GWAS is usually conducted using Single Nucleotide Polymorphisms (SNPs) (Bush and Moore, 2012; Visscher et al., 2017). SNPs are generally biallelic, i.e. there are two possible choices within a population (Bush and Moore, 2012). SNPs, as a unit of "*common* genetic variation," are typically defined via some minor allele frequency threshold for a given population (Bush and Moore, 2012). Thresholds for delineating common and rare variants are usually somewhere between 1% and 5% (Visscher et al., 2017; Budu-Aggrey et al., 2017).

Despite their names, common variants are less commonly found in the human genome, and rare variants more commonly. Analysis from the 1000 Genomes Project suggested that a typical human genome differs from the reference genome at 4 to 5 million sites. Out of the roughly 84 million autosomal variants, about 64 million were found to have a frequency of less than 0.5%, 12 million between 0.5% and 5.0%, and only 8 million above 5.0% (Auton, A. and Abecasis, G. and Altshuler, D. et al., 2015).

There are some practical and theoretic considerations behind the popular use of SNPs.

A given common variant captures more of the variation extant in a population than a given rare variant. Most ostensibly, a rare variant may have zero variance in a data-set, where all individuals in a given data-set are homozygous for a biallelic rare variant. The Common Disease/Common Variant Hypothesis (CD/CV) suggests common diseases are likely influenced by common genetic variation (Bush and Moore, 2012; Wright et al., 2007). Although this hypothesis has ample empirical support, it does not preclude the possibly of rare variants having a material impact, as well (Bush and Moore, 2012).

A usual feature of GWAS is some means by which to adjust for population stratification and linkage disequilibrium; neglecting to do so can result in a considerable number of false-positive GWAS hits (Bush and Moore, 2012). Different populations may have differing minor allele frequencies for a given SNP. Furthermore, different populations may have different correlations for nearby SNPs. For these reasons, Principal Component Analysis (PCA), a dimension-reduction technique, is a popular inclusion in GWAS and other approaches in statistical genetics (Abraham and Inouye, 2014; Price et al., 2006). Principal components are typically included in a design matrix as covariates alongside the locus allele count.

The design matrix for a GWAS at a given locus is typically composed of rows or columns corresponding to individual observations, and the other dimension fields corresponding to an intercept, covariates (some of which may be principal components), and the genotype data. The genotype for an individual at a particular SNP can be recorded as a ternary outcome to represent the number of a copies of a referent allele, i.e. 0, 1, or 2. For a phased haplotype, the corresponding entry would be a 0 or 1, and the sum of the two haplotype entries yields the genotype entry. Haplotype "phasing" refers to the process of allocating individual genotypes to haplotypes, research of which has attracted much attention in bioinformatics (Browning and Browning, 2011). The response (typically univariate), contains the phenotype data. Common data types include binary observations (e.g. disease status of 1 or 0) and quantitative/continuous (e.g. height or weight).

For GWAS, the null hypothesis is generally that a given SNP has an effect size of 0. Because GWAS proceeds SNP by SNP, this can readily lead to issues of multiple-testing, if left unaddressed. There are a variety of approaches to account for multiple-testing. An ex-

ample of which is the Bonferroni correction, an approach that is one of the simplest and most conservative (Bush and Moore, 2012).

Sometimes researchers may be interested in pleiotropy, where a SNP can potentially be associated with more than one phenotype. PheWAS is a popular method for performing a "reverse GWAS," scanning multiple phenotypes for associations with a particular SNP (Denny et al., 2010). PheWAS has been used to suggest associations between *HLA* variants with different outcomes—for instance, an association between the *HLA-B* genotype and spondylopathies, uveitis, variability in platelet count, and also other conditions, e.g. mastoiditis (Hebbring, 2014).

Some methods in recent years have attempted to model multiple phenotypes simultaneously in a multivariate GWAS (Galesloot et al., 2014). Instead of testing one phenotype against a SNP at a time, multiple phenotypes are simultaneously tested. Increasing dimensionality in the phenotype naturally becomes increasingly difficult, computationally and statistically.

1.2.2 Association Studies in Recent Years

Although increasingly common since 2007, GWAS have also engendered controversy in terms of their value in genetics (Visscher et al., 2012; Yang et al., 2015; Bush and Moore, 2012). Historical criticism of GWAS can be found aplenty, particularly with respect to "missing heritability" and multiple-testing (Visscher et al., 2012). The "missing heritability" refers to the then-purported lack of ability in GWAS to recover the implied heritability claimed by approaches such as twin, adoption, and/or other familial studies. These issues have, in turn, invited skepticism over whether such studies over-stated their claims about heritability, and/or whether GWAS is fundamentally flawed as a methodology.

However, it has been argued that increasing sample sizes in modern-day data-sets alone recover much of the "missing heritability" in complex traits (Visscher et al., 2012; Yang et al., 2015), especially through polygenic prediction methods. Although computationally intensive—and perhaps not as immediately targeted toward candidate loci identification—these methods have largely corroborated the heritability ascribed by twin, adoption, and/or other familial

studies. Typically, polygenic prediction methods are machine-learning and/or Bayesian in nature, as the number of possible predictors tends to be (vastly) greater than the number of observations in modern-day genomic data-sets (Vattikuti et al., 2014; Moser et al., 2015).

For GWAS, when effect sizes for individual loci are small, as is the case for human phenotypes such as cognitive ability, height, and diabetes, larger sample sizes are needed for identifying loci candidates. The relationship between the number of GWAS hits found and the sample size is not always linear, for example, as seen in human height (Visscher et al., 2012). Increases in sample size can lead to relatively sudden increases in the number of hits.

In retrospect, GWAS, as an approach, has been useful in highlighting issues in statistical genetics, such as statistical power, multiple-testing, and replicability. The "robust and overwhelming" empirical results from GWAS have largely validated it as a methodology, and GWAS has since grown in acceptance. To this day, GWAS continues to play a central role in identifying candidate loci, characterizing genetic architecture, uncovering pleiotropy, and developing polygenic scores. Each year sees an increasing number of candidate SNPs identified by GWAS (Visscher et al., 2017).

1.3 Summary

We present some likelihood-based approaches for statistical genetics, particularly with respect to genotype-phenotype associations. Our approaches have, as target applications, modern-day large-scale genomic data-sets and research goals—seeing as modern data-sets should have a sufficient number of observations to leverage asymptotic results.

In our approaches we prioritize computational performance and feasibility, while aiming to retain desirable statistical properties. Although clearly motivated by modern-day genomics—the spirit and concepts behind these methods can be applicable elsewhere. Likelihood-based in nature, they are well-suited to common formulations of modeling, testing, and inference.

We lead-off with HAP-SAMPLE2, a successor to HAP-SAMPLE (Wright et al., 2007).

HAP-SAMPLE2 is a haplotype resampling technique, intended as a data-generation approach for researchers looking to utilize simulated data-sets. Attention is paid to scalability with respect to number of loci, population admixture, and type of study. Four simulation types are available, formed by the two-by-two combination of rare and common variants, with a case/control or quantitative phenotype design.

Next, we describe a Fast Estimation for the Multivariate Probit, accomplished via a Two-Stage Composite Likelihood. The multivariate probit is a popular method by which to model correlated binary responses. In many settings, including clinical data, phenotypes are recorded as binary traits. Despite its popularity, multivariate probit estimation can be cumbersome as the number of observations and/or predictors grows. Correlated binary responses have received attention in fields as varied as genomics, epidemiology, economics, and finance.

Our approach on the Multivariate Probit transitions to our subsequent presentation on the Simultaneous Modeling of Multivariate Heterogeneous Responses & Heteroskedasticity via a Two-Stage Composite Likelihood. It is an extension of the Two-Stage Composite Likelihood methodology as applied to the multivariate probit. More targeted toward genetics in its presentation, it incorporates continuous traits alongside binary ones, as well as the potential for different covariance structures for different locus allele counts. We have in mind a fast-scanning method to detect mean and variance associations, simultaneously testing multivariate phenotypes with both binary and continuous components.

2

Haplotype Resampling with Admixed Populations

2.1 Introduction

Given developments in statistical genetics, simulation methods have been receiving greater attention in complementing genotype-phenotype association studies (Wright et al., 2007; Su et al., 2011). Simulated data-sets can be of use in performing power analyses and reviewing different statistical learning methods (Yuan et al., 2012). We present HAP-SAMPLE2, expanding the methodologies previously found in HAP-SAMPLE (Wright et al., 2007). HAP-SAMPLE2 simulates data-sets for admixed populations by resampling from real data-sets, using either common or rare variants, with phenotypes for use in a case/control or a quantitative trait design.

Over the years there have been many software packages made available to perform data-set simulations. These packages can often be loosely characterized as deploying backward or forwards simulations, or using resampling approaches (Yuan et al., 2012; Wright et al., 2007). Backward or forward simulations refers to emulating an evolutionary process going forward or backward in time. Modeling evolutionary processes often entails a material degree of hyperparameter-specification and monitoring; resampling approaches, requiring less parameter-tuning and ad-hoc supervision, tend to be more computationally favorable (Yuan et al., 2012). By its nature, resampling also retains realistic linkage equilibrium patterns and accompanying allele frequencies (Wright et al., 2007; Su et al., 2011; Yuan et al., 2012). Some examples of bioinformatics software are briefly described below.

HAPSIMU (Zhang et al., 2008) is a C++ simulation platform for association studies with HapMap data, including recombination fractions from HapMap distances. It supports qualitative or qualitative data under both the continuous migration or discrete models, simulating both genotype and phenotype data. A central focus of HAPSIMU is that of population

structure and inter-population heterogeneity, which is deemed to be a deleterious factor in association studies. The user can specify two-way admixture proportions and other input values, such as genotype relative risk and disease prevalence.

HAPGEN2 (Su et al., 2011)—the successor to HAPGEN—a C++ package focused on case/control disease data, allows multiple disease loci on the same chromosome. It incorporates SNP-level recombination rates, and specification of heterozygote and homozygote relative risks. The outputs would be simulated haplotype and genotype data for cases and controls. The motivating considerations behind HAPGEN2 include better leveraging of pre-existing data, avoiding the computational burdens involved with forward and/or backward simulations, and providing simulated data that realistically characterizes extant linkage disequilibrium in populations. HAPGEN2 is considered a resampling approach, using data from HapMap3 and 1000 Genomes to provide fast simulations.

Another simulation software package is simuPOP (Peng and Kimmel, 2005), alongside simuGWAS. It is a Python script that uses HapMap data (including fine-scale recombination rates), for evolving a population forward or backward. It can also handle "GxE" models through user-specified markers. Target final allele frequencies are also specified, with fitness information expressed in terms of fitness, additive interactive or none under a logistic regression model. Key considerations for simuPOP are flexibility and extensibility of simulations. No ex-ante limits are placed upon factors such as genome size, ploidy number, or mating type.

SimuPOP is provided in the form of a suite of Python libraries, leaving it "the user's responsibility to write a Python script to glue these pieces together and form a simulation" (Peng and Amos, 2010). simuGWAS leverages the functions provided by simuPOP to provide simulations and post-processing steps targeted toward aiding association studies. An example of a post-processing step would be a rejection-sampling algorithm for case/control samples and/or parent-child trios in situations where disease prevalence is low.

GWAsimulator (Li and Li, 2008) is a C++ program, with Linux, Windows, and Mac executables, that uses a "rapid window algorithm" to simulate case/control or population samples, with Hardy-Weinberg Equilibrium assumed. The moving window refers to sampling contigu-

ous SNPs in a Markovian fashion to mimic recombination. It can use HapMap data and can receive user-specified inputs for a locus-specific, multi-locus disease model (at most one per chromosome) with support for two-way interactions. The user also provides disease prevalence and genotype relative risk. GWAsimulator assumes disease prevalence is a function of genotypes, and ultimately converts the inputs into logistic equation form.

SEQSIMLA/SEQSIMLA2 (Ren-Hua et al., 2014) is a download-able C++ program, with an online version also available. A top priority is shared environment and correlated traits (such as blood pressure, BMI, insulin, and glucose levels) under large pedigree structures—it simulates sequence data under both prevalence and population-attributable risk models, for family or unrelated individuals case/control or quantitative data. Computation time is also a material consideration. Currently—due to computation concerns—two correlated traits at a time are supported, albeit the model could hypothetically be extended to three or more traits. Pairwise interactions can be specified via odds-ratios or disease prevalence. The underlying model is that of logistic penetrance, where the log-odds of the probability of being affected by a disease is a function of an intercept, the additive effects of genotype, and the interactive/epistatic effects of genotype.

simuRare (Xu et al., 2013) is an R package for regression-based resampling with 1000 Genomes data, focused on rare variants. The idea is to use dataset A with both rare and common variants to impute on dataset B to form dataset C, then resample from C with recombination. The probability of a rare allele being chosen is a function of nearby common variants with respect to the rare allele. The disease model is based on the logistic regression, inputted via disease prevalence and odds ratios. Retaining realistic LD and minor allele frequency are high priorities. For these purposes, the authors found the performance of simuRare favorable to that of GWAsimulator, simuGWAS, and HAPGEN2.

Under a threshold model for binary traits, GPOPSIM (Zhang et al., 2015) deploys forward simulation of multiple correlated genetic traits, based on the mutation-drift equilibrium model. Continuous traits are supported as well. It can incorporate various assumptions such as different population sizes, male/female breeding ratios (or differentials), selection rules, genetic drift, etc. Chromosomal structure is “arbitrarily assigned” and the locations of markers

and QTLs are determined via the assumption of a mixture of uniform and exponential distributions, to emulate real SNP data in a realistic manner. GPOPSIM is implemented in Fortran 90, tested on Windows and Linux.

GCTA-Simu (Yang et al., 2011) of popular software GCTA simulates a GWAS based on actual genotype data for an arbitrary number of causal variants for a given number of individuals. It can do so under an additive quantitative model, or case/control liability threshold model. Arguments can include assumptions such as effect size for loci, heritability, and disease prevalence. Causal SNPs can be inputted as a list via a file, or effects can be generated from the standard normal distribution.

Some of these features are summarized in Table 2.1, including the new features incorporated into the software HAP-SAMPLE2 described here. Features include logistic regression and absolute genotype user specification, recombination, and admixture functionality.

Table 2.1: Overview of simulation software by software name and estimated year made available. "+" denotes presence of functionality; "-" denotes absence or unknown. Functionality shown in the columns are case/control traits, quantitative traits, resampling, recombination, and admixture.

Name	Year	C/C	Quant.	Resamp.	Recomb.	Adm.
HAP-SAMPLE2	2020	+	+	+	+	+
HAP-SAMPLE	2007	+	-	+	+	-
HAPSIMU	2008	+	+	-	+	+
HAPGEN2	2011	+	-	+	+	-
simuGWAS	2011	+	-	-	+	+
GWAsimulator	2008	+	-	+	+	-
SEQSIMLA2	2013	+	+	-	+	-
simuRare	2013	+	-	+	+	-
GPOPSIM	2015	+	+	-	+	-
GCTA-Simu	2014	+	+	+	-	-

2.1.1 HAP-SAMPLE2 and Admixture

HAP-SAMPLE (Wright et al., 2007), a program with a web-based interface, performed resampling of real autosomal chromosome data to yield simulated data sets for case/control studies. The simulated population that resulted thus had realistic linkage disequilibrium patterns and allele frequencies. Multiple disease loci were permitted, with at most one per

chromosome. HAP-SAMPLE also included a crossover process, to provide a potential source of novelty (Wright et al., 2007).

In HAP-SAMPLE, simulated genotypes were conditioned upon the specified phenotype for a given individual, whether it be case or control. For case/control designs, specifying the number of case and control samples first and simulating the genotypes upon the phenotypes can be more efficient than the converse, i.e. simulating the genotype first and then constructing the phenotype post-hoc. An example would be a rare disease, where simulating the genotypes first may necessitate a large number of draws to attain a given level of cases, and many of the simulated controls may end-up unused in downstream analyses.

However, the random sampling involved in HAP-SAMPLE was not designed to incorporate population substructure (Wright et al., 2007). We present HAP-SAMPLE2, an updated version of HAP-SAMPLE, that incorporates the admixture simulation process described by Zhou et al. (2017) for generating simulated individuals with targeted admixture proportions of differing ancestral populations. HAP-SAMPLE2 thus performs haplotype resampling to simulate individuals with user-specified ancestral population proportions. Each subset of a simulated crossover, where the boundary regions of each subset are stochastically-determined by the artificial recombination process, can be chosen from different source haplotype populations based upon the target admixture proportions (Zhou et al., 2017). An example is shown in Figure 2.1.

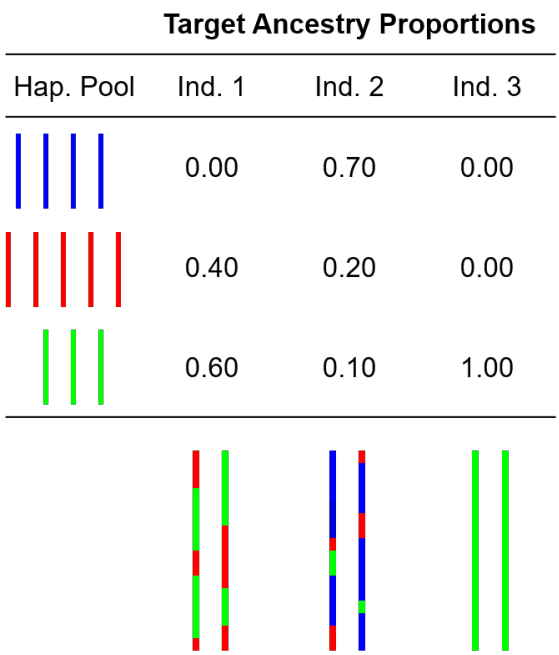

Figure 2.1: Example chromosomes for three individuals with admixture from three potential source populations. Individual 1 has two-way admixture between two source populations, Individual 2 has admixture from all three source populations, and Individual 3 has ancestry from one source population.

To better align with common forms of user-specifications, we transition to the logistic regression odds-ratio specification, where the user inputs what are essentially coefficients to a logistic regression. This is ultimately converted into the absolute genotype specification, i.e. the genotype probabilities conditioned on the trait.

2.1.2 HAP-SAMPLE2 and Rare Variants

It has been long noted that many phenotypes—including complex diseases—tend to be associated with common, ancient variants (Altshuler et al., 2005; Lohmueller et al., 2003; Peng and Kimmel, 2007; Wright et al., 2007). Variants with low penetrance and/or are only associated with age-related diseases can persist given the lack of selection pressure, and thus can be observed at material levels across different continental populations (Altshuler et al., 2005; Gibbs et al., 2003; Wright et al., 2007). However, with whole-genome sequencing data becoming increasingly available, researchers have been considering rare variants.

Rare variants pose some distinct challenges vis a vis common variants (Auer and Lettre,

2015). They are more numerous than common variants, and are less mutually correlated, which renders any LD-pruning more difficult and exacerbates issues with multiple-testing. Furthermore, by definition, their minor allele frequencies are lower, which tends to decrease statistical power.

Burden-testing is a common approach that aggregates minor allele counts across loci (Lee et al., 2014). Aggregating minor allele frequencies helps alleviate headwinds caused by low minor allele frequencies. By construction, burden-testing assumes that the effects of rare alleles are in the same direction (e.g. typically increasing the odds of a disease). According to evolutionary theory, rare variants should generally have deleterious effects on fitness; thus, ignoring directionality may not be as concerning as may seem on first glance.

Although the primary focus for HAP-SAMPLE2 remains on common variants, we also introduce burden-testing functionality for rare variants, under both a case/control and quantitative trait design. This largely leverages the approach for common variants in spirit, as well as much of the infrastructure.

2.1.3 The Four Simulation Types in HAP-SAMPLE2

In addition to binary phenotypes for case/control designs, as was done in HAP-SAMPLE, HAP-SAMPLE2 can also simulate quantitative phenotypes. Thus, HAP-SAMPLE2 offers four simulation types, for the two-by-two combination of common and rare variants, and case/control and quantitative designs. This is an expansion of HAP-SAMPLE, which provided one. The four simulation types and the expansion of capabilities since HAP-SAMPLE, is summarized in Figure 2.2.

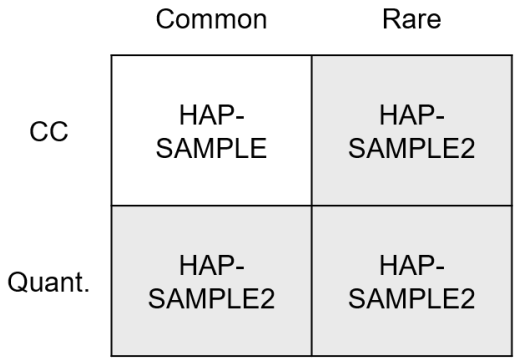

Figure 2.2: Comparison of simulation types provided by HAP-SAMPLE and HAP-SAMPLE2, by response type (case/control versus quantitative) and variant type (common versus rare).

2.2 Methods

Each of the four simulation types has their idiosyncrasies, but they share many core aspects in methodology and infrastructure. An overview of the general process is depicted in Figure 2.3. HAP-SAMPLE2 receives as inputs user-specified settings and target admixture proportions, along with two data sources—phased haplotype data and a genetic map. The phased haplotype data serves as the source from which to resample in constructing simulated chromosomes. The genetic map provides genetic distances, which after interpolation and extrapolation yields recombination probabilities at the inter-loci level. The output from HAP-SAMPLE2 is simulated genotype/phenotype data. For a quantitative trait design, the phenotype in the output is explicitly simulated. For a case/control design, the binary phenotype is implied, as the data-sets for cases and controls are separate objects.

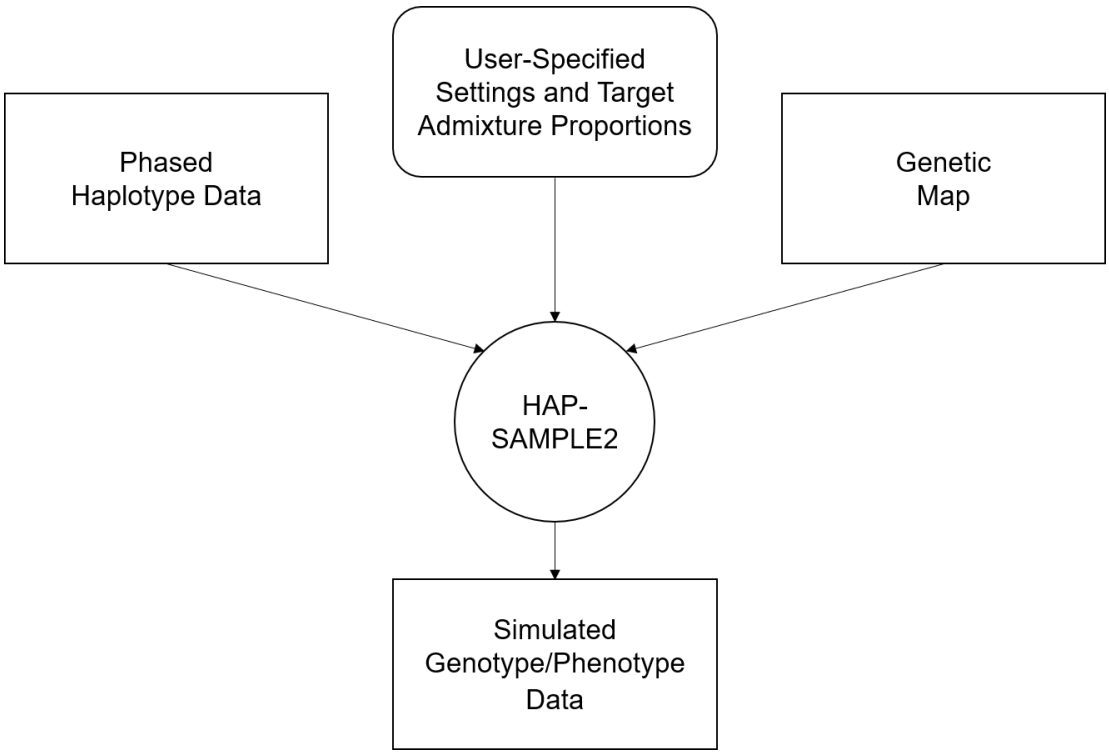

Figure 2.3: Process schematic for HAP-SAMPLE2. The user provides settings for the simulations, as well as target admixture proportions for the simulated individuals. Phased haplotype data serves as the source data for resampling, and the genetic map is the main input in determining recombination probabilities. The output is simulated genotype/phenotype data.

2.2.1 Phased Haplotype Data

In developing HAP-SAMPLE2, we used data downloaded from the 1000 Genomes Project (Auton, A. and Abecasis, G. and Altshuler, D. et al., 2015). However, the methodology could theoretically be applied using different input data. 1000 Genomes provides Variant Call Format (VCF) files for each chromosome that contain phased autosomal haplotype data for 2,504 individuals. There is also an associated information file that labels each of these 2,504 individuals with a population code, which is what we used to determine ancestral backgrounds, for 26 populations in total (Auton, A. and Abecasis, G. and Altshuler, D. et al., 2015).

In our approach for HAP-SAMPLE2 we consider biallelic loci, i.e. loci where a 0 or 1 can be recorded to denote presence of the reference allele or alternative allele, respectively. We can then apply the Hardy-Weinberg assumptions throughout, in moving between allele frequencies

and genotype frequencies. This also allows us to store source and simulated haplotypes in sparse vector representation, enabling faster run-times and entailing less RAM usage.

Prior to haplotype construction, an optional pre-processing step can be taken in HAP-SAMPLE2 to manage the number of loci considered. This step includes filtering for minor allele frequencies above a user-specified threshold, retaining only every n'th locus, and/or retaining particular loci by reference SNP identification (rs ID).

2.2.2 Genetic Map for Recombination Probabilities

Starting with a genetic map containing locus-specific base positions and corresponding relative distances in centimorgans (cM), one can estimate relative distances for loci in the 1000 Genomes data via interpolation and/or extrapolation. Such maps can provide guidance as to recombination hotspots, where recombination is more likely to occur (Khil and Camerini-Otero, 2010).

We used genetic maps downloaded from the IMPUTE/IMPUTE2 website (Howie and Marchini, 2015). For 1000 Genomes loci that lay within the range of the downloaded map, linear interpolation was performed using contiguous points from the map. For loci that lay outside the range, extrapolation was done using a simple linear regression informed by all of the points in the downloaded map, performed for each chromosome.

Once this interpolation/extrapolation is done to create a genetic map for the 1000 Genomes loci for each desired chromosome, it is straightforward to then obtain locus-specific recombination probabilities by taking the difference between relative distances and converting the centimorgans to probabilities (1 cM = .01 expected crossovers).

2.2.3 Case/Control Design for Common Variants

We will begin our description of the methods with the case/control design for common variants, which was the simulation type provided in HAP-SAMPLE. Since all four simulation types share core aspects of the methodology and infrastructure, we will walk-through

the case/control design for common variants, and then describe the others by comparing and contrasting.

Typically in case/control designs, individuals recorded as cases are bearers of some condition—e.g. a disease—and controls the converse. Let π_i be the probability of disease for a given individual i. Then, in a logistic regression model we can write:

$$\ln(\frac{\pi_i}{1-\pi_i}) = \boldsymbol{a_i^T \alpha_s} + g_{is}\beta_s$$

Where $\boldsymbol{a_i}$ is a vector (that sums to 1) containing values of the individual's admixture proportions for each source population, $\boldsymbol{\alpha_s}$ is a vector of coefficients corresponding to each source population for locus s, g_{is} is the individual's genotype for locus s ($g_{is} \in \{0, 1, 2\}$), and β_s is an element of $\boldsymbol{\beta}$, a vector of effect sizes for each locus. Equivalently, the above can also be stated as:

$$\pi_i = \frac{exp(\boldsymbol{a_i^T \alpha_s} + g_{is}\beta_s)}{1 + exp(\boldsymbol{a_i^T \alpha_s} + g_{is}\beta_s)} = \frac{exp(z_i)}{1 + exp(z_i)}$$

With z_i representing the natural logarithm of the odds, or the "log-odds."

In HAP-SAMPLE2, the user specifies case prevalence (commonly a disease prevalence) for each source population, causal/effective SNPs (SNPs associated with trait values, maximum of one per chromosome), logistic regression coefficients corresponding to slopes against the allele values for those SNPs ($\boldsymbol{\beta}$), and desired admixture proportions for each simulated individual i ($\boldsymbol{a_i}$). HAP-SAMPLE2 will then calculate values for the ancestry vector for each SNP ($\boldsymbol{\alpha_s}$), and genotype values for each SNP (g_{is}) will be drawn for each simulated individual.

2.2.3.1 Ancestry Coefficients

The ancestry coefficients in the logistic regression model can be solved for, given disease prevalence, the alternative allele frequency, and the coefficient against genotype. Alternative allele frequencies for each ancestral population can be calculated from the source haplotype pool. Genotype frequencies for the source populations can be obtained from the allele frequencies via the Hardy-Weinberg assumptions (Wright et al., 2007). Let p_k be the user-specified

disease prevalence for population k and let f_{ks} be the alternative allele frequency for SNP s for population k, while $P_k(g_{is})$ is the population k's genotype frequency for genotype g_{is}, then :

$$P_k(g_{is} = 0) = (1 - f_{ks})^2$$

$$P_k(g_{is} = 2) = f_{ks}^2$$

$$P_k(g_{is} = 1) = 1 - P_k(g_{is} = 0) - P_k(g_{is} = 2)$$

The probability of a given individual from this population with genotype g_{is} being a case is:

$$\frac{exp(\alpha_{ks} + g_{is}\beta_s)}{1 + exp(\alpha_{ks} + g_{is}\beta_s)}$$

Where α_{ks} corresponds to an intercept. Thus, we can write:

$$p_k = \frac{1}{N_k}\sum_{1}^{N_k}\pi_i = \frac{1}{N_k}\sum_{1}^{N_k}\frac{exp(z_i)}{1 + exp(z_i)} = \frac{1}{N_k}\sum_{1}^{N_k}\frac{exp(\alpha_{ks} + g_{is}\beta_s)}{1 + exp(\alpha_{ks} + g_{is}\beta_s)}$$

This leaves α_{ks} as the only unknown. A value for each α_{ks} can then attained via numeric optimization, e.g. via R's `optimize`. Repeating this for each α_{ks} allows us to estimate $\boldsymbol{\alpha_s}$, each of which has length K for K total populations.

2.2.3.2 Absolute Genotype, Cases

Recall that, according to the logistic regression model, we have as an individual's probability of disease: $\ln(\frac{\pi_i}{1-\pi_i}) = \boldsymbol{a_i^T\alpha_s} + g_{is}\beta_s$. That is, each individual has a probability of disease given ancestry and genotype. However, in HAP-SAMPLE2, we wish to have an Absolute Genotype specification—the probability of genotype given disease and ancestry—as simulated individuals have pre-specified disease statuses and user-provided ancestry proportions.

Now with the ancestry vector solutions in hand for each combination of ancestral population and SNP, we can proceed to converting the inputs into the Absolute Genotype specification via typical conditional probability calculations. Since we assume loci are independent

across chromosomes and limit each chromosome to one causal SNP, this formulation allows us to efficiently draw genotypes, i.e. determine whether an individual has a 0, 1, or 2 for each SNP.

Using each individual's probability of being a case given genotype and ancestry, $P(D_i = 1|g_{is}, \boldsymbol{a_i})$, we can derive the probability of genotype given disease and ancestry, $P(g_{is}|D_i = 1, \boldsymbol{a_i})$. Following the logistic regression model:

$$\pi_i = P(D_i = 1|g_{is}, \boldsymbol{a_i})$$

$$\ln(\frac{P(D_i = 1|g_{is}, \boldsymbol{a_i})}{1 - P(D_i = 1|g_{is}, \boldsymbol{a_i})}) = \boldsymbol{a_i}^T \boldsymbol{\alpha_s} + g_{is}\beta_s = z_i$$

$$\frac{P(D_i = 1|g_{is}, \boldsymbol{a_i})}{1 - P(D_i = 1|g_{is}, \boldsymbol{a_i})} = exp(z_i)$$

$$P(D_i = 1|g_{is}, \boldsymbol{a_i}) = \frac{exp(z_i)}{1 + exp(z_i)} = \frac{P(D_i = 1, g_{is}, \boldsymbol{a_i})}{P(g_{is}|\boldsymbol{a_i})P(\boldsymbol{a_i})}$$

$$= \frac{P(g_{is}|D_i = 1, \boldsymbol{a_i})P(D_i = 1|\boldsymbol{a_i})P(\boldsymbol{a_i})}{P(g_{is}|\boldsymbol{a_i})P(\boldsymbol{a_i})} = \frac{P(g_{is}|D_i = 1, \boldsymbol{a_i})P(D_i = 1|\boldsymbol{a_i})}{P(g_{is}|\boldsymbol{a_i})}$$

$$P(g_{is}|D_i = 1, \boldsymbol{a_i}) = \frac{P(D_i = 1|g_{is}, \boldsymbol{a_i})P(g_{is}|\boldsymbol{a_i})}{P(D_i = 1|\boldsymbol{a_i})} = \frac{exp(z_i)P(g_{is}|\boldsymbol{a_i})}{(1 + exp(z_i))P(D_i = 1|\boldsymbol{a_i})}$$

Note that $P(D_i = 1|\boldsymbol{a_i})$ can be estimated from the user input for disease prevalence and a given individual's ancestry proportions, and $P(g_{is}|\boldsymbol{a_i})$ can be calculated from the data.

2.2.3.3 Absolute Genotype, Controls

Absolute genotype values for the controls can be obtained via a similar formulation as that of cases. Recalling the last equation previously from cases and substituting in $D_i = 0$ for $D_i = 1$:

$$P(D_i = 0|g_{is}, \boldsymbol{a_i}) = \frac{P(D_i = 0|g_i s, \boldsymbol{a_i})P(g_{is}|\boldsymbol{a_i})}{P(D_i = 0|\boldsymbol{a_i})} =$$

$$\frac{(1 - P(D_i = 1|g_{is}, \boldsymbol{a_i}))P(g_{is}|\boldsymbol{a_i})}{1 - P(D_i = 1|\boldsymbol{a_i})} = \left(1 - \frac{e^{z_i}}{1 + e^{z_i}}\right)\left(\frac{P(g_{is}|\boldsymbol{a_i})}{1 - P(D_i = 1|\boldsymbol{a_i})}\right)$$

2.2.3.4 Haplotype Construction

A Poisson approximation can be used to stochastically determine the number of crossovers for each haplotype for each individual. Let L_h represent the number of loci on a given chromosome h, and θ_{hl} represent the recombination probability between locus l and $l+1$ on that chromosome. Then for each individual's haplotype of chromosome h, the number of assigned crossovers for that chromosome, r_h, could be determined via a random draw from the Poisson distribution with Poisson parameter $\lambda_h = \sum_{l=1}^{L_h - 1} \theta_{hl}$.

Furthermore, we introduce a floor of 1 such that each simulated chromosome experiences at least one crossover, and adjust the Poisson parameter accordingly to leave the mean number of crossovers unchanged. Although this requirement does not have a large impact on any downstream simulation results, this more closely matches the expectation that there occurs at least one crossover per human chromosome (Wilson and Hunt, 2002). For each λ_h, we solve for a λ_h^* such that $E[r_h] = 1 + E[r_h^*]$, where each r_h^* is also Poisson distributed with Poisson parameter λ^*. The number of crossovers for a given individual on a given chromosome, r_h^*, can then be determined via a random draw from this adjusted distribution.

Suppose a given haplotype for a particular individual was assigned to have r_h^* crossovers on a given chromosome. r_h^* sites can be drawn from the $L_h - 1$ possibilities to serve as the crossover locations, with weighted probabilities informed by the aforementioned locus-specific probabilities. The simulated haplotype will then have $r_h^* + 1$ subsets, where the boundaries of the subsets are either crossover sites or one of the two ends of the chromosome. $r_h^* + 1$ haplotypes are drawn from the source pool to construct the haplotype, where the probability of each of the $r_h^* + 1$ source haplotypes coming from a given source population k is specified by the individual's target admixture proportion for population k.

For disease chromosome, the subset out of the $r_h^* + 1$ subsets that contains the disease locus for a given constructed haplotype will be limited to sampling from haplotypes of a given source population based on the simulated individual's assigned genotype. If the individual was assigned to be homozygous for the alternative allele, only haplotypes of that source population containing the alternative allele are eligible to be sampled for each of the individual's two sim-

ulated chromosomes, and likewise for the reference allele. If the individual was assigned to be heterozygous, only haplotypes containing the alternative allele are eligible for one chromosome, and only haplotypes containing the reference allele for the other chromosome.

Note that this requires each ancestral population in the source chromosome pool to have non-zero alternative (reference) frequencies for all causal SNPs in the case/control design. For subsets of the disease chromosome that do not contain the disease locus, or for non-disease chromosomes, all source haplotypes of the assigned source population for a given subset are eligible to be sampled. This process is illustrated in Figure 2.4.

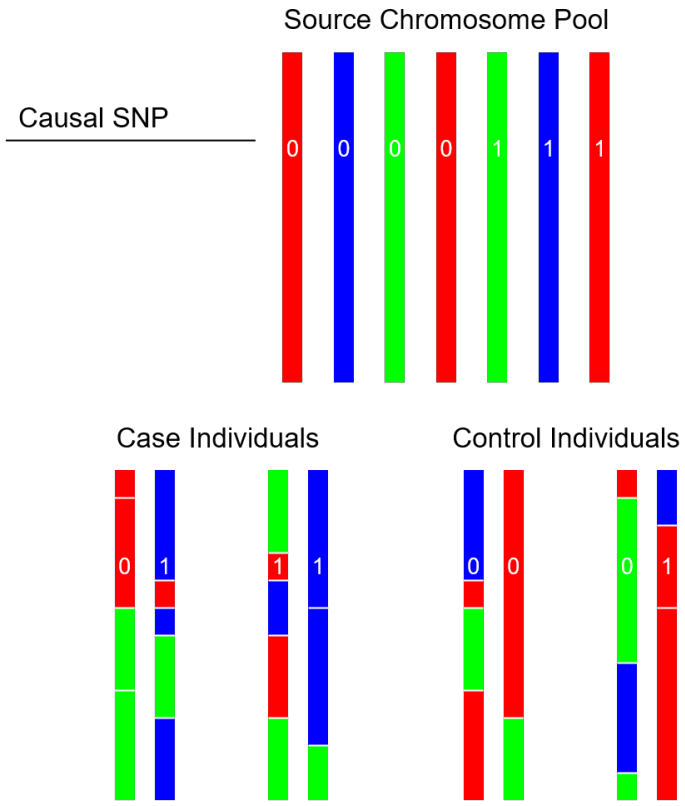

Figure 2.4: Stylized illustration of simulating chromosomes for cases and controls with common variants, using three ancestral populations. A positive effect size is assumed for the alternative allele, thus the alternative allele is found with higher frequency in the two case individuals than in the two control individuals. White borders in the simulated chromosomes represent locations where crossovers occurred.

2.2.4 Quantitative Phenotype for Common Variants

Unlike for the case/control design, we are not conditioning upon the quantitative phenotype in drawing genotypes. Suppose an individual i has a phenotype value of z_i for a given trait.

Assuming normality of the trait and equal variance across sub-populations, we can write:

$$z_i \sim \phi(\boldsymbol{a_i^T}\boldsymbol{\alpha} + \boldsymbol{g_i^T}\boldsymbol{\beta}, \sigma^2)$$

Where ϕ is the normal probability density function, $\boldsymbol{a_i}$ and $\boldsymbol{g_i}$ are the individual's ancestry proportions and genotype, respectively, $\boldsymbol{\alpha}$ contains ancestry-specific intercepts, $\boldsymbol{\beta}$ contains the user-specified loci effect sizes, and σ^2 the user-specified variance. In contrast to the case/control design, multiple loci per chromosome can be specified as causal SNPs for quantitative phenotypes. Analogous to disease prevalences for the case/control design, given user-specified source population means $\boldsymbol{\mu}$ and source population-specific genotype frequencies, one can solve for each α_k in $\boldsymbol{\alpha}$.

2.2.4.1 Ancestry Coefficients

Similar to the case/control design for common variants, we will solve for ancestry coefficients. However, instead of vector of coefficients for each SNP s, we will now just have one α_k for each ancestral population k. Suppose we have S SNPs and thus $3^S = M$ combinations of genotypes. Let $\boldsymbol{g_m}, m \in \{1, ..., M\}$, be a particular combination of genotypes, such that $\boldsymbol{g_1} = \{0, ..., 0\}$ and $\boldsymbol{g_M} = \{2, ..., 2\}$.

If μ_k is the user-specified mean trait value for population k, f_{ks} is the alternative allele frequency for SNP s for population k, $P_k(\boldsymbol{g_{im}})$ is the genotype frequency for genotype $\boldsymbol{g_m}$ among population k, and $\boldsymbol{\beta}$ is a vector of slopes against genotype g_{is}, corresponding to genotype $\{0, 1, 2\}$, then:

$$P_k(g_{is} = 0) = (1 - f_{ks})^2$$

$$P_k(g_{is} = 2) = f_{ks}^2$$

$$P_k(g_{is} = 1) = 1 - P_k(g_{is} = 0) - P_k(g_{is} = 2)$$

$$P_k(\boldsymbol{g_{im}}) = \prod_{s=1}^{S} P_k(g_{is})$$

$$z_{km} = \alpha_k + \boldsymbol{g_m}\boldsymbol{\beta}$$

$$\mu_k = \sum_{m=1}^{3^S} z_{km} P_k(\boldsymbol{g_{im}})$$

Which leaves α_k as the only unknown. Optimizing for each α_k provides our estimate for the vector of ancestry coefficients $\boldsymbol{\alpha}$.

2.2.4.2 Haplotype Construction

For a quantitative trait model, the simulated haplotypes are constructed similarly to those under the case/control design. However, there is no longer a need for consideration of causal SNPs nor special handling of particular subsets, since the quantitative phenotype values for each individual are determined after the genotypes are constructed. This is illustrated in Figure 2.5.

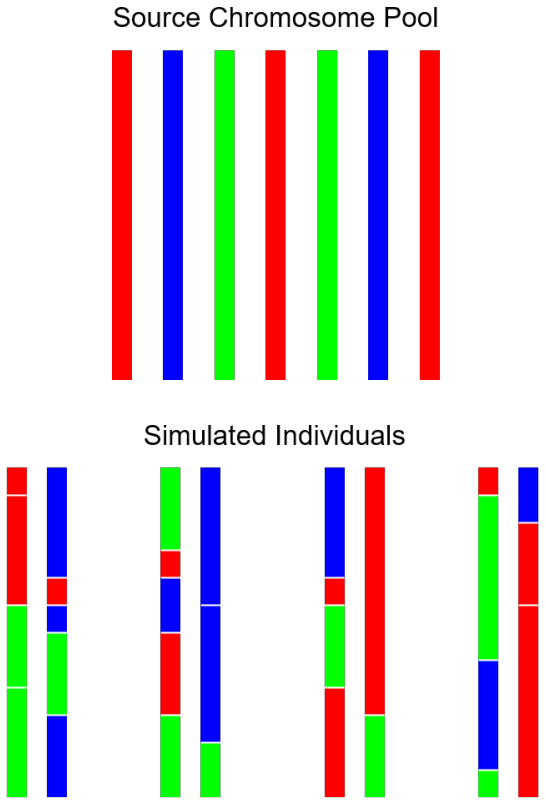

Figure 2.5: Stylized illustration of simulating chromosomes for a quantitative phenotype with common variants, using three ancestral populations. White borders in the simulated chromosomes represent locations where crossovers occurred.

2.2.5 Case/Control Design for Rare Variants

For a rare variants case/control design, we have the user specify ancestral population-specific case prevalences, akin to the common variants set-up. The user would also specify the chromosomes to use, mean effect size of a rare variant allele, the weighting scheme, and the minor allele threshold. Note in contrast to the functionality for common variants, that no longer is there a specification of effect sizes for particular loci.

2.2.5.1 Calculation of Rare Variant Weights and Minor Allele Frequencies

Currently, HAP-SAMPLE2 supports two variant weighting schemes: Fixed and Madsen-Browning. If the weighting scheme is fixed, any rare allele counts will have the same weight (the count, for a particular locus, being 0, 1, or 2 per individual), where rare is defined by the minor allele threshold:

$$w_s = \begin{cases} 1 & f_s < f_{thresh} \\ 0 & f_s \geq f_{thresh} \end{cases}$$

For Madsen-Browning, rare allele counts are weighted by $w_s = \frac{1}{[(f_s)/(1-f_s)]^{\frac{1}{2}}}$, f_s being the minor allele frequency for a particular locus s (Lee et al., 2014; Madsen and Browning, 2009). Whether the weighting scheme is fixed or Madsen-Browning, the weights are then adjusted by the sum of the weights and the number of loci as to set the mean adjusted weight equal to the user specification:

$$\beta_s = \beta_* S w_s \sum_{1}^{S} w_s$$

We calculate minor allele frequencies by taking the overall mean across the 26 population-specific minor allele frequencies. However, we acknowledge that choosing the reference population(s) upon which calculate the minor allele frequencies can be a noteworthy consideration, as allele frequencies can be quite different across different populations (Persyn et al., 2018). An allele that is rare in one population may not necessarily be rare in another

population, and in fact may be the major allele. Different populations can have greater or fewer rare variant loci than others; this could be related to differing mutational load across populations. Different populations may have experienced different effect population sizes in evolutionary history, with attendant implications for mutational load accumulation (Henn et al., 2015).

2.2.5.2 Scoring Individuals in the Source Pool for Rare Variant Burdens

Suppose we have weights for each locus s, calculated as above, whether they be weights calculated via the Fixed or Madsen-Browning weighting scheme. We can now score each individual in the source population for their rare variant burden. As per Lee et al. (2014), each individual i has score c_i:

$$c_i = \sum_{s=1}^{S} g_{is}\beta_s = \boldsymbol{g_i^T}\boldsymbol{\beta}$$

Where each g_{is} is the genotype value (0, 1, or 2) for individual i at locus s, and β_s is the adjusted weight.

2.2.5.3 Calculating Ancestry Coefficients for the Case/Control Design

Because each individual in the source pool is associated with a population tag, we can use the individuals of each population tag to calculate ancestry-specific coefficients. This is analogous to the handling of ancestry coefficients for common variants.

Given an ancestry-specific coefficient α_k for ancestral population k, for source individual i we can calculate the implied probability π_i of being a case through the logistic regression model:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_k + \boldsymbol{g_i^T}\boldsymbol{\beta} = \alpha_k + c_i$$

Or equivalently:

$$\pi_i = \frac{exp(\alpha_k + c_i)}{1 + exp(\alpha_k + c_i)}$$

Recall that the user specifies ancestry-specific prevalences. Suppose p_k is the user sup-

plied prevalence for population k. It is necessary that the mean implied probabilities of being a case for individuals of population k to equal the specified ancestry-specific prevalence p_k. Suppose N_k is the number of individuals belonging to population k. We can then set:

$$p_k = \frac{1}{N_k}\sum_{i=1}^{N_k}\pi_i = \frac{1}{N_k}\sum_{i=1}^{N_k}\frac{exp(\alpha_k + c_i)}{1 + exp(\alpha_k + c_i)}$$

This leaves α_k as the only free parameter, and thus we can solve for a value of α_k to satisfy the above equivalency. As discussed previously, this type of uni-dimensional optimization can be well-performed by R's `optimize`.

2.2.5.4 Calculating Probability of a Source Haplotype Being Selected for Constructing a Simulated Haplotype, Given Case/Control Status

Now all the inputs are available for calculating each source individual's implied probability of being a case, π_i. As would be expected, a given individual's probability of being a control would be $1 - \pi_i$. We can then calculate a given individual's probability share among that ancestral population, for both cases and controls:

$$\psi_{i,case} = \frac{\pi_i}{\sum_{i=1}^{N_k}\pi_i}$$

$$\psi_{i,cont} = \frac{(1-\pi_i)}{\sum_{i=1}^{N_k}(1-\pi_i)}$$

As was the case for common variants, in constructing a simulated haplotype, the ancestry for a given region in a simulated haplotype is first determined (stochastically selected via the specified target admixture proportions) before a particular source haplotype is selected for construction. Suppose a given population is chosen for a given region. Then, for source individuals, their probability of being chosen to construct that region is their probability share, whether it be for cases or controls.

One chromosome for a given individual may have a heavier burden than the other. Let us arbitrarily designate one chromosome as chromosome 1, and the other 2. So for chromosome

1 for a given individual, its probability of being selected when the mean rare variant burden effect size is positive is:

$$\eta_{i1,case} = \psi_{i,case}\left(\frac{c_{i1}}{c_{i1}+c_{i2}}\right)$$

$$\eta_{i1,cont} = \psi_{i,cont}\left(1-\frac{c_{i1}}{c_{i1}+c_{i2}}\right)$$

When the mean effect size is negative:

$$\eta_{i1,case} = \psi_{i,case}\left(1-\frac{c_{i1}}{c_{i1}+c_{i2}}\right)$$

$$\eta_{i1,cont} = \psi_{i,cont}\left(\frac{c_{i1}}{c_{i1}+c_{i2}}\right)$$

Likewise for chromosome 2, when the mean rare variant burden effect size is positive is:

$$\eta_{i2,case} = \psi_{i,case}\left(\frac{c_{i2}}{c_{i1}+c_{i2}}\right)$$

$$\eta_{i2,cont} = \psi_{i,cont}\left(1-\frac{c_{i2}}{c_{i1}+c_{i2}}\right)$$

When it is negative:

$$\eta_{i2,case} = \psi_{i,case}\left(1-\frac{c_{i2}}{c_{i1}+c_{i2}}\right)$$

$$\eta_{i2,cont} = \psi_{i,cont}\left(\frac{c_{i2}}{c_{i1}+c_{i2}}\right)$$

2.2.6 Quantitative Phenotype for Rare Variants

Analogous to the common variant approaches for case/control and quantitative phenotype designs, the rare variant process for quantitative phenotypes is, in many ways, simpler than the case/control design.

We can calculate minor allele frequencies and solve for ancestry coefficients in the same way as the for case/control individuals for rare variants. Akin to the methodology for the quantitative phenotype with common variants, we can perform haplotype construction first, and then determine the simulated individuals' phenotypes.

Suppose we have rare variant scores of c_i for each individual and intercept adjustments corresponding to each ancestral sub-population $\boldsymbol{\alpha}$. Similar to the quantitative phenotypes using common variants, these individuals' phenotypes can be determined by draws from the normal distribution:

$$z_i \sim \phi(\boldsymbol{a_i^T}\boldsymbol{\alpha} + c_i, \sigma^2)$$

The ancestry coefficients, $\boldsymbol{\alpha}$ can be solved, using the user-specified values for the targeted ancestral population means, $\boldsymbol{\mu}$.

$$\mu_k = \frac{1}{N_k}\sum_{i=1}^{N_k}\alpha_k + c_i$$

Thus, using each individual's rare variant score, c_i, we can assign phenotype scores after haplotype construction:

$$z_i = \boldsymbol{a_i^T}\boldsymbol{\alpha} + c_i + \epsilon_i$$

Where each ϵ_i is an independent draw from the normal distribution with a mean of 0 and a variance of σ^2. σ is the user-specified standard deviation.

2.3 Example Simulations

Example simulations were performed using data from 1000 Genomes (Auton, A. and Abecasis, G. and Altshuler, D. et al., 2015; IGSR: The International Genome Sample Resource, 2015) to illustrate the functionalities of HAP-SAMPLE2, for each of the four simulations. Similar to the Methods section, we will begin with the Case/Control Design for Common Variants, especially since the admixture aspects are largely the same across the four simulation types.

2.3.1 Case/Control Design for Common Variants

1,000 cases and 1,000 controls were simulated. Every fifth locus was used to form the simulated haplotypes and genotypes; a minor allele threshold of 0.02 was used.

A disease prevalence was set at 0.2, 0.1, and 0.3 for CEU, CHB, and YRI, respectively.

For cases, 300 individuals were designated to have target admixture proportions of 100%, 100 each for CEU, CHB, and YRI. 300 individuals were targeted to have 50% admixture proportions, 100 for each pairwise combination between CEU, CHB, and YRI. 100 individuals were designated to have 1/3 admixture between the three. 300 individuals had target admixture proportions generated from the Dirichlet distribution with means of 1/3.

Target admixture proportions for the controls were specified the same way. The target admixture proportions are summarized in Table 2.2.

Table 2.2: Target mean admixture proportions for example case/control simulations, for each of 1,000 cases and 1,000 controls.

Group	Type	Inds.	CEU	CHB	YRI
1	Fixed	100	1	0	0
2	Fixed	100	0	1	0
3	Fixed	100	0	0	1
4	Fixed	100	$\frac{1}{2}$	$\frac{1}{2}$	0
5	Fixed	100	$\frac{1}{2}$	0	$\frac{1}{2}$
6	Fixed	100	0	$\frac{1}{2}$	$\frac{1}{2}$
7	Fixed	100	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
8	Dirichlet	300	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

22 chromosomes were used, with effect sizes of $\{1, 0, -1, 0\}$ for the first 20, and $\{0, 0\}$ for the last two chromosomes. This is shown in Table 2.3, along with the mean allele counts for case and control individuals tabulated post-simulation. As expected, for SNPs with positive effect sizes, cases had noticeably higher mean counts than that of controls. For SNPs with negative effect sizes, controls analogously had higher mean counts than that of cases. When the effect sizes for SNPs were 0, the mean counts were similar between that of cases and controls.

Table 2.3: Comparison of mean alternative allele counts between 1,000 simulated case and 1,000 simulated control individuals.

Chr.	SNP	Effect Size	Cases	Controls
1	rs72638544	1	0.22	0.07
2	rs7579608	0	0.19	0.20
3	rs10510310	-1	0.09	0.22
4	rs7685911	0	0.38	0.34
5	rs1901498	1	1.34	0.87
6	rs4960131	0	1.02	1.05
7	rs10237521	-1	0.90	1.38
8	rs4433167	0	1.30	1.29
9	rs7865441	1	1.21	0.78
10	rs4584478	0	0.53	0.53
11	rs9804657	-1	0.51	0.94
12	rs55768228	0	1.48	1.40
13	rs7991629	1	0.55	0.26
14	rs28721413	0	1.22	1.22
15	rs7180505	-1	1.08	1.52
16	rs9935770	0	0.79	0.84
17	rs11247560	1	1.34	0.89
18	rs62078854	0	0.49	0.49
19	rs12980335	-1	0.50	0.97
20	rs156338	0	1.24	1.23
21	rs9980561	0	0.72	0.81
22	rs5746362	0	1.29	1.32

The estimated admixture proportions for cases, calculated from the simulated genotypes, are shown in Figure 2.6. Controls looked very similar. Function `sNMF` from R package `LEA`, which provides `STRUCTURE`-like functionality for admixture, was used to calculate the estimated admixture proportions (Francois, 2016; Frichot and Francois, 2014). `STRUCTURE` is one of the most popular tools for admixture estimation in population genetics analyses (Pritchard et al., 2000; Porras-Hurtado et al., 2013).

The admixture calculations were done in unsupervised fashion—i.e. no reference populations were designated as the three ancestral populations. The mean estimated admixture proportions by group indeed reflected the target admixture proportions. There is some variance across individuals present due to the stochastic nature of the artificial recombination

process and perhaps some noise from the admixture estimation procedure.

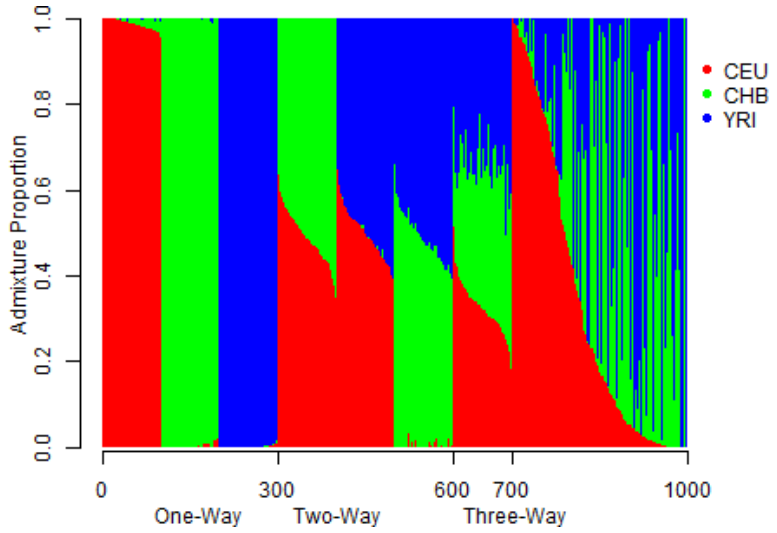

Figure 2.6: Admixture plots for case individuals. "One-Way" refers to individuals with target admixture proportions of 1 for a given population. "Two-Way" refers to individuals with target admixture proportions of $\frac{1}{2}$ for each of two populations, and likewise "Three-Way" for three populations of $\frac{1}{3}$ each. 1,500 loci were randomly sampled from each of the chromosomes. The admixture proportions were calculated in an unsupervised fashion by function `sNMF` from R package `LEA` (Francois, 2016; Frichot and Francois, 2014)

PCA plots were also constructed using 1,500 loci randomly sampled from each of the 22 chromosomes, as another way to look at admixture results across groups. The first two principal components for cases are shown in Figure 2.7. The individuals with targeted proportions of 100% of the ancestral populations indeed tightly cluster amongst themselves and comprise the nodes of a triangle, and the individuals with 50% of two populations form the edges. The individuals with targeted proportions of $\frac{1}{3}$ each are found in the center of the triangle, and the Dirichlet-targeted individuals are scattered across the triangle. PCA results for controls looked largely the same.

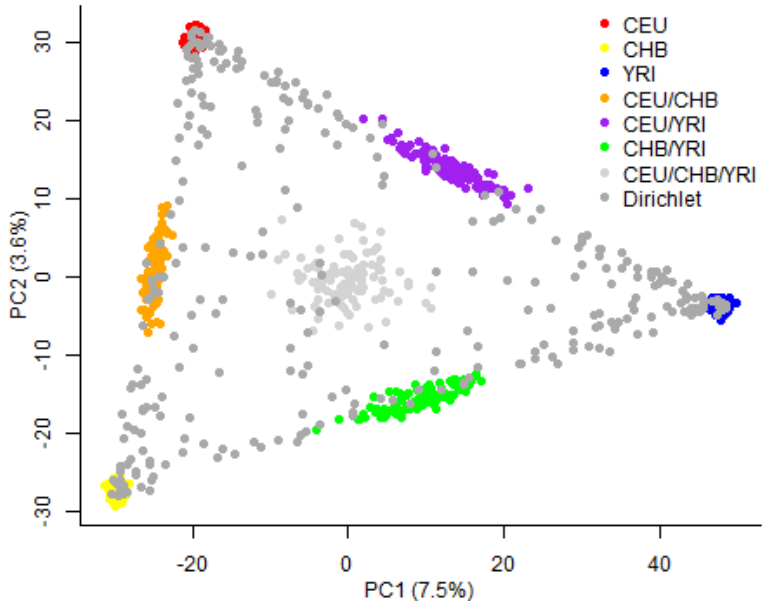

Figure 2.7: PCA plot for 1,000 case individuals, using 22 chromosomes. 1,500 loci were randomly sampled from each of the chromosomes to calculate the principal components.

The 300 case individuals with Dirichlet-generated admixture proportions are depicted in Figure 2.8. Each plot compares targeted versus realized admixture for a given ancestral source population. Targeted proportions were those drawn from the Dirichlet distribution, and the realized admixture proportions were provided by `sNMF`.

The Manhattan plot in Figure 2.9 displays GWAS results, conducted upon the simulated genotypes and phenotypes (a binary variable with 1 for case individuals, 0 for controls). The first four principal components were used as covariates in logistic regressions. 2,500 SNPs were randomly chosen to compute the first four principal components. The 1,000 closest loci to each causal SNP were used in the GWAS to form windows.

We see that in each window containing a causal SNP with an effect size of -1 or 1, the -log10(p-values) peak at that SNP. In contrast, no peaks are found in the windows where the causal SNPs had an effect size of 0. Loci closer to the causal SNPs had higher rates of LD, as conveyed via R^2 values.

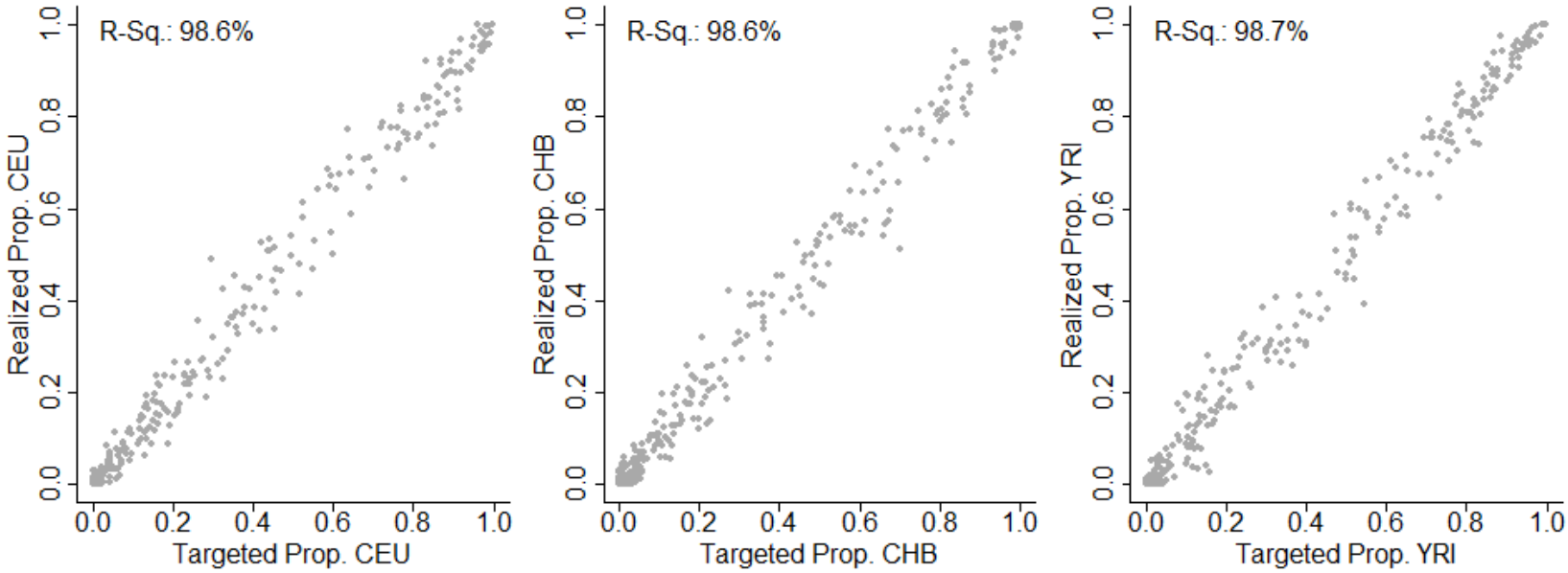

Figure 2.8: Targeted vs. realized admixture proportions for the 300 case individuals with Dirichlet-drawn targeted admixture proportions, by the three ancestral source populations.

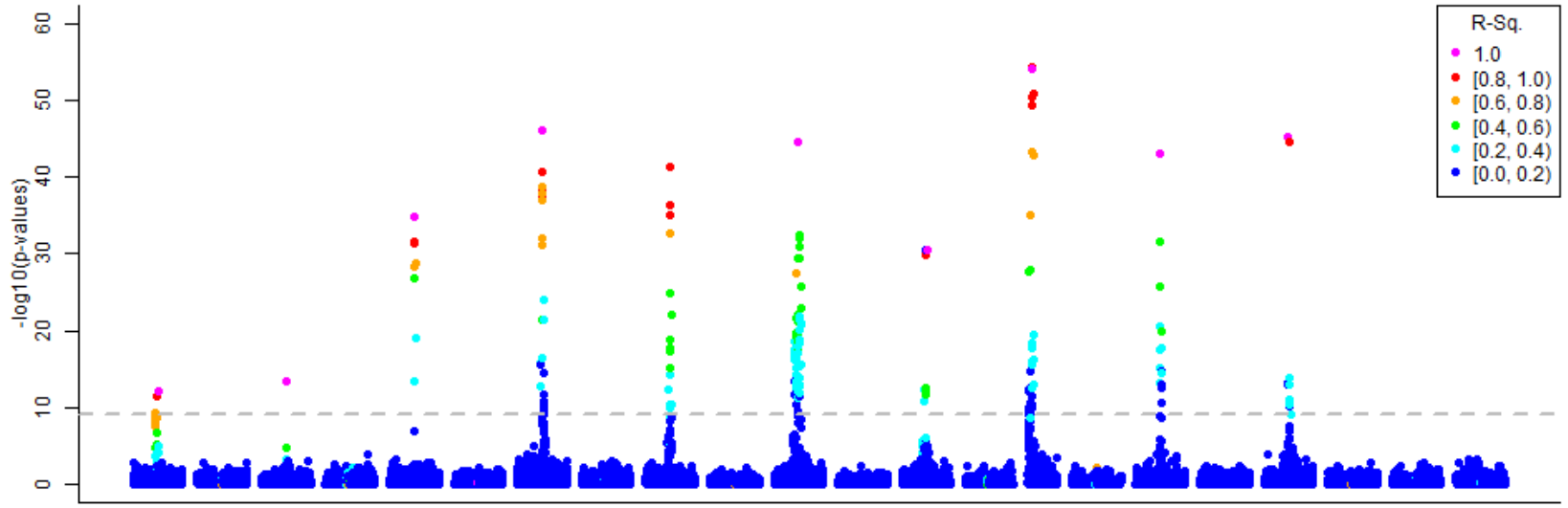

Figure 2.9: Manhattan plot of simulated GWAS results for the case/control design using common variants, with the first four principal components as covariates alongside the SNP values. The causal SNP is highlighted in magenta. The 1,000 loci closest to the causal SNP are depicted in each window. The grey dashed line corresponds to a Bonferroni-corrected p-value of 0.05, applied at the genome-wide level of 84.4 million. The R^2 values correspond to the degree of linkage disequilibrium for the loci with the nearest causal SNP.

2.3.2 Quantitative Phenotype for Common Variants

Using a similar admixture set-up as for the case/control design, 2,000 individuals were simulated for a quantitative phenotype. Loci rs12980335, rs156338, rs9980561, rs6516850, rs2838791, and rs5746362 were specified as the causal loci, with effect sizes of 5, 0, 10, 0, and 15, respectively. The specified standard deviation was 15. The specified population means for CEU, CHB, and YRI were 100, 105, and 85, respectively.

The Manhattan plot in Fig. 2.10 displays GWAS results, conducted upon the simulated genotypes and phenotypes, using the first four principal components as covariates. 2,000 SNPs were randomly chosen to compute the first four principal components. As seen for the case/control design for common variants, the points in the windows peak where causal SNPs with non-zero effect sizes are found. Similarly, proximity of loci to the causal SNPs is positively associated with higher rates of LD.

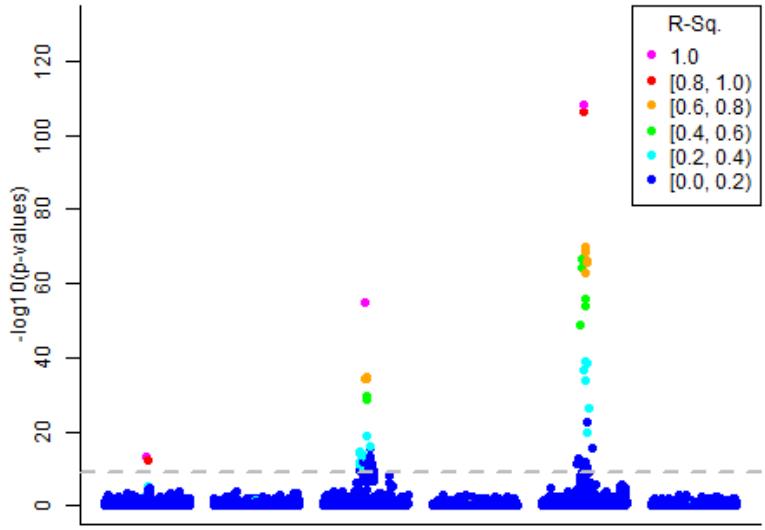

Figure 2.10: Manhattan plot of simulated GWAS results for the quantitative phenotype using common variants, with the first four principal components as covariates alongside the SNP values. The causal SNP is highlighted in magenta. The 1,000 loci closest to the causal SNP are depicted in each window. The grey dashed line corresponds to a Bonferroni-corrected p-value of 0.05, applied at the genome-wide level of 84.4 million. The R^2 values correspond to the degree of linkage disequilibrium for the loci with the nearest causal SNP.

The simulation for a quantitative phenotype with common variants was re-run, this

time with the effect sizes for the SNPs being all 0, and the other settings left unchanged. The q-q plot in Figure 2.11 illustrates that neglecting to account for population stratification can result in more false positives, as the grey line (corresponding to -log10 p-values of the GWAS without the first four PCs as covariates) diverges from the unit line much sooner than does the blue line (corresponding to -log10 p-values of the GWAS including the first four PCs as covariates).

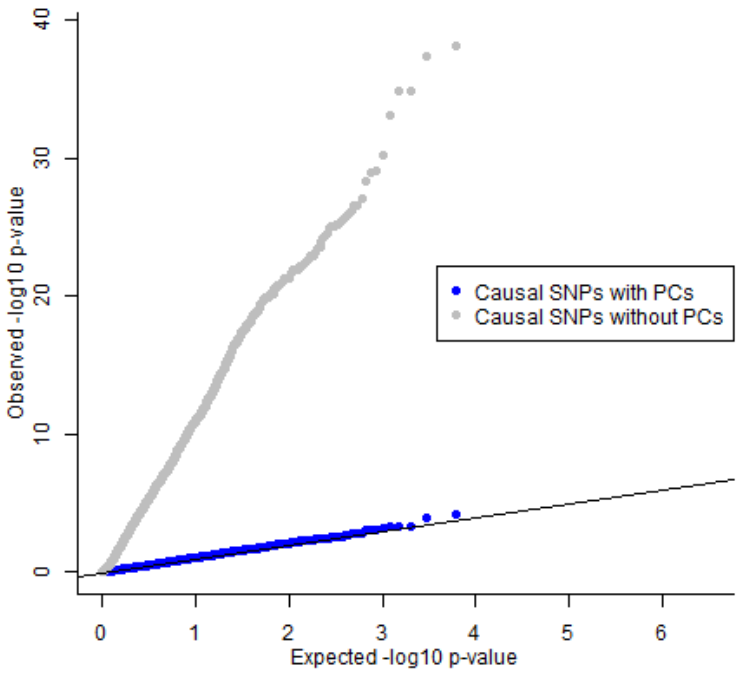

Figure 2.11: q-q plot of simulated GWAS results, comparing observed and expected -log10 p-values. The effect sizes for the SNPs were set at 0.

2.3.3 Case/Control Design for Rare Variants

Using chromosomes 19, 20, 21, and 22, 1,000 individuals were simulated each for cases and controls, where their target admixture set-up was the same as that used previously for common variants. Every fifth locus was used. The prevalences were 0.2, 0.1, and 0.3 for CEU, CHB, and YRI, respectively. The mean effect size for rare variants was set at 0.01, under a Madsen-Browning weighting scheme. The results are shown in Table 2.4.

Table 2.4: Comparison of probabilities of being a case, calculated ex-post for 1,000 case and 1,000 control individuals, by admixture group and case/control status.

Group	Desc.	Case	Cont.
1	CEU	0.26	0.17
2	CHB	0.15	0.07
3	YRI	0.46	0.18
4	CEU/CHB	0.21	0.12
5	CEU/YRI	0.50	0.38
6	CHB/YRI	0.43	0.26
7	CEU/CHB/YRI	0.35	0.34
8	Dirichlet	0.33	0.22

We can note a few general trends, e.g. that the cases have higher scores than the controls. This would be expected, as the specified mean effect size was positive. The admixed groups generally fall in between their prescribed ancestral groups.

2.3.4 Quantitative Phenotype for Rare Variants

Similar to the example simulations for case/control burden-testing, 2,000 individuals were simulated for a quantitative trait phenotype, where their target admixture set-up were the same as that used previously for common variants. Every fifth locus was used from chromosomes 19, 20, 21, and 22 . The specified mean values for each ancestral population were 100, 105, and 85 for CEU, CHB, and YRI, respectively, with a standard deviation of 15. The mean effect size was set at 0.01, under a Madsen-Browning weighting scheme. The mean phenotype scores are shown in Table 2.5.

Table 2.5: Comparison of mean quantitative traits for 2,000 individuals scores by group.

Group	Desc.	Score
1	CEU	100.69
2	CHB	105.67
3	YRI	87.80
4	CEU/CHB	103.60
5	CEU/YRI	91.65
6	CHB/YRI	94.05
7	CEU/CHB/YRI	95.44
8	Dirichlet	96.17

Despite some noise from the simulation process, the mean phenotype scores of the non-admixed groups (groups 1-3) are generally close to the specified target means for those ancestral populations. The mean phenotype scores for the admixed groups lay in between that of their ancestral populations.

2.4 Discussion

The updated HAP-SAMPLE, HAP-SAMPLE2, provides an efficient approach to simulating genotype-phenotype data-sets, for both case/control designs and quantitative phenotypes, and is compatible with large-scale data-sets such as 1000 Genomes. Data-sets can be pre-processed to reduce the computation burden. The admixture functionality in HAP-SAMPLE2 can create admixed populations and/or preserve population substructure, while introducing novelty via the artificial recombination process. Targeted admixture proportions can be specified at the individual-level. HAP-SAMPLE2 also now supports burden-testing for rare-variants, under both the fixed and Madsen-Browning weighting schemes.

Currently in our approach, chromosomes are limited to one causal SNP per chromosome for the case/control design for common variants. We use the Hardy-Weinberg assumptions in handling loci on different chromosomes. Under the current framework, this would likely not be an appropriate assumption for loci on the same chromosome, which are likely to exhibit some linkage disequilibrium. Future work could involve adding the capability to allow more than

one causal SNP(s) per chromosome.

For quantitative traits, the standard deviations for different ancestral populations are currently assumed to be equivalent. This assumption could perhaps be relaxed, and the user could specify population-specific standard deviations.

Support for dominance and interactive effects could be of interest. This would be an area of further development. Particularly for interactive effects, careful consideration would be needed to not exacerbate computation time and/or space.

Assortative mating could be another area of future attention. Assortative mating in humans "is almost universally positive" for similarity (Robinson et al., 2017), with a material genetic basis—for many traits, mates tend to be more similar genetically than would be expected on average. This has implications for human phenotypes as broad and varied as stature (Robinson et al., 2017), educational attainment (Robinson et al., 2017; Domingue et al., 2014), and longevity (Ruby et al., 2018).

Currently in HAP-SAMPLE2, mating is assumed to be random. Source haplotypes, conditional on population (and causal SNP value if it is a disease chromosome in the case/control design), have equal chances of being selected to construct a given region of a simulated chromosome. Functionality could be added to the simulation process to have greater probabilities of selecting mutually similar haplotypes to construct simulated chromosomes. This could largely leverage existing conceptual knowhow and infrastructure, but would necessitate additional development. Haplotype-similarity would need to be defined, either as stored pre-calculated values, or calculated on the fly.

The current focus of HAP-SAMPLE2 is upon autosomal chromosomes, as 22 of the 23 human chromosomes are autosomal. However, it would certainly be interesting to explore sex chromosomes, as well—seeing as there is a plethora of sex-linked traits in humans, even encompassing complex cognitive traits (Trent and Davies, 2012; Griffiths et al., 2000; Skuse, 2005). Some traits have received attention from various fields, due to potential sexual dimorphism in phenotypic expression, clinical diagnoses, and etiology. One notable example would be autism (Retico et al., 2016), which coincidentally has been suspected to be also implicated

with rare variants (Brandler et al., 2018). For HAP-SAMPLE2, incorporation of sex chromosomes could have implications for different male/female variances in phenotypes (Brown et al., 2009). To the extent different human populations may have had experienced varied differentials in historical and/or pre-historical female-to-male breeding ratios (Labuda et al., 2010), this could have intriguing tie-ins with the admixture functionality of HAP-SAMPLE2.

Our current implementation for rare variants performs burden-testing under the Madsen-Browning and fixed weighting schemes. Further work could involve adding new weighting schemes. For example, Wu et al. (2011) suggested beta densities, i.e. $w_s = beta(f_s, a_1, a_2)$, of which the Madsen-Browning weighting scheme would be a special case (Lee et al., 2014). Wu et al. (2011) recommended values of $a_1 = 1$ and $a_2 = 25$. Another possible pathway would be introducing more flexible testing, e.g. in the spirit of SKAT and similar approaches—to accommodate situations where the effects of rare variants may not be all in the same direction (Lee et al., 2014). Another avenue would be to allow the effect of rare variants to be different across groups—for example, rare variants in different regions of chromosome and/or across different chromosomes may have different effect sizes. This would essentially entail accepting user-specification of different mean rare variant effect sizes for groups.

As HAP-SAMPLE2 is a resampling-based approach, the number of rare variants per chromosome a simulated individual possesses is, on average, limited by the number of extant rare variants in the source haplotypes. Further development could involve introducing more rare variants per simulated individual via perturbation. That is, each simulated haplotype could be perturbed to include more rare variant alleles than would otherwise be expected, as prescribed by the source haplotypes. This could generate simulated individuals with "fat-tailed" phenotypes, which could be of value to researchers studying more extreme traits (Peloso et al., 2016).

Now with both functionality for common variants and rare variants separately, a natural next-step would be to integrate haplotype sampling and construction for phenotypes that potentially are influenced by both common and rare variants. Many traits, such as cognitive ability, may have an association with both common and rare variants (Luciano et al., 2015).

Multivariate phenotypes could also be another area for further development, especially leveraging existing infrastructure for common and/or rare variants. For example, a multivariate phenotype could be generated from user-specified values of coefficients and covariance/correlation values for a multivariate latent variable model. The user would also specify which components of the multivariate phenotype are binary, and which are continuous. Although this approach would be unable to take advantage of some of the beneficial aspects of the case/control framework, it could be achievable via an adaptation of the current framework for quantitative traits.

3

Fast Multivariate Probit Estimation via a Two-Stage Composite Likelihood

3.1 Introduction

The multivariate probit is a standard model for correlated binary data, with advantages due to its flexibility in handling correlation structures and interpretability of parameters (Chib and Greenberg, 1998). The approach is conceptually simple, as the underlying multivariate latent normal requires specification of only means and covariances. However, for likelihood parameter estimation, the integrals for calculating the likelihood from the multivariate cumulative normal distribution are computationally intensive (Chib and Greenberg, 1998; Moffa and Kuipers, 2014; Genz, 1992), e.g. as detailed in documentation for software such as the R package `mvProbit` (Henningsen, 2019).

We consider the standard multivariate probit, where binary components of the multivariate response Y are modeled as the binarized result of a latent multivariate distribution. For identifiability, we assume unit marginal latent variances, i.e. the covariance matrix is a correlation matrix. With K binary response components, this implies $\binom{K}{2}$ correlation values. We consider the N by P design matrix $\boldsymbol{X}$ as shared across components, as well as the $P \times K$ coefficient matrix $\boldsymbol{B}$, where N, P, and K are the number of observations, predictors, and components, respectively. The role of $\boldsymbol{X}$ is to serve as a predictor matrix in a regression framework for the latent outcome. For the multivariate probit, the role of the coefficients (in conjunction with the design matrix) can be viewed as specifying the mean for the latent multivariate normal probability, with the region of integration being $(-\infty, 0]$ or $(0, \infty)$ for a given component depending on whether the response is 0 or 1. Commonly, as we will do here, the mean is fixed at $\mathbf{0}$ and instead the coefficients help determine the boundaries of integration—a numerically-equivalent representation.

Thus, for multivariate binary response $\boldsymbol{Y}$ with K components the full likelihood is:

$$L_{full}(\boldsymbol{\theta}; \boldsymbol{y_i}) = \int_{A_{i1}} \dots \int_{A_{iK}} \phi(\boldsymbol{z_i}, \boldsymbol{0}, \boldsymbol{\Sigma}) \boldsymbol{dz_i}$$

$$A_{ik} = \begin{cases} (-\infty, \boldsymbol{x_i \beta_k}] & y_{ik} = 1 \\ (\boldsymbol{x_i \beta_k}, \infty) & y_{ik} = 0 \end{cases}$$

$$\boldsymbol{\theta} = \{\boldsymbol{B}, \boldsymbol{\Sigma}\}$$

Where i corresponds to a given observation, and k a given component of $\boldsymbol{Y}$. The latent multivariate normal variable is assumed to have a mean vector of $\boldsymbol{0}$, with the constants of integration determined by whether observed values of the multivariate binary response are 0 or 1.

We do not apply constraints to the correlations, and issues of positive-definiteness are addressed below. With PK coefficient parameters and $K(K-1)/2$ correlation parameters, the number of parameters grows quickly with increasing number of components.

Moffa and Kuipers (2014) proposed a sequential expectation-maximization Monte Carlo method to estimate parameters in the multivariate probit. The approach builds on (Chib and Greenberg, 1998) and utilizes the truncated multivariate T distribution, with heavier tails than the normal. The approach can be computationally intensive, with variability in results due to the stochastic sampling.

Mullahy (2016) proposed that multivariate probit estimation be performed via "chained" bivariate probits. Each element in the correlation matrix is estimated pairwise for components in the response, and coefficient estimates are obtained by averaging over coefficient estimates obtained from the bivariate pairings. The approach is computationally attractive, but statistical efficiency and other properties remain unclear. The chained bivariate probit approach is implemented in Mata's `bvpmvp`, as a faster alternative to Stata's `mvprobit` (Mullahy, 2016). Stata's `mvprobit` (Cappellari and Jenkins, 2003) and R's `mvProbit` both use the GHK (Geweke, Hajivassiliou and Keane) approach to simulate multivariate normal probabilities, and both can be computationally inefficient.

Feddag (2013) suggested using a composite pairwise likelihood approach in the context of estimating multivariate probit longitudinal models. In this formulation an unconstrained covariance matrix used (instead of correlation), and identifiability assured by including a mean term and constraining coefficients to sum up to 0 across components. Simulations were performed with response variables of 3, 5, and 8 components with 50, 100, and 300 observations. Feddag (2013) noted empirically that the general pairwise likelihood results retained nearly full statistical efficiency compared to using the full likelihood, but was much faster computationally. The standard deviations of coefficient estimates across simulations were similar between the full likelihood and their composite likelihood. For an example with 3 components and 300 observations, maximizing the pairwise likelihood took 0.16 minutes a desktop computer to converge, whereas the full likelihood took 27.3 minutes.

Jin (2009) also found good performance of a composite pairwise likelihood for binary data using a different model. Pairwise likelihood also performed well in efficiency and computation time.

In a larger exploration of a multivariate normal model, Zhao and Joe (2005) included a similar two-Stage composite likelihood for multivariate probit in simulations, where the first stage consisted of univariate marginals, and in the second stage bivariate marginals. However, they were chiefly concerned with analysis of data in familial units. Thus, their simulations were performed using models in which there were two mutual coefficients $\{\beta_0, \beta_1\}$ across components (that is, coefficients across components were constrained to equality) and where correlation parameters across components were either one ρ_1 or ρ_2, corresponding to parent-offspring or sibling-sibling correlations. Each of their simulations used 2,000 families. The authors noted that the two-stage composite likelihood is faster to compute than the pairwise composite likelihood, and both are more computationally efficient than full maximum likelihood.

Ghosh et al. (2013) introduced a bivariate logistic model that includes an intermediate latent probit step. The approach, originally designed to handle bias in outcome-dependent sampling situation, has considerable flexibility in handling nuisance covariates. However, it is not easily extensible beyond $K = 2$.

To address issues of computational efficiency while retaining a balance of simplicity and flexibility, we introduce a novel two-stage composite likelihood approach for multivariate probit estimation. This approach is designed to be fast, and thus amenable to situations where many potential predictors are screened, such as with genome-wide association studies. Coefficient standard errors are obtained using a sandwich estimator appropriate for a composite likelihood. In contrast to Zhao and Joe (2005), we focus upon multivariate probit models with unconstrained parameters, and show that our model can achieve impressive reduction in computation time while largely maintaining statistical efficiency.

3.2 Methods

3.2.1 Two-Stage Estimation

Two-stage, or "two-step," likelihood estimation (Hardin, 2002) is an option for analytically or computationally difficult likelihood and/or log-likelihood functions. In two-stage estimation, an original model is essentially split into two models, with the first embedded in the second. The first stage estimates parameters associated with the first likelihood, and the second stage, the second likelihood.

Following (Hardin, 2002), suppose we start with a full log-likelihood function and n independent observations:

$$\ln L(\theta_1, \theta_2) = \sum_{1}^{n} \ln f(y_{1i}, y_{2i} | x_{1i}, x_{2i}, \theta_1, \theta_2)$$

The parameter vector θ_1 is associated with data x_1, y_1, and in the first stage, the parameters in θ_1 are estimated by maximizing:

$$\ln L(\theta_1) = \sum_{1}^{n} \ln f_1(y_{1i} | x_{1i}, \theta_1)$$

In the second stage, the estimates of θ_1 from the first stage can be used as fixed inputs to

maximize the conditional likelihood:

$$\ln L(\theta_2) = \sum_{1}^{n} \ln f_2(y_{2i}|x_{2i}, \theta_2, (x_{1i}, \hat{\theta_1}))$$

y_{2i} is a subset of responses from the i'th observation of response y, and y_{1i} is another subset. x_{1i} and x_{2i} are their counterparts in the design matrix. θ_1 and θ_2 partition the full parameter vector θ. Either or both stages can be considered misspecified likelihoods, and for certain problems, maximizing the conditional likelihood is equivalent to maximizing the full likelihood.

3.2.2 Composite Likelihood

A composite likelihood is formed by the product of so-called "associated" or "'sub"-likelihoods, which are individually proper likelihoods. Like two-stage estimation, composite likelihoods are a popular alternative when maximizing the full likelihood is computationally difficult. Lindsay et al. (2011) provide an overview of theoretical properties and construction strategies. For proper associated likelihoods, the composite likelihood is generally consistent, but may suffer a loss in efficiency compared to the full likelihood (Lindsay et al., 2011). Suppose there are A associated likelihoods. For each observation i we write the composite likelihood:

$$L_{comp}(\boldsymbol{\theta}; \boldsymbol{y_i}) = \prod_{a=1}^{A} f(\boldsymbol{y_i}; \boldsymbol{\theta})^{w_a}$$

Where $\boldsymbol{\theta}$ is the parameter vector and w_a denotes a weight for the a'th associated likelihood. This weight parameter is fixed in advance, and can be 1 for all associated likelihoods. Varin et al. (Varin et al., 2011) provided a review of composite likelihood methods and remarked upon their practicality and high statistical efficiency. However, the efficiency and asymptotic properties can depend on the full likelihood and the particular composite likelihood set-up (e.g. marginal vs. conditional likelihoods and the complexity of the associated likelihoods) (Varin et al., 2011; Cattelan and Sartori, 2016).

3.2.3 Two-Stage Composite Likelihood for the Multivariate Probit

We estimate parameters for the multivariate probit likelihood using a composite likelihood, and divide the composite likelihood estimation process into two stages. In the first stage, we obtain coefficient estimates from a composite likelihood consisting of univariate marginals. Each associated likelihood involves one component from the response, which for our setting is the univariate probit:

$$\ln L_{uni}(\boldsymbol{B}; \boldsymbol{y_i}, \boldsymbol{x_i}) = \sum_{k=1}^{K} \ln f(y_{ik}, \boldsymbol{x_i}; \boldsymbol{\beta_k})$$

As no parameters are shared across associated likelihoods here, for estimation we can write:

$$max_{\boldsymbol{B}} \sum_{i=1}^{N} \ln L_{uni}(\boldsymbol{B}; \boldsymbol{y_i}, \boldsymbol{x_i}) = max_{\boldsymbol{B}} \sum_{i=1}^{N} \sum_{k=1}^{K} \ln f(y_{ik}, \boldsymbol{x_i}; \boldsymbol{\beta_k}) =$$

$$max_{\boldsymbol{\beta_1}} \sum_{i=1}^{N} \ln f(y_{i1}, \boldsymbol{x_i}; \boldsymbol{\beta_1}) + ... + max_{\boldsymbol{\beta_K}} \sum_{i=1}^{N} \ln f(y_{iK}, \boldsymbol{x_i}; \boldsymbol{\beta_k})$$

Since these associated likelihoods can be estimated independently, this set-up simplifies the computational process. For example, R's `glm()` can provide coefficient estimates for each of the components.

For each associated likelihood in the first stage we have univariate probit likelihoods:

$$f(y_{ik}, \boldsymbol{x_i}; \boldsymbol{\beta_k}) = \int_{A_{ik}} \phi(z_{ik}, 0, 1) dz_{ik}$$

$$A_{im} = \begin{cases} (-\infty, \boldsymbol{x_i \beta_k}] & y_{ik} = 1 \\ (\boldsymbol{x_i \beta_k}, \infty) & y_{ik} = 0 \end{cases}$$

The area of integration is determined by the i'th observation of the design matrix, the coefficients for the component, β_k, and the binary observation for that component, y_{ik}. As with the general multivariate case, the mean is fixed at 0 here for the univariate probit, and the variance at 1.

In the second stage, we estimate the correlation parameters, using as inputs the coef-

ficient estimates from the first stage. Here each associated likelihood involves a pair of components, each a bivariate probit:

$$\ln L_{pair}(\boldsymbol{\Sigma}; \boldsymbol{y_i}, \boldsymbol{x_i}, \hat{\boldsymbol{B}}) = \sum_{j=1}^{K-1} \sum_{k=j+1}^{K} \ln f(y_{ij}, y_{ik}, \boldsymbol{x_i}; \rho_{jk}, \hat{\boldsymbol{\beta}}_{\boldsymbol{j}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{k}}).$$

Again, no parameters are shared across associated likelihoods, so we approach maximization one associated likelihood at a time:

$$max_{\boldsymbol{\Sigma}} \sum_{i=1}^{N} \ln L_{pair}(\boldsymbol{\Sigma}; \boldsymbol{y_i}, \boldsymbol{x_i}, \hat{\boldsymbol{B}}) = max_{\boldsymbol{\Sigma}} \sum_{i=1}^{N} \sum_{j=1}^{K-1} \sum_{k=j+1}^{K} \ln f(y_{ij}, y_{ik}, \boldsymbol{x_i}; \rho_{jk}, \hat{\boldsymbol{\beta}}_{\boldsymbol{j}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{k}}) =$$

$$\Big(max_{\rho_{1,2}} \sum_{i=1}^{N} \ln f(y_{i1}, y_{i2}, \boldsymbol{x_i}; \rho_{1,2}, , \hat{\boldsymbol{\beta}}_{\mathbf{1}}, \hat{\boldsymbol{\beta}}_{\mathbf{2}}) + ... +$$

$$max_{\rho_{K-1,K}} \sum_{i=1}^{N} \ln f(y_{i(K-1)}, y_{iK}, \boldsymbol{x_i}; \rho_{K-1,K}, \hat{\boldsymbol{\beta}}_{\boldsymbol{K-1}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{K}}) \Big)$$

Each bivariate probit inherits coefficient estimates from the first stage. For each associate likelihood involving a pair that consists of components j and k, we have:

$$f(y_{im}, \boldsymbol{x_i}; \rho_{jk}, \hat{\boldsymbol{\beta}}_{\boldsymbol{j}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{k}}) = \int_{A_{ij}} \int_{A_{ik}} \phi(\boldsymbol{z_i}, \mathbf{0}, \boldsymbol{\Sigma_{jk}}) \boldsymbol{dz_i}$$

$$A_{ij} = \begin{cases} (-\infty, \boldsymbol{x_i}\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}] & y_{ij} = 1 \\ (\boldsymbol{x_i}\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}, \infty) & y_{ij} = 0 \end{cases}$$

$$A_{ik} = \begin{cases} (-\infty, \boldsymbol{x_i}\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}] & y_{ik} = 1 \\ (\boldsymbol{x_i}\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}, \infty) & y_{ik} = 0 \end{cases}$$

$$\boldsymbol{\Sigma_{jk}} = \begin{pmatrix} 1 & \rho_{jk} \\ \rho_{jk} & 1 \end{pmatrix}$$

The areas of integration for a particular observation i of the two components are determined by the observation of the design matrix, their coefficient estimates (inherited from the

first stage) and the observed binary responses. The mean for both components is fixed at 0. The variances for the two components are both fixed at one, leaving their pairwise correlations, ρ_{jk} for components j and k, to be estimated.

The primary gain in computational efficiency arises from this component-wise estimation, which we can implement using simple maximization routines such as R's `optim()`. In contrast to models where parameters are constrained, such as the familial models explored by Zhao and Joe (2005), here maximization can proceed independently.

3.2.4 The Robust Murphy-Topel Estimator of Variance for the Two-Stage Composite Likelihood

Composite likelihoods can be considered misspecified likelihoods, as the associated likelihoods do not fully reflect data dependencies. Hardin (2002) describes a robust variance estimator to account for both misspecification and the two-stage nature of the estimation process– essentially a "sandwich" version of the Murphy-Topel variance estimator (Murphy and Topel, 1985) for two-stage models. Following Hardin (2002), let $\boldsymbol{V_S(\theta_1)}$ denote the robust variance estimator for $\hat{\boldsymbol{\theta}}_\mathbf{1}$ estimated in the first stage, and $\boldsymbol{V_S(\theta_2)}$ for $\hat{\boldsymbol{\theta}}_\mathbf{2}$ in the second stage, with $\boldsymbol{Cov_S(\theta_1, \theta_2)}$ the covariance between them. We have

$$
\begin{aligned}
\boldsymbol{V_S(\theta_1)} &= \boldsymbol{V_1 V_1^{*-1} V_1} = \boldsymbol{V_{S1}} \\
\boldsymbol{Cov_S(\theta_1, \theta_2)} &= \boldsymbol{V_1 R^T V_2} - \boldsymbol{V_{S1} C^{*T} V_2} \\
\boldsymbol{V_S(\theta_2)} &= \boldsymbol{V_2 V_2^{*-1} V_2} + \boldsymbol{V_2}(\boldsymbol{C^* V_{S1} C^{*T}} - \boldsymbol{R V_1 C^{*T}} - \boldsymbol{C^* V_1 R^T})\boldsymbol{V_2} \\
&= \boldsymbol{V_{S2}} + \boldsymbol{V_2}(\boldsymbol{C^* V_{s1} C^{*T}} - \boldsymbol{R V_1 C^{*T}} - \boldsymbol{C^* V_1 R^T})\boldsymbol{V_2}
\end{aligned}
$$

where $\boldsymbol{V_1}$ is the non-robust (naive) likelihood-based asymptotic variance estimator for the stage one parameters $\boldsymbol{\theta_1}$ based upon the stage one log-likelihood $\ln L_1(\boldsymbol{\theta_1})$, e.g. the expected value of the negative second derivatives, and $\boldsymbol{V_1^*}$ the expected value of the matrix of outer product of gradients. Similarly, $\boldsymbol{V_2}$ is the non-robust asymptotic variance estimator for the stage two parameters $\boldsymbol{\theta_2}$ based upon the stage two conditional log-likelihood $\ln L_2(\boldsymbol{\theta_2}|\boldsymbol{\theta_1})$, and $\boldsymbol{V_2^*}$ the expected value of the matrix of outer gradients. $\boldsymbol{C^*}$ is the sub-matrix of the expected value

of the negative second derivatives based on $\ln L_2(\boldsymbol{\theta_2}|\boldsymbol{\theta_1})$, the rows corresponding to $\boldsymbol{\theta_2}$ and the columns corresponding to $\boldsymbol{\theta_1}$. $\boldsymbol{R}$ is be the sub-matrix of the expected value of the negative second derivatives based on $\ln L_2(\boldsymbol{\theta_2}|\boldsymbol{\theta_1})$ and $\ln L_1(\boldsymbol{\theta_1})$, the rows corresponding to $\boldsymbol{\theta_2}$ and the columns corresponding to $\boldsymbol{\theta_1}$. $\boldsymbol{V_{s1}}$ is be the usual sandwich estimator for the stage one parameters, and $\boldsymbol{V_{s2}}$ that of the second stage parameters (treating stage one parameters as fixed). Empirical plug-in estimates for the matrix elements are obtained by taking the mean across observations (using the final parameter estimates as inputs). Once estimated, these matrices can be used to calculate the robust Murphy-Topel estimate of variance (Murphy and Topel, 1985):

$$\boldsymbol{V_S}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{V_S}(\boldsymbol{\theta_1}) & \boldsymbol{Cov_S}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) \\ \boldsymbol{Cov_S}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) & \boldsymbol{V_S}(\boldsymbol{\theta_2}) \end{pmatrix}$$

3.3 Examples

3.3.1 Six Cities

The *Six Cities* data set has been a popular choice for comparing multivariate probit estimation methodologies. We performed our two-stage composite likelihood estimation on this data set, and compared it to the results of Chib and Greenberg (1998) and Moffa and Kuipers (2014). We were chiefly concerned with run-time and statistical efficiency, as judged by coefficient standard errors.

In the *Six Cities* data, wheezing status at ages 7, 8, 9, and 10 for 537 children were recorded as 0 or 1 to serve as the multivariate response, for four components with binary observations. Coefficients (shared across all components) included the intercept, age centered at 9, maternal smoking status (1 or 0), and an interactive variable between maternal smoking status and age. These were represented by β_0, β_1, β_2 and β_3, respectively. Note that the four components share coefficients, i.e. $\boldsymbol{\beta_1} = \boldsymbol{\beta_2} = \boldsymbol{\beta_3} = \boldsymbol{\beta_4} = \{\beta_0, \beta_1, \beta_2, \beta_3\}$. The covariance

(correlation) matrix has 6 off-diagonal entries, corresponding to correlations between wheezing status between various ages. Standard errors were calculated using the robust approach described in the previous section. 250 bootstrapped replications were performed for comparisons.

Using our two-stage composite likelihood, the *Six Cities* coefficient and correlation estimates are very similar to those previously published by Chib and Greenberg (1998) and Moffa and Kuipers (2014). Parameter estimation took about $\frac{1}{40}$ of a second for our model on a Windows 2.70 GHz Intel i7-7500 laptop. The bootstrapped standard errors and empirical standard errors obtained from the original data are similar to each other, and also similar to those provided by Chib and Greenberg (1998) and Moffa and Kuipers (2014). In summary, for these data the estimates do not reflect apparent loss in statistical efficiency, and the correspondence with bootstrapped standard errors indicates appropriateness of the robust variance estimates.

Table 3.1: Comparison of *Six Cities* mean parameter estimates between Chib and Greenberg (1998), Moffa and Kuipers (2014), and the Two-Stage Composite Likelihood. 250 replications were done for the Two-Stage Composite Likelihood to calculate mean parameter estimates and the mean log-likelihood value (its empirical standard error), and 250 bootstrapped replications for the bootstrapped standard errors. Empirical standard errors were calculated using the robust Murphy-Topel variance estimator. Parameter estimation took about 0.025 seconds for the Two-Stage Composite Likelihood.

	Chib & Greenberg (1998)		Moffa & Kuipers (2014)		Two-Stage CL (2019)		
Param.	Est.	SE	Est.	SE	Est.	BSE	ESE
β_0	-111.8	(6.5)	-112.3	(6.2)	-112.6	(6.5)	(6.3)
β_1	-7.9	(3.3)	-7.9	(3.1)	-7.7	(3.1)	(3.1)
β_2	15.2	(10.2)	15.9	(10.1)	17.1	(10.6)	(10.1)
β_3	3.9	(5.2)	3.8	(5.1)	3.7	(4.9)	(4.9)
σ_{12}	58.4	(6.8)	58.3	(6.6)	59.1	(6.5)	(6.6)
σ_{13}	52.1	(7.6)	52.2	(7.1)	53.1	(7.3)	(7.2)
σ_{14}	58.6	(9.5)	57.8	(7.4)	59.1	(7.5)	(7.2)
σ_{23}	68.8	(5.1)	68.6	(5.6)	69.2	(5.7)	(5.6)
σ_{24}	56.2	(7.7)	55.8	(7.4)	57.5	(7.5)	(7.3)
σ_{34}	63.1	(7.7)	62.7	(6.7)	64.1	(6.4)	(6.6)
Est. Log-Lik	-794.94	(0.69)	-794.95	(0.82)	-794.76	(—)	(0.00)
Corr. Log-Lik	-794.70		-794.61		-794.76		

Interestingly, the two-stage composite likelihood produces estimates that achieved a higher log-likelihood when inputted into the full information likelihood than did the log-

likelihoods from Chib and Greenberg (1998) or Moffa and Kuipers (2014). However, as pointed out by Moffa and Kuipers (2014), the stochastic nature of their processes (leading to noticeable variance across replications of log-likelihood estimations upon the same data set) may reduce the log-likelihood. They thus supply a correction calculation for the log-likelihood. The two-Stage composite likelihood does not require such a correction.

3.3.2 MEPS

Started in 1996, The Medical Expenditure Panel Survey (MEPS) is a set of surveys containing data on how American families and individuals use health services (for Healthcare Research and Quality, 2018). The R package `GJRM` (Marra and Radice, 2019) provides a 2008 MEPS data-set with 18,592 observations of individual characteristics and their binary-coded disease statuses for diabetes, hyperlipidemia, and hypertension. The function `gjrm()` of `GJRM` can perform a variety of semi-parametric model estimations, including fully parametric univariate, bivariate, and trivariate probit estimations using a general penalized maximum likelihood approach in conjunction with smoothing set-up via penalized regression splines (Marra and Radice, 2019).

Here we use the MEPS data-set to compare `gjrm()` and the Two-Stage Composite Likelihood in estimating a trivariate probit. The trivariate response consists of the three disease statuses, and individual characteristics serve as predictors: body mass index (BMI), age (in years), sex (1 for male, 0 for female), education (in years), income (log-transformed), race (coded as white, black, Native American, or other), and region (northeast, midwest, south, or west). 18,273 observations were retained after excluding those with incomes listed as zero.

250 replications were performed to record the bootstrapped standard errors. The mean run-time for `gjrm` on a Windows 2.70 GHz Intel i7-7500 laptop was about 32 seconds, whereas it was about 12 seconds for the Two-Stage Composite Likelihood. The coefficient estimated for the three components of Diabetes, Hyperlipidemia, and Hypertension are displayed below in Table 3.2. The correlation parameter estimates are included at the bottom.

The parameter estimates were generally similar, and the bootstrapped standard errors

for the two methods were nearly identical for both coefficients and correlation parameters. The reported standard errors from `gjrm()` and the empirical standard errors from the Two-Stage CL were generally close for coefficients. However, the reported standard errors from `gjrm()` were noticeably higher than the `gjrm()` bootstrapped standard errors, as well as both the empirical and bootstrapped standard errors from the Two-Stage CL.

Table 3.2: Comparison of estimates produced by `gjrm()` and the Two-Stage Composite Likelihood, displayed by component and coefficient, along with the correlation parameter estimates. The standard errors from `gjrm()` come from its native reporting; empirical standard errors for the Two-Stage Composite Likelihood were calculated using the robust Murphy-Topel variance estimator. Coefficient estimates multiplied by 100; correlation estimates displayed without rounding.

		gjrm()			Two-Stage CL		
Comp.	Param.	Est.	SE	BSE	Est.	ESE	BSE
Diabetes	Intercept	-375.4	20.6	21.6	-360.9	21.0	22.0
	BMI	5.3	0.2	0.2	5.1	0.2	0.3
	Age	3.9	0.1	0.2	3.9	0.1	0.2
	Sex	4.9	3.1	3.1	4.5	3.1	3.1
	Education	-3.7	0.5	0.5	-4.1	0.5	0.5
	Income (10,000s)	-5.5	1.8	1.7	-5.9	1.9	1.7
	Race (Black)	14.9	3.9	4.8	15.4	3.9	4.8
	Race (Nat. Amer.)	44.6	12.5	12.8	43.5	12.8	12.8
	Race (Other)	25.4	5.8	5.8	25.3	5.9	5.9
	Region (Midwest)	-13.8	5.3	5.4	-13.4	5.4	5.3
	Region (South)	-3.1	4.5	4.4	-3.0	4.6	4.4
	Region (West)	-3.3	4.9	4.9	-2.6	5.0	4.9
Hyperlipidemia	Intercept	-400.3	15.3	14.6	-400.4	15.3	14.7
	BMI	3.4	0.2	0.2	3.5	0.2	0.2
	Age	4.8	0.1	0.1	4.8	0.1	0.1
	Sex	13.5	2.2	2.4	14	2.2	2.4
	Education	0.9	0.4	0.4	0.8	0.4	0.4
	Income (10,000s)	1.2	1.3	1.3	1.2	1.3	1.3
	Race (Black)	-13.0	3.1	3.4	-12.7	3.0	3.4
	Race (Nat. Amer.)	13.1	11.0	10.7	13.2	11.3	10.7
	Race (Other)	15.6	4.1	4.5	15.7	4.1	4.5
	Region (Midwest)	-7.9	3.8	3.6	-8.2	3.8	3.6
	Region (South)	0.9	3.3	3.2	0.8	3.3	3.2
	Region (West)	-7.3	3.6	3.8	-7.3	3.6	3.8
Hypertension	Intercept	-351.1	15.1	15.6	-350	14.8	15.6
	BMI	5.7	0.2	0.2	5.7	0.2	0.2
	Age	4.8	0.1	0.1	4.8	0.1	0.1
	Sex	11.7	2.3	2.3	12.1	2.3	2.3
	Education	-0.6	0.4	0.4	-0.7	0.4	0.4
	Income (10,000s)	-8.5	1.3	1.2	-8.5	1.3	1.2
	Race (Black)	27.9	2.9	2.8	27.8	2.9	2.8
	Race (Nat. Amer.)	25.9	10.8	11.1	26.3	10.5	11.2
	Race (Other)	15.1	4.3	4.1	15.0	4.4	4.1
	Region (Midwest)	-7.2	3.9	4.0	-7.5	3.9	4.1
	Region (South)	3.1	3.4	3.2	2.9	3.3	3.3
	Region (West)	-9.2	3.7	3.5	-9.3	3.7	3.5
Diabetes	Hyperlipidemia	0.41	0.04	0.02	0.41	0.02	0.02
Diabetes	Hypertension	0.35	0.03	0.02	0.35	0.02	0.02
Hyperlipidemia	Hypertension	0.41	0.03	0.02	0.41	0.01	0.02

3.4 Simulations

3.4.1 Run-Time Comparisons with the Chained Bivariate Probit

Simulations were performed following the set-up described by Mullahy (2016), who found that the chained bivariate probit approach was much faster than a simulation-based maximum likelihood approach. The number of components considered was 4 and 8, the number of predictors considered was 5 and 9 (including intercept), and the number of observations varied among 2,000, 5,000, and 10,000. 100 simulations were run for each of these 12 combinations. For coefficients, the intercepts ranged step-wise from -0.2 to -1.6 from the first component to the eighth, assuming the values 0.0 or 0.5. Other coefficient parameters were set to 0.0. When only four components were needed for a given combination of settings, the first four components were used. For our analysis, design matrices were constructed using an intercept and the first 3 or 7 principal component values from a genetic dataset consisting of ternary data, obtained by subsampling a random set of genetic markers from HapMap data (Gibbs et al., 2003), and one randomly chosen of the ternary values.

For each combination of number of observations, number of components, and number of predictors, a design matrix was sampled (with replacement) from the original 728 observations. The multivariate response was then simulated anew for each of the 100 simulations per combination. Run-time comparisons were made versus reported values from Mullahy (2016), in which the previous author used a desktop computer with a higher clock speed (iMac 3.4 GHz Intel Core i7) than used here.

Table 3.3: Mean run-time comparisons between Mullahy (2016) & the Two-Stage Composite Likelihood, by combinations of number of observations, number of components, and number of predictors. Run-times are rounded to the nearest second; each combination was simulated 100 times. Mullahy's results were performed on an iMac 3.4 GHz Intel Core i7. The re-coded chained bivariate probit and the Two-Stage CL were ran on a Windows 2.70 GHz Intel i7-7500. The re-coded chained bivariate probit used R's `zelig` (Choirat et al., 2018; Imai et al., 2008) from the `Zelig` package to perform the bivariate probits.

# Obs	# Comp	# Pred	Mullahy (2016)	Re-coded `bvpmvp`	Two-Stage CL
2,000	4	5	1	1	0
2,000	4	9	1	2	0
2,000	8	5	5	7	1
2,000	8	9	8	7	1
10,000	4	5	2	6	1
10,000	4	9	3	6	1
10,000	8	5	14	28	5
10,000	8	9	19	41	5
50,000	4	5	12	28	6
50,000	4	9	18	30	6
50,000	8	5	65	136	23
50,000	8	9	86	145	23

For each combination of settings, the two-stage composite likelihood was the fastest. Mullahy's reported results were faster than the re-coded version of the chained bivariate probit, possibly at least partly due to processor specifications. As also observed by Mullahy, the number of components had the greatest effect on run-time, followed by number of observations, and the number of predictors had the least effect.

3.4.2 Coverage Percentages for the Two-Stage Composite Likelihood

Using a similar set-up (and the same original data) as for the run-time comparisons, simulations were performed to gauge the 95% coverage percentages for the two-stage composite likelihood. The number of predictors was fixed at four, and the number of observations and components were varied for the simulations. 500 simulations of each combination of observations and number of components were run. The 95% coverage percentages for the coefficients are displayed below. Coverage is near the target 95% in all instances.

Table 3.4: 95% Coverage Percentages for the Two-Stage Composite Likelihood, using 500 simulations. The number of predictors was fixed at four, with the number of observations and components varied. $B_{m,p}$ corresponds to the coefficient associated with the m'th component and p'th predictor.

Obs	Comp	$B_{1,1}$	$B_{1,2}$	$B_{1,3}$	$B_{1,4}$	$B_{2,1}$	$B_{2,2}$	$B_{2,3}$	$B_{2,4}$	$B_{3,1}$	$B_{3,2}$	$B_{3,3}$	$B_{3,4}$	$B_{4,1}$	$B_{4,2}$	$B_{4,3}$	$B_{4,4}$	$B_{5,1}$	$B_{5,2}$	$B_{5,3}$	$B_{5,4}$
200	3	96.8	95.2	96.0	93.2	96.2	93.6	96.0	96.2	96.2	94.8	95.6	96.2								
200	4	94.8	93.4	94.8	94.8	95.2	94.8	95.4	94.0	96.0	95.6	96.4	94.8	94.4	94.4	95.4	96.2				
200	5	94.4	94.4	94.8	95.2	96.0	94.8	94.8	93.8	94.6	96.4	96.4	93.4	93.8	96.2	96.0	96.2	95.2	96.0	94.8	95.2
400	3	95.0	95.4	94.2	95.8	94.6	94.4	92.4	94.8	96.2	94.0	93.2	94.6								
400	4	95.2	95.6	95.2	95.6	94.4	97.4	94.6	95.4	96.6	94.8	96.6	95.0	94.6	95.2	95.2	95.4				
400	5	95.8	96.0	96.4	95.0	94.0	95.2	94.8	94.6	94.8	96.6	95.6	95.4	94.8	94.8	95.4	96.2	93.6	96.8	94.4	96.0
800	3	96.0	94.0	94.2	95.8	95.0	96.4	95.6	95.4	94.8	94.4	94.6	95.4								
800	4	95.6	94.8	96.0	95.4	96.2	94.0	93.6	96.8	94.6	96.2	97.6	94.0	95.2	93.2	96.0	96.0				
800	5	95.8	96.4	93.0	94.6	95.0	95.0	95.4	94.8	96.0	95.4	96.2	94.0	96.2	94.0	95.8	95.2	94.4	94.6	95.0	94.6
1,600	3	96.4	96.4	94.8	95.4	96.4	94.0	93.2	96.2	96.0	93.6	94.4	96.2								
1,600	4	95.8	94.4	94.2	95.8	95.0	96.0	95.6	95.4	93.0	95.0	96.2	96.0	96.2	95.6	95.0	95.6				
1,600	5	94.4	94.0	95.4	96.0	94.0	96.2	95.6	94.2	92.8	96.0	96.4	95.6	93.6	95.2	92.8	95.4	94.2	96.0	94.8	96.6
3,200	3	94.0	94.8	95.4	94.2	92.6	95.2	95.4	94.4	94.8	94.4	93.4	94.6								
3,200	4	94.4	95.2	96.6	95.4	96.6	95.2	93.4	95.4	95.8	94.0	95.4	95.4	93.8	94.2	94.4	95.0				
3,200	5	95.4	94.6	94.4	95.0	93.2	94.8	94.8	93.8	96.8	95.8	94.0	96.0	94.0	94.0	93.8	96.4	94.6	93.6	93.8	94.8
6,400	3	94.6	94.4	96.0	96.6	95.0	94.4	94.4	95.8	93.0	94.0	97.0	92.2								
6,400	4	93.6	95.8	92.2	96.4	95.8	93.8	95.2	94.4	95.2	95.2	96.8	95.0	96.8	96.0	96.4	95.0				
6,400	5	95.6	95.4	97.0	94.2	92.8	96.6	94.0	94.8	94.4	92.8	95.8	93.4	96.0	95.6	95.4	94.2	96.2	94.8	95.2	95.2

3.5 Discussion

Our proposed two-stage composite likelihood for the multivariate probit produces results similar to those previously published, for both parameter estimations and standard errors. Bootstrap comparisons show that the robust variance estimates provide accurate standard errors. Furthermore, these standard errors are comparable to those of full likelihood or those of alternate methods, suggesting little loss in statistical efficiency.

Run-times for the two-stage composite likelihood compare favorably to the chained bivariate probit approach, which was already much faster than the approach using simulated maximum likelihood. The effects of increasing settings such as the number of observations, number of components, and number of predictors has similar effects to that experienced by the chained bivariate probit. Under simulation, our approach produces near nominal confidence coverage.

A possible next step would be to extend this approach to incorporate heterogeneous multivariate responses, i.e., where the response can include both binary and continuous components. Such an approach would include bivariate normal densities for continuous-continuous pairs, and likelihoods for binary-continuous pairs. For binary-continuous pairs, the joint likelihood can be re-stated as the product of the marginal density of the continuous component multiplied against the conditional density of the binary component upon the continuous component.

Further considerations could include heteroskedasticity, i.e. non-constant variance across predictor values, which would further expand the range of applications. In genomics, for example, this could manifest in different variances depending on allele counts. In fields such as finance and/or epidemiology, this could entail time-dependent variances, perhaps expressed through a time-series model.

Along similar lines, one avenue would be to explore more “fat-tailed” latent distributions. Normal distributions may be inappropriate when the response has a greater propensity for extreme outcomes, especially combined with interdependence among response components

(Ly et al., 2019). Insufficient attention to tail outcomes has been an issue in finance for price and risk modeling for asset-backed securities, and has been implicated as a contributor to the 2007-2008 financial crisis (Salmon, 2012; Flavin and Sheenan, 2015; MacKenzie and Spears, 2014). This could have implications for human phenotypes, as well, such as cognitive ability (Lubinski, 2009).

4

Simultaneous Modeling of Multivariate Heterogeneous Responses & Heteroskedasticity via a Two-Stage Composite Likelihood

4.1 Introduction

Historically, most GWAS are primarily focused only on univariate responses, i.e. one phenotype at a time. Simultaneous consideration of phenotypes in a multivariate framework could be insightful, as there may be correlated errors across phenotypes. Analysis of multivariate phenotypes together may confer improvements in statistical power (Park et al., 2011; Salinas et al., 2018) and more fully leverage genomic data sets (Park et al., 2011). Furthermore, observational data for these multivariate phenotypes may be heterogeneous—e.g. binary or continuous. As previously discussed, multivariate binary choice models can already be difficult to estimate in themselves, much less with the addition of continuous responses.

Typically, GWAS test mean effects associated with locus allele counts. However, loci could be associated with variance changes, as well—e.g. heteroskedasticity across groups, where each group is composed of individuals with allele counts of either 0, 1, or 2. Thus, a test for loci associated with variance changes would be needed and incorporated with a test for mean changes into an overall framework. Properly characterizing heteroskedasticity could be helpful in detecting significant SNP effects in both mean and variance.

Modeling of multivariate phenotypes could have particularly interesting tie-ins with pleiotropy (Hackinger and Zeggini, 2017; Yang and Wang, 2012; Salinas et al., 2018; Park et al., 2011; XiaoCan et al., 2019), where a locus affects multiple phenotypes (Salinas et al., 2018). If a SNP is associated with mean changes in more than one phenotype, some of the overall inter-phenotype covariance could potentially be accounted for by that SNP. Better understanding of pleiotropy could aid in revisiting phenotype definitions and disease diagnoses (Hackinger and Zeggini, 2017), and/or identify loci as target intervention candidates for multiple diseases

simultaneously.

As an example, psychiatric traits have received considerable attention from clinicians and researchers with regard to pleiotropy. Human cognitive traits in general have long been known to have substantial genetic influences (Bouchard, 2004), including many psychiatric disorders (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013; XiaoCan et al., 2019; Demkow and Wolańczyk, 2017). Yet, characterization and diagnoses of psychiatric disorders remain a challenge—much less understanding their etiology and designing effective interventions (Demkow and Wolańczyk, 2017). Previous work has indeed confirmed the high heritability of major psychiatric disorders, and furthermore, has suggested that material pleiotropy may be present in the genetic architecture of such disorders (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013; XiaoCan et al., 2019).

Pleiotropy could very well be "ubiquitous" in human genetics (Visscher et al., 2017). Widespread pleiotropy in the human genome would have profound impacts on research in precision medicine and genome editing, in particular upon drug development and genome "fixing" via CRISPR-Cas approaches (Gratten and Visscher, 2016).

4.1.1 Approaches to Multivariate Phenotypes

Dimension-reduction techniques can re-express multivariate phenotypes more parsimoniously, e.g. into a univariate response. Methods such as Principal Components of Heritability Association Tests (PCHAT) do this through PCA to create independent components in a newly-formed multivariate response (Klei et al., 2008). In PCHAT, the dimension-reduction in the multivariate phenotype is performed to maximize the heritability. However, a bagging-procedure is needed to algorithmically form these principal components. The distribution of resulting test-statistics can be awkward, necessitating permutation-testing to better characterize the distributions for hypothesis-testing. By reducing a multivariate phenotype into fewer dimensions, component-specific information is naturally lost.

Another tactic for multivariate phenotypes is through Canonical Correlation Analysis (CCA), such as that implemented by `MV-PLINK` (Galesloot et al., 2014). CCA aims to extract

the linear combinations of phenotypes that create the greatest covariance against SNP data (Galesloot et al., 2014), although there is some diversity in how CCA-based tests are conducted (Tang and Ferreira, 2012). The implicit assumption is that the multivariate response is multivariate normal, and thus does not account for heterogeneity in the multivariate components (O'Reilly et al., 2012).

mvBIMBAM is a multivariate extension of Bayesian Imputation-Based Association Mapping (BIMBAM) (Servin and Stephens, 2007). In mvBIMBAM, multivariate regression is performed through partitioning; genotypes are partitioned as direct, indirect, or no effect, corresponding to the inferred effect of a variant upon a given phenotype (Porter and O'Reilly, 2017). Shim et al. (2015) used mvBIMBAM to conduct a GWAS investigating response to statin treatment in 1,868 individuals of Caucasian descent, where the phenotypes were 7 subfractions of low density lipoproteins (LDLs) and 3 subfractions of intermediate density lipoproteins (IDLs). mvBIMBAM shows solid performance in statistical power (Porter and O'Reilly, 2017; Shim et al., 2015). mvBIMBAM is only tailored toward multivariate normal phenotypes (Stephens, 2013).

MultiPhen, with an eponymous R package, is an approach that"reverses" or "inverts" the usual GWAS formulation such that the genotype at a given locus is the response in a proportional-odds logistic regression, and the components of the multivariate phenotype serve as the predictors. Using the multivariate phenotype as the predictors, MultiPhen neatly accommodates any heterogeneity in a multivariate phenotype (O'Reilly et al., 2012).

Following O'Reilly et al. (2012), we begin by reviewing a univariate GWAS for a multivariate phenotype:

$$Y_{ik} = \alpha_k + \beta_{gk} X_{ig} + \epsilon_{igk}$$

i denotes a given observation, k a component of the multivariate phenotype, and g a given SNP. α_k would then be the intercept for a given component, β_{gk} the effect size for a given SNP upon a given component, and X_{ig} the allele count (0, 1, or 2). ϵ_{igk} is a normally-distributed error.

Let us now invert the GWAS set-up into a proportional-odds logistic regression:

$$P(X_{ig} \leq m) = \frac{1}{1 + \exp(-\alpha_{gm} - \sum_{k=1}^{K} \beta_{gk} Y_{ik})}$$

Where m in $\{0, 1, 2\}$. A likelihood ratio test can then be performed, as to test the null hypothesis that $\beta_{g1} = ... = \beta_{gK} = 0$.

This approach performs well, computationally and statistically-speaking. However, it is unclear as to how additional covariates would be handled, and/or their interpretability. Addressing of heteroskedasticity also remains.

4.1.2 Approaches to Continuous and Binary Responses

Teixeira-Pinto and Normand (2009) reviewed some approaches for handling bivariate responses with one continuous component, and one bivariate. One of which is a "factorization model" suggested by Fitzmaurice and Laird (1995), conditioning the continuous outcome upon the binary outcome (Teixeira-Pinto and Normand, 2009).

Suppose we have a binary component y_b and a continuous component y_c. Then we can write $f(y_b, y_c) = f(y_b)f(y_c|y_b)$. $f(y_b)$ can then be the usual univariate probit, and y_{ci} has conditional normal distribution:

$$\phi(\boldsymbol{x_{ci}^T \beta_c} + \tau(y_{bi} - \mu_{bi}), \sigma_c^2)$$

Where $\mu_{bi} = \boldsymbol{x_{bi}^T \beta_b}$ is the expected value of y_{bi} and σ_c^2 is the variance of the continuous component. τ is obtained by regressing y_{ci} upon y_{bi}. Ceteris paribus, larger absolute values of τ suggest a stronger association between the two components (Teixeira-Pinto and Normand, 2009). The correlation parameter, ρ, can then be obtained as:

$$\rho = \begin{cases} \frac{sign(\tau)}{\sqrt{1 + \frac{\sigma_c^2}{\tau^2 Var(y_{bi}|\boldsymbol{x_{bi}})}}} & \tau \neq 0 \\ 0 & \tau = 0 \end{cases}$$

Maximum likelihood can then be used to estimate the parameters. Teixeira-Pinto and Normand (2009) remarked that a drawback is that the marginal distribution of the continuous component is then an implied mixture of two normal distributions, even potentially bimodal. Conditioning the binary upon the continuous is also mentioned, as an alternative: $f(y_b, y_c) = f(y_b|y_c)f(y_c)$. Now the continuous component can retain its univariate marginal distribution, and the binary component has the latent distribution:

$$\phi(\boldsymbol{x_{bi}^T}\boldsymbol{\beta_b} + \tau^{'}(y_{bi} - \mu_{ci}), 1)$$

Where $\mu_{ci} = \boldsymbol{x_{ci}^T}\boldsymbol{\beta_c}$ and $\tau^{'}$ is obtained by regressing y_{bi} upon y_{ci}.

de Leon and Wu (2011) proposed a copula-based approach for handling bivariate responses where one component is binary, and one is continuous. It remains to be seen to how this could be extended to settings where there could be more than one binary and/or continuous component.

4.1.3 Approaches to Heteroskedasticity

For detecting quantitative trait loci (QTL) affecting the variance of a univariate phenotype, Rönnegård and Valdar (2011) proposed a double GLM approach using REML to perform the estimation, where a QTL can affect both the mean and variance of a phenotype. They mention applications such as predictability of response to medicinal dosages, and selecting uniform individuals from livestock.

Young et al. (2018) suggested a log-linear model for a potential SNP effect on variance for a quantitative phenotype. Following Young et al. (2018), suppose we have $\sigma_g^2 = Var(Y|G = g)$ as the conditional variance of the phenotype given genotype, where $g \in \{0, 1, 2\}$. Then we set $\ln(\sigma_g^2) = \mu_v + \alpha_v g$, where μ_v is akin to an intercept, and α_v a coefficient upon g for log-variance. Then along with an additive effect upon the mean, the additive variance model would be:

$$Y|G = g \sim \phi\big(\mu + \alpha g, \exp\left(\mu_v + \alpha_v g\right)\big)$$

For more complicated effects on variance, they also provide a general variance model, where now:

$$\ln\left(\sigma_g^2\right) = \mu_v + \alpha_v g + \delta_v\left(g - 2f\right)^2\right)$$

And f is the allele frequency. For a log-linear variance effect, they note that there is a 1 degree of freedom test—and for the log-linear and general variance effect, a 2 degree of freedom test. The latter, similar to a Bartlett's Test, was also proposed by Cao et al. (2014) as a simultaneous test for mean and variance (Young et al., 2018). Young et al. (2018) extend this approach to account for non-normality in a phenotype, and apply it to body mass index among the approximately 408,000 individuals with British ancestry in the UK Biobank data.

Bartlett's Test is a common test for equality of variance between groups, i.e. if there is any heteroskedasticity across groups (Snedecor and Cochran, 1989). The multivariate Bartlett's Test, as briefly described by Schott (2001), is the analogous multivariate version for equality of covariance matrices between groups. Assuming multivariate Y is normally distributed with m components, the null hypothesis is then:

$$H_0 : \Sigma_1 = ... = \Sigma_g = \Sigma$$

Where Σ_i is the covariance matrix of the i'th group of g total groups, and Σ is the common covariance matrix. Let $n_i + 1$ be the sample size of the i'th group, and thus $S_i = \frac{1}{n_i}\sum_{j=1}^{n_i+1}(x_{ij} - \bar{x_i})(x_{ij} - \bar{x_i})^t$ would be the usual unbiased sample covariance estimate of each group i, with each $\bar{x_i} = \frac{1}{n_i+1}\sum_{j=1}^{n_i+1} x_{ij}$ being the sample mean for that group. Naturally, $\sum_1^g n_i = n$ and the test statistic is then:

$$M = n \ln|S| - \sum_1^g n_i \ln|S_i|$$

Where $S = \frac{1}{n}\sum_1^g n_i S_i$. Under the null hypothesis, M is asymptotically χ^2 with $\frac{(g-1)(m)(m+1)}{2}$ degrees of freedom. The multivariate Bartlett's Test is apparently also sometimes referred to as Box's M Test (D'Agostino and Russell, 2005). The Bartlett's Test is appealing for producing a straightforward test statistic. However, it is unclear to what extent this test can accommodate a multivariate response with at least some binary components. Although more general, this

test has been used in the context of finding variance QTLs, as well—but it may be underpowered if we are interested in heteroskedasticity of a particular type, e.g. in GWAS, modeling the covariance as a linear function of an SNP-value (0, 1, or 2).

Zhu et al. (2017) explored testing of covariance matrices of type $\boldsymbol{D} = \boldsymbol{\Sigma_2} - \boldsymbol{\Sigma_1}$ across 2 groups, with null hypothesis $H_0 : \boldsymbol{D} = \boldsymbol{0}$, with an application for gene expression data from post-mortem schizophrenia cases and controls. They were chiefly concerned with the high-dimensional situation, particularly where the number of components in the response exceeded the number of observations.

With concerns over loss of power using homoskedastic models in the presence of heteroskedasticity, Quintero and Lesaffre (2017) presented an extension of a model discussed by Li et al. in 2013, which in turn was an extension of a model from Hoff and Niu (2012). Hoff & Niu proposed a covariance regression model that parameterizes the multivariate response as a quadratic function of the predictors, reminiscent of a factor analysis model.

Quintero and Lesaffre (2017) performed simulation studies, and also applied the model to data on nurse burnout in Belgium. The components of the response were emotional exhaustion, depersonalization, and reduced personal accomplishment on an integer scale ranging from 0, 0, and 0 to 54, 30, and 48, respectively. They noted from their simulation studies that neglecting heteroskedasticity resulted in higher root mean squared errors.

As described by Quintero and Lesaffre (2017), the covariance regression model of Hoff and Niu (2012) reparameterized heteroskedasticity in the form of a mixed effects model. For a given observation i: $\boldsymbol{Y}_i = \boldsymbol{A}\boldsymbol{x}_i + \gamma_i \boldsymbol{B}\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$, where $\boldsymbol{Y}$ is the multivariate response, $\boldsymbol{A}$ and $\boldsymbol{B}$ are matrices corresponding respectively to coefficients and factor loadings, $\boldsymbol{x_i}$ contains covariate values for the i'th observation, and γ_i is a latent standard normal variable that serves as a scalar for additional variation in the response. This yields a covariance matrix that is a quadratic function of the covariates: $\boldsymbol{\Sigma}(\boldsymbol{x}_i) = \boldsymbol{\Psi} + \boldsymbol{B}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{B}^T$, with $\boldsymbol{\Psi}$ being the covariance matrix of $\boldsymbol{\epsilon}_i$.

This set-up lends itself well to maximum likelihood estimation using approaches such as Expectation Maximization or Gibbs Sampling, as it is represented in a familiar form (Quin-

tero and Lesaffre, 2017). However, due to the nature of the scalar, embedded is the assumption that "extreme residuals for one component of $\boldsymbol{Y}_i$ are accompanied with extreme residuals of the other elements in the vector response" (Quintero and Lesaffre, 2017).

4.1.4 Simultaneous Modeling of Multivariate Heterogeneous Responses & Heteroskedasticity

We present a straightforward, unified framework by which to perform two-stage composite likelihood estimations for regressions with multivariate responses that include both binary and continuous components, and heteroskedasticity where the covariance matrix depends on choice of a predictor—i.e. the covariance is group-dependent. We leverage the lessons-learned and knowhow from previous work done on multivariate probit estimations, and extend it to incorporate heterogeneity (multivariate responses with both binary and continuous components) and heteroskedasticity. In performing estimations under this framework, we can use common functions such as `optim` and `optimize` in R for performing optimizations, and `pbivnorm` for fast evaluations of bivariate probit probabilities (Kenkel, 2015).

Such a framework lends itself well to hypothesis testing in genome-wide association studies—testing, in a flexible manner, for associations between loci and the mean and/or covariance of the phenotype. Although we are primarily concerned with the GWAS situation with three groups (SNP values of 0, 1, or 2), an arbitrary number of groups can be included–albeit we expect the number of observations to be greater than that of the number of components.

As with the multivariate probit, the set-up of the Two-Stage Composite Likelihood engenders a divide-and-conquer approach. This breaks-up an otherwise formidable likelihood into more tractable associated likelihoods—allowing more favorable computational performance—while retaining useful asymptotic properties for hypothesis testing. With modern genomic data-sets having thousands or more of observations, we have in mind data-sets with sample sizes that number in the multiple thousands at the minimum.

4.2 Methods

4.2.1 The Multivariate Probit Likelihood

Let us begin by recalling the multivariate probit likelihood, as our approach for multivariate heterogeneous responses with heteroskedasticity is essentially a generalization of our approach for the multivariate probit. For the i'th observation and R components, all of which binary, we have:

$$L(\boldsymbol{\theta}; \boldsymbol{y_i}) = \int_{A_{i1}} \dots \int_{A_{iR}} \phi(\boldsymbol{z_i}, \mathbf{0}, \boldsymbol{V}) \boldsymbol{dz_i}$$

$$A_{ir} = \begin{cases} (-\infty, \boldsymbol{x_i \beta_r}] & y_{ir} = 1 \\ (\boldsymbol{x_i \beta_r}, \infty) & y_{ir} = 0 \end{cases}$$

$$\boldsymbol{\theta} = \{\boldsymbol{B}, \boldsymbol{V}\}$$

$\boldsymbol{B}$ is a matrix of coefficients, $\boldsymbol{V}$ is the covariance matrix—which in this case, is a correlation matrix (since all component variances are set to 1, for identifiability purposes). $\boldsymbol{\beta_r}$ is a column of $\boldsymbol{B}$ that pertains to a particular component r. $\boldsymbol{X}$ is the design matrix, with a particular observation being vector $\boldsymbol{x_i}$. A_{ir} defines the boundaries of integration, depending on the cut-point determined by $\boldsymbol{x_i \beta_r}$ and whether y_{ir}, a given observation for a given component, is a 1 or 0.

4.2.2 The Multivariate Likelihood for Multivariate Heterogeneous Responses with Heteroskedasticity

With the formulation of the multivariate probit in mind, let us now extend the multivariate probit likelihood to incorporate heterogeneity and heteroskedasticity. For observation i we have:

$$L(\boldsymbol{\theta}; \boldsymbol{y_i}) = \int_{A_{i1}} \dots \int_{A_{iR}} \phi(\boldsymbol{z_i}, \boldsymbol{\mu_i}, \boldsymbol{V_i}) \boldsymbol{dz_i}$$

$$A_{ir} = \begin{cases} (-\infty, \boldsymbol{x_i\beta_r}] & y_{ir} = 1 \\ (\boldsymbol{x_i\beta_r}, \infty) & y_{ir} = 0 \end{cases}$$

$$\mu_{im} = \begin{cases} 0 & y_{im} \text{ binary} \\ \boldsymbol{x_i\beta_m} & y_{im} \text{ continuous} \end{cases}$$

$$\boldsymbol{V_i} = \begin{cases} \boldsymbol{\Sigma} & x_{iP} = 0 \\ \boldsymbol{T} & x_{iP} = 1 \\ \boldsymbol{\Omega} & x_{iP} = 2 \end{cases}$$

$$\boldsymbol{\theta} = \{\boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{T}, \boldsymbol{\Omega}\}$$

Where P is the number of predictors; the P'th predictor would correspond to a given candidate SNP in a GWAS setting. $\phi()$ is the normal density function. Similar to the multivariate probit, the area of integration, A_{ir} for a given observation is determined by $\boldsymbol{x_i\beta_r}$ and whether the response for that component, y_{ir}, is 1 or 0. The mean remains 0 for binary components; the vector of means for a given observation, $\boldsymbol{\mu_i}$ is zero for binary components, but potentially non-zero for continuous components. Now instead of one covariance matrix $\boldsymbol{V}$, the covariance $\boldsymbol{V_i}$ is dependent upon the value of x_{iP}. Thus we have, essentially, three covariance matrices: $\boldsymbol{\Sigma}, \boldsymbol{T}, \boldsymbol{\Omega}$, for whether x_{iP} has a value of 0, 1, or 2, respectively.

With no loss of generalizability, we will arrange the first R components to be binary out of the M total components, leaving the remaining $M - R$ as continuous. Thus, for a given observation, the covariance matrix will take the form:

$$\boldsymbol{V_i} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ v_{i11} & v_{i12} & \dots & v_{iMM} \end{pmatrix}$$

4.2.3 Two-Stage Composite Likelihood for Multivariate Heterogeneous Responses with Heteroskedasticity

Recall that in the first stage of two-stage estimation, we estimate a subset of the parameters:

$$\ln L_1(\boldsymbol{\theta_1}|\boldsymbol{Y}) = \sum_{1}^{N} \ln f_1(\boldsymbol{y_{1i}}|\boldsymbol{x_{1i}}, \boldsymbol{\theta_1})$$

Then in the second stage, the estimates of θ_1 from the first stage can be used as fixed inputs to maximize the conditional likelihood:

$$\ln L_2(\boldsymbol{\theta_2}|\hat{\boldsymbol{\theta_1}}, \boldsymbol{Y}) = \sum_{1}^{N} \ln f_2(\boldsymbol{y_{2i}}|\boldsymbol{x_{2i}}, \boldsymbol{\theta_2}, (\boldsymbol{x_{1i}}, \hat{\boldsymbol{\theta_1}}))$$

For the Two-Stage Composite Likelihood for Multivariate Heterogeneous Responses with Heteroskedasticity, our choice for the subset of parameters to estimate will be the coefficients and continuous variances, i.e.:

$$\boldsymbol{\theta_1} = \{\{\boldsymbol{B}\}, ..., \sigma_{(r+1)(r+1)}, ..., \tau_{mm}, ..., \omega_{mm}\}$$

Thus, in the second stage, we estimate the remaining parameters, i.e. the pairwise correlations and covariances:

$$\boldsymbol{\theta_2} = \{\sigma_{12}, ..., \sigma_{(r-1)r}, \sigma_{1(r+1)}, ..., \tau_{rm}, ..., \omega_{rm}, \sigma_{(r+1)m}, ..., \omega_{(m-1)m}\}$$

Let us now combine the Composite Likelihood approach with Two-Stage estimation, and consider a given observation, i. In the Composite Likelihood, we re-formulate a given likelihood into separate sub-likelihoods, or associated likelihoods. For example, with A associated likelihoods, we have:

$$\ln L_{comp}(\boldsymbol{\theta}|\boldsymbol{y_i}) = \sum_{a=1}^{A} \ln \boldsymbol{f_a}(\boldsymbol{y_i}|\boldsymbol{\theta})$$

For each of the stages, we will set-up a Composite Likelihood. That is, in the first stage we will

have:

$$L_{uni}(\boldsymbol{\theta_1}; \boldsymbol{y_i}, \boldsymbol{x_i})$$

For the second, the Composite Likelihood will be:

$$L_{pair}(\boldsymbol{\theta_2^*}; \boldsymbol{y_i}, \boldsymbol{x_i})$$

$\boldsymbol{\theta_2^*}$ contains both parameter estimates from the first stage, $\hat{\boldsymbol{\theta}}_\mathbf{1}$, as well as parameters to be estimated in the second stage, $\boldsymbol{\theta_2}$.

4.2.4 Setting-up and Estimating the First Stage

In the first stage of Two-Stage estimation we can set-up a Composite Likelihood with parameters involving only one component in each associated likelihood:

$$L_{uni}(\boldsymbol{\theta_1}; \boldsymbol{y_i}, \boldsymbol{x_i}) = \prod_{m=1}^{M} f(y_{im}, \boldsymbol{x_i}; \boldsymbol{\theta_{1m}})$$

Where we have M total components in the multivariate response, each with first stage parameter subset $\boldsymbol{\theta_{1m}}$, and $\boldsymbol{\theta_1} = \{\boldsymbol{\theta_{11}}, ..., \boldsymbol{\theta_{1M}}\}$. For binary components, $\boldsymbol{\theta_{1m}} = \boldsymbol{\beta_m}$. In contrast to binary components, where variances are fixed at 1, for continuous components we have $\boldsymbol{\theta_{1m}} = \{\boldsymbol{\beta_m}, \sigma_{mm}, \tau_{mm}, \omega_{mm}\}$. Each associated likelihood in the first stage involves a heteroskedastic univariate probit or normal density, depending on whether a given component is binary or continuous. Binary components have variances fixed at one, but continuous components have heteroskedastic variances to be estimated, depending on whether x_{iP} has a value of 0, 1, or 2.

For a binary component we have:

$$f(y_{im}, \boldsymbol{x_i}; \boldsymbol{\theta_{1m}}) = \int_{A_{im}} \phi(z_{im}, 0, 1) dz_{im}$$

$$A_{im} = \begin{cases} (-\infty, \boldsymbol{x_i}\boldsymbol{\beta_m}] & y_{im} = 1 \\ (\boldsymbol{x_i}\boldsymbol{\beta_m}, \infty) & y_{im} = 0 \end{cases}$$

The area of integration is determined by the i'th observation of the design matrix, the coefficients for the component, β_m, and the binary observation for that component, y_{im}. The mean is fixed at 0, and the variance at 1.

For a continuous component we have:

$$f(y_{im}, \boldsymbol{x_i}; \boldsymbol{\theta_{1m}}) = \phi(z_{im}, \mu_{im}, v_{imm})$$

$$v_{imm} = \begin{cases} \sigma_{mm} & x_{iP} = 0 \\ \tau_{mm} & x_{iP} = 1 \\ \omega_{mm} & x_{iP} = 2 \end{cases}$$

Analogous to the area of integration for binary components, the mean for continuous components is determined by the design matrix and coefficients, $\mu_{im} = \boldsymbol{x_i}\boldsymbol{\beta_m}$. As represented by v_{imm} for each observation, we now have heteroskedastic variances for each continuous component, which can take on values of $\{\sigma_{mm}, \tau_{mm}, \omega_{mm}\}$ depending on whether the observation corresponding to a given SNP has a value of 0, 1, or 2.

Since no parameters are shared across associated likelihoods, we can write:

$$max_{\boldsymbol{\theta_1}} \sum_{i=1}^{N} \ln L_{uni}(\boldsymbol{\theta_1}; \boldsymbol{y_i}, \boldsymbol{x_i}) =$$

$$max_{\boldsymbol{\theta_{11}}} \sum_{i=1}^{N} \ln L_{uni}(\boldsymbol{\theta_{11}}; \boldsymbol{y_i}, \boldsymbol{x_i}) + ... + max_{\boldsymbol{\theta_{1M}}} \sum_{i=1}^{N} \ln L_{uni}(\boldsymbol{\theta_{1M}}; \boldsymbol{y_i}, \boldsymbol{x_i})$$

Subsequently, we can perform the optimization for each associated likelihood, $max_{\boldsymbol{\theta_{1m}}} \sum_{i=1}^{N} \ln L_{uni}(\boldsymbol{\theta_{1m}}; \boldsymbol{y_i}, \boldsymbol{x_i})$, independently.

4.2.5 Setting-up and Estimating the Second Stage

In the second stage we can similarly set-up a Composite Likelihood involving only two components, i.e. a pair, in each associated likelihood:

$$L_{pair}(\boldsymbol{\theta_2^*}; \boldsymbol{y_i}, \boldsymbol{x_i}) = \prod_{j=1}^{M-1} \prod_{k=j+1}^{M} f\left(y_{ij}, y_{ik}, \boldsymbol{x_i}; \boldsymbol{\theta_{2jk}^*}\right)$$

Where $\boldsymbol{\theta_2} = \{\boldsymbol{\theta_{2(1)(1)}}, ..., \boldsymbol{\theta_{2(M-1)(M)}}\}$. Each $\boldsymbol{\theta_{2jk}^*}$ contains the parameters to be estimated, $\boldsymbol{\theta_{2jk}}$, and parameters inherited from the first stage. The coefficients for each component $\boldsymbol{\beta_m}$ are inherited from the first stage, as are the heteroskedastic variances for continuous components. Thus, the parameters inherited for a given pair depends on the whether the pair consists of two binary components, one binary and one continuous, or two continuous components:

$$\boldsymbol{\theta_{2jk}^*} = \begin{cases} \{\sigma_{jk}, \tau_{jk}, \omega_{jk}, \hat{\boldsymbol{\beta}}_{\boldsymbol{j}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{k}}\} & \text{both binary} \\ \{\sigma_{jk}, \tau_{jk}, \omega_{jk}, \hat{\sigma}_{kk}, \hat{\tau}_{kk}, \hat{\omega}_{kk}, \hat{\boldsymbol{\beta}}_{\boldsymbol{j}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{k}}\} & j \text{ is binary, } k \text{ is continuous} \\ \{\sigma_{jk}, \tau_{jk}, \omega_{jk}, \hat{\sigma}_{jj}, \hat{\tau}_{jj}, \hat{\omega}_{jj}, \hat{\sigma}_{kk}, \hat{\tau}_{kk}, \hat{\omega}_{kk}, \hat{\boldsymbol{\beta}}_{\boldsymbol{j}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{k}}\} & \text{both continuous} \end{cases}$$

Each associated likelihood in the second stage involves a heteroskedastic bivariate probit, mixed binary/normal density, or bivariate normal density.

4.2.5.1 Associated Likelihood for a Binary-Binary Pair

For two binary components j and k we have:

$$f(y_{im}, \boldsymbol{x_i}; \boldsymbol{\theta_{2jk}^*}) = \int_{A_{ij}} \int_{A_{ik}} \phi(\boldsymbol{z_i}, \boldsymbol{0}, \boldsymbol{V_{ijk}}) \boldsymbol{dz_i}$$

$$A_{ij} = \begin{cases} (-\infty, \boldsymbol{x_i}\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}] & y_{ij} = 1 \\ (\boldsymbol{x_i}\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}, \infty) & y_{ij} = 0 \end{cases}$$

$$A_{ik} = \begin{cases} (-\infty, \boldsymbol{x_i}\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}] & y_{ik} = 1 \\ (\boldsymbol{x_i}\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}, \infty) & y_{ik} = 0 \end{cases}$$

$$\boldsymbol{V_{ijk}} = \begin{pmatrix} 1 & v_{ijk} \\ v_{ijk} & 1 \end{pmatrix}$$

$$v_{ijk} = \begin{cases} \sigma_{jk} & x_{iP} = 0 \\ \tau_{jk} & x_{iP} = 1 \\ \omega_{jk} & x_{iP} = 2 \end{cases}$$

The areas of integration for a particular observation i of the two components are determined by the observation of design matrix, their coefficient estimates (inherited from the first stage), and the observed binary responses. The mean for both components is fixed at 0. The variances for the two components are both fixed at one, leaving their heteroskedastic pairwise correlations $\{\sigma_{jk}, \tau_{jk}, \omega_{jk}\}$ to be estimated.

4.2.5.2 Associated Likelihood for a Binary-Continuous Pair

Suppose we have j binary and k continuous:

$$f(y_{im}, \boldsymbol{x_i}; \boldsymbol{\theta^*_{2jk}}) = \int_{A_{ij}} \phi(z_i, \hat{\boldsymbol{\mu}}_{\boldsymbol{ijk}}, \boldsymbol{V_{ijk}}) dz_i$$

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{ijk}} = \{0, \boldsymbol{x_i}\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}\}$$

$$A_{ij} = \begin{cases} (-\infty, \boldsymbol{x_i}\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}] & y_{ij} = 1 \\ (\boldsymbol{x_i}\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}, \infty) & y_{ij} = 0 \end{cases}$$

$$\boldsymbol{V_{ijk}} = \begin{pmatrix} 1 & v_{ijk} \\ v_{ijk} & \hat{v}_{ikk} \end{pmatrix}$$

$$= \begin{cases} \boldsymbol{\Sigma_{jk}} & x_{iP} = 0 \\ \boldsymbol{T_{jk}} & x_{iP} = 1 \\ \boldsymbol{\Omega_{jk}} & x_{iP} = 2 \end{cases}$$

Now, in contrast to a pair of both binary components, there is now just one integral (correspondingly to the binary component), instead of two. The mean of the continuous component is $\boldsymbol{x_i}\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}$, rather than fixed at 0. The continuous component also now inherits a variance estimate, $\hat{v}_{ikk}$, from the first stage, instead of being fixed at 1. Similarly, we estimate the heteroskedastic pairwise covariances $\{\sigma_{jk}, \tau_{jk}, \omega_{jk}\}$. The associated likelihood for a binary-continuous pair can also be written as the product of a conditional heteroskedastic univariate probit and a heteroskedastic univariate normal density, as observed by Teixeira-Pinto and Normand (2009) for the non-heteroskedastic case.

4.2.5.3 Associated Likelihood for a Continuous-Continuous Pair

When both components are continuous:

$$f(y_{im}, \boldsymbol{x_i}; \boldsymbol{\theta^*_{2jk}}) = \phi(\boldsymbol{x_i}\hat{\boldsymbol{B}}_{\boldsymbol{(jk)}}, \boldsymbol{V_{ijk}})$$

$$\hat{\boldsymbol{B}}_{\boldsymbol{(jk)}} = \{\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{j}}\}$$

$$\boldsymbol{V_{ijk}} = \begin{pmatrix} \hat{v}_{ijj} & v_{ijk} \\ v_{ijk} & \hat{v}_{ikk} \end{pmatrix}$$

$$= \begin{cases} \boldsymbol{\Sigma_{jk}} & x_{iP} = 0 \\ \boldsymbol{T_{jk}} & x_{iP} = 1 \\ \boldsymbol{\Omega_{jk}} & x_{iP} = 2 \end{cases}$$

Previously we saw changes to the handling of component k in moving from binary k to continuous k, i.e. from a binary-binary pair to a binary-continuous pair. These can be similarly seen in moving from binary j to continuous j, i.e. from a binary-continuous pair to a continuous-

continuous pair.

Since no parameters are shared across associated likelihoods, analogous to the first stage we can write:

$$max_{\boldsymbol{\theta_2}} \sum_{i=1}^{N} \ln L_{pair}(\boldsymbol{\theta_2^*}; \boldsymbol{y_i}, \boldsymbol{x_i}) =$$

$$max_{\boldsymbol{\theta_{2(1)(1)}}} \sum_{i=1}^{N} \ln L_{pair}(\boldsymbol{\theta^*_{2(1)(1)}}; \boldsymbol{y_i}, \boldsymbol{x_i}) + ... + max_{\boldsymbol{\theta_{2(M-1)(M)}}} \sum_{i=1}^{N} \ln L_{pair}(\boldsymbol{\theta^*_{2(M-1)(M)}}; \boldsymbol{y_i}, \boldsymbol{x_i})$$

As in the first stage, we can perform the optimization for each associated likelihood, $max_{\boldsymbol{\theta_{2jk}}} \sum_{i=1}^{N} \ln L_{pair}(\boldsymbol{\theta^*_{2jk}}; \boldsymbol{y_i}, \boldsymbol{x_i})$, independently.

Note that this divide-and-conquer strategy allows us to estimate each associated likelihood independently. Independent estimation of associated likelihoods allows us to achieve much better computational performance than would be attainable under estimating the full likelihood.

4.2.6 The Robust Murphy-Topel Estimator of Variance for the Two-Stage Composite Likelihood, Revisited

By construction, the Two-Stage Composite Likelihood applies a misspecified likelihood in estimation. We will apply the same equations from Hardin (2002) for robust variance estimation as we did for Two-Stage Composite likelihood, extending the approach as applied to the multivariate probit. Recall that we have $\boldsymbol{V_S}(\boldsymbol{\theta_1})$ as the robust variance estimator for the parameters estimated in the first stage of the two-stage model, and $\boldsymbol{V_S}(\boldsymbol{\theta_2})$ the second stage, with $\boldsymbol{Cov_S}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ being the covariance between them.

$$\begin{aligned}
\boldsymbol{V_S}(\boldsymbol{\theta_1}) &= \boldsymbol{V_1}\boldsymbol{V_1^{*-1}}\boldsymbol{V_1} = \boldsymbol{V_{S1}} \\
\boldsymbol{Cov_S}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) &= \boldsymbol{V_1}\boldsymbol{R^T}\boldsymbol{V_2} - \boldsymbol{V_{S1}}\boldsymbol{C^{*T}}\boldsymbol{V_2} \\
\boldsymbol{V_S}(\boldsymbol{\theta_2}) &= \boldsymbol{V_2}\boldsymbol{V_2^{*-1}}\boldsymbol{V_2} + \boldsymbol{V_2}(\boldsymbol{C^*}\boldsymbol{V_{S1}}\boldsymbol{C^{*T}} - \boldsymbol{R}\boldsymbol{V_1}\boldsymbol{C^{*T}} - \boldsymbol{C^*}\boldsymbol{V_1}\boldsymbol{R^T})\boldsymbol{V_2} \\
&= \boldsymbol{V_{S2}} + \boldsymbol{V_2}(\boldsymbol{C^*}\boldsymbol{V_{s1}}\boldsymbol{C^{*T}} - \boldsymbol{R}\boldsymbol{V_1}\boldsymbol{C^{*T}} - \boldsymbol{C^*}\boldsymbol{V_1}\boldsymbol{R^T})\boldsymbol{V_2}
\end{aligned}$$

Then we have:

$$\boldsymbol{V_S(\theta)} = \begin{pmatrix} \boldsymbol{V_S(\theta_1)} & \boldsymbol{Cov_S(\theta_1, \theta_2)} \\ \boldsymbol{Cov_S(\theta_1, \theta_2)} & \boldsymbol{V_S(\theta_2)} \end{pmatrix}$$

4.2.7 The Wald Test for the Two-Stage Composite Likelihood

Given the ability to test for both mean changes and variance changes, a natural desired outcome would be to simultaneously test both mean and variance changes in conjunction. In a GWAS setting, this would be applied to loci identifying candidates for further downstream analyses.

Using the robust Murphy-Topel variance estimate as per Hardin (2002) and with the findings of Cattelan and Sartori (2016) in mind, we present a Wald Test for Genome Wide Association Studies for simultaneous testing of both mean and variance changes. Suppose we wish to apply the hypothesis test to some subset of our parameters, say $\boldsymbol{\theta_d}$, using H by Q matrix $\boldsymbol{D}$, i.e. $H_0 : \boldsymbol{D\theta_d} = \boldsymbol{0}$. H corresponds to the number of parameters of interest, and Q the number of hypotheses to be tested.

Let $\boldsymbol{\delta} = \boldsymbol{D\theta_d}$. Our test statistic is then (Harrell, 2001):

$$W = \boldsymbol{\delta^T} \left[\boldsymbol{D} \left(\hat{\boldsymbol{V}}_{\boldsymbol{s}}(\boldsymbol{\theta_d})/N \right) \boldsymbol{D^T} \right]^{-1} \boldsymbol{\delta}$$

Where $\hat{\boldsymbol{V}}_{\boldsymbol{s}}(\boldsymbol{\theta_d})$ is the robust Murphy Topel estimates of variance for parameter subset $\boldsymbol{\theta_d}$ and N is the number of observations. Our test-statistic W is asymptotically χ^2 with Q degrees of freedom (Harrell, 2001).

For simultaneous testing of mean-changes and heteroskedasticity in GWAS, a common parameter subset that would be of interest would be the coefficients corresponding to the effect sizes of a given locus, and the covariance matrices for when the locus value is 0, 1, or 2. Suppose, as before, we have M components and P predictors, such that the P'th predictor in design matrix $\boldsymbol{X}$ corresponds to the allele count for a particular locus. Thus we have P by M matrix of coefficients $\boldsymbol{B}$ for that locus, and M by M covariance matrices $\boldsymbol{\Sigma}$, $\boldsymbol{T}$, or $\boldsymbol{\Omega}$ for when the i'th observation $\boldsymbol{x_{iP}}$ has a value of 0, 1, or 2 respectively. Our parameter subset of interest

would then be:

$$\boldsymbol{\theta_d} = \{\{\boldsymbol{B_{P\cdot}}\}, \boldsymbol{\Sigma}, \boldsymbol{T}, \boldsymbol{\Omega}\} = \{\beta_{P1}, ..., \beta_{PM}, ..., \sigma_{11}, ..., \sigma_{MM}, \tau_{11}, ..., \tau_{MM}, \omega_{11}, ..., \omega_{MM}\}$$

Where $\boldsymbol{\theta_d}$ has length $H = M + 3\left(\frac{M(M+1)}{2} - R\right)$, if R of the M components are binary. Then we have matrix $\boldsymbol{D}$ with H rows and Q columns, where $Q = M + 2\left(\frac{M(M+1)}{2} - R\right)$. We can represent $\boldsymbol{D}$ as a block matrix, with blocks pertaining to entries corresponding to coefficients $\boldsymbol{D_B}$ and covariance parameters $\boldsymbol{D_V}$, with $\boldsymbol{D_0}$ being matrices (composed of entries of zeroes) of appropriate size:

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{D_B} & \boldsymbol{D_0} \\ \boldsymbol{D_0^T} & \boldsymbol{D_V} \end{pmatrix}$$

For GWAS, our hypothesis test is such that each $\beta_{Pm} = 0$, and that $\sigma_{mm} = \tau_{mm} = \omega_{mm}$, i.e. $\sigma_{mm} - \tau_{mm} = 0$ and $\tau_{mm} - \omega_{mm} = 0$. Thus, $\boldsymbol{D_B}$ is an M by M identity matrix. $\boldsymbol{D_V}$ is a $2\left(\frac{M(M+1)}{2} - R\right)$ by $3\left(\frac{M(M+1)}{2} - R\right)$ matrix that consists of one 1 and one -1 on each row, with 0s otherwise:

$$\boldsymbol{D_V} = \begin{pmatrix} 1 & 0 & \dots & 0 & -1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & 0 & 0 & \dots & -1 \end{pmatrix}$$

4.3 Simulations

4.3.1 Coverage Regions

Using genotype data generated by HAP-SAMPLE2, we further simulated multivariate phenotype data in various scenarios under the null to investigate the performance of the Two-Stage Composite Likelihood across different input settings. We varied the phenotype specifications by the number of components (two to four), the number of binary/continuous components

(1 binary component to one less than the number of components), and the number of observations (2,500; 5,000; 7,500; and 10,000). We also used 6 arbitrarily chosen SNPs: rs375027891, rs536422311, rs540691877, rs577080509, rs71243915, and rs77267992. The four components had intercept values of -1.0, -0.5, 0.0, and 0.5.

For each combination of these specifications, we ran 500 replications and recorded coverage proportions at the 90%, 95%, and 99% intervals. This results in 36 tables, each of which has 3 confidence levels for 4 observation values. In each replication, the multivariate phenotype was simulated anew under the null. The resulting tables are contained later in the section under Supplementary Tables.

In addition to the Supplementary Tables, the Wald test-statistic coverage percentages are also displayed in Tables 4.1, 4.2, and 4.3. For the Wald test-statistic, the coverage proportions were calculated by recording the frequency by which the empirical Wald test-statistic was less than the χ^2 quantile values implied by the confidence probabilities (90%, 95%, and 99%) under the null. For the coefficient and covariance parameters, the coverage proportions were calculated by tabulating the percentage of confidence intervals that contained the true parameter value. A given confidence interval was formed via the parameter estimations for that replication and the accompanying robust Murphy-Topel standard errors.

Table 4.1: Coverage percentages for the Wald test-statistic at the 90% targeted conference region.

Comp.	Bin.	df	Obs.	rs375027891	rs536422311	rs540691877	rs577080509	rs71243915	rs77267992
2	1	6	2,500	90.2	91.0	88.0	91.0	89.8	87.8
2	1	6	5,000	90.6	88.0	89.8	88.8	89.2	90.4
2	1	6	7,500	91.0	92.4	88.6	89.2	88.6	91.6
2	1	6	10,000	90.2	89.8	89.2	89.0	90.0	91.6
3	2	11	2,500	88.2	88.6	88.4	87.0	87.6	88.4
3	2	11	5,000	86.0	89.4	90.2	91.6	89.6	86.6
3	2	11	7,500	91.6	87.6	89.8	90.8	90.0	90.4
3	2	11	10,000	88.4	89.0	89.6	89.2	88.8	90.0
3	1	13	2,500	88.0	89.6	87.6	87.4	88.6	88.2
3	1	13	5,000	88.0	89.6	88.6	89.2	88.8	89.4
3	1	13	7,500	90.2	90.2	86.2	87.6	89.8	87.2
3	1	13	10,000	92.2	90.4	90.8	89.8	89.4	91.4
4	3	18	2,500	87.8	85.4	87.8	88.0	87.2	85.4
4	3	18	5,000	90.8	88.8	89.4	88.6	89.4	88.6
4	3	18	7,500	90.6	89.0	90.2	88.8	88.8	91.6
4	3	18	10,000	90.4	89.6	89.6	91.8	89.4	90.2
4	2	20	2,500	86.8	85.4	86.2	88.2	86.0	87.8
4	2	20	5,000	87.8	88.0	90.8	88.8	88.6	88.4
4	2	20	7,500	89.6	89.2	86.4	88.4	88.4	91.6
4	2	20	10,000	88.8	89.6	87.8	88.4	92.0	90.6
4	1	22	2,500	84.4	87.6	87.4	84.4	83.4	86.6
4	1	22	5,000	89.4	90.2	88.2	84.8	85.0	86.8
4	1	22	7,500	90.4	89.2	86.6	90.4	89.4	88.6
4	1	22	10,000	88.2	90.0	88.4	90.2	85.2	91.4

Table 4.2: Coverage percentages for the Wald test-statistic at the 95% targeted conference region.

Comp.	Bin.	df	Obs.	rs375027891	rs536422311	rs540691877	rs577080509	rs71243915	rs77267992
2	1	6	2,500	93.8	95.4	94.4	95.4	93.6	93.4
2	1	6	5,000	95.8	94.4	93.8	93.4	96.0	94.2
2	1	6	7,500	96.4	96.2	93.6	96.0	94.8	94.4
2	1	6	10,000	95.4	95.4	95.0	95.0	96.0	95.0
3	2	11	2,500	93.0	95.0	94.6	91.6	93.8	94.4
3	2	11	5,000	93.0	94.4	94.4	94.2	95.6	93.2
3	2	11	7,500	94.4	93.4	95.8	95.4	94.6	95.6
3	2	11	10,000	94.0	93.6	95.0	93.6	95.4	94.8
3	1	13	2,500	93.6	93.8	94.4	93.2	93.6	94.2
3	1	13	5,000	95.0	94.4	95.6	94.0	94.2	94.6
3	1	13	7,500	96.0	94.6	93.2	95.2	94.4	94.0
3	1	13	10,000	96.6	95.8	96.4	95.8	95.0	95.4
4	3	18	2,500	94.2	93.6	92.8	93.6	93.8	93.6
4	3	18	5,000	95.6	92.6	93.6	94.4	93.6	93.2
4	3	18	7,500	96.2	95.6	95.2	94.2	94.8	95.4
4	3	18	10,000	96.2	94.6	95.8	94.8	95.2	95.8
4	2	20	2,500	92.8	91.8	93.4	93.0	91.0	94.2
4	2	20	5,000	93.4	95.0	95.2	92.8	93.8	95.0
4	2	20	7,500	93.6	94.4	92.8	94.4	93.4	95.6
4	2	20	10,000	95.6	93.6	94.8	95.0	96.0	95.4
4	1	22	2,500	92.4	93.6	92.6	90.4	92.6	92.4
4	1	22	5,000	94.8	94.4	93.0	92.0	92.4	93.6
4	1	22	7,500	95.6	94.4	92.0	94.0	93.8	94.2
4	1	22	10,000	93.0	95.0	94.8	95.6	93.2	95.2

Table 4.3: Coverage percentages for the Wald test-statistic at the 99% targeted conference region.

Comp.	Bin.	df	Obs.	rs375027891	rs536422311	rs540691877	rs577080509	rs71243915	rs77267992
2	1	6	2,500	98.2	99.0	99.0	98.8	98.6	98.8
2	1	6	5,000	99.0	98.4	98.8	98.8	99.2	99.4
2	1	6	7,500	99.4	99.4	99.0	98.6	98.6	99.6
2	1	6	10,000	99.0	98.2	98.8	99.2	99.0	98.6
3	2	11	2,500	98.6	98.6	98.2	97.6	99.0	98.6
3	2	11	5,000	98.8	99.0	99.0	98.6	99.4	98.2
3	2	11	7,500	99.2	99.0	99.4	98.4	99.0	98.8
3	2	11	10,000	98.2	98.4	99.6	99.4	99.2	99.0
3	1	13	2,500	98.6	97.8	98.2	97.6	98.8	98.6
3	1	13	5,000	99.2	98.8	99.4	98.8	98.8	99.4
3	1	13	7,500	99.4	98.8	99.0	99.2	98.8	99.0
3	1	13	10,000	99.6	98.8	99.4	99.6	99.0	99.0
4	3	18	2,500	98.4	98.2	98.2	97.8	99.0	98.4
4	3	18	5,000	99.6	98.4	97.6	98.4	98.2	99.4
4	3	18	7,500	99.2	99.0	99.2	98.2	97.8	98.6
4	3	18	10,000	99.4	99.0	99.0	99.6	98.6	98.8
4	2	20	2,500	98.0	97.8	98.4	98.2	98.4	98.8
4	2	20	5,000	98.8	98.8	98.6	98.2	98.6	98.8
4	2	20	7,500	98.4	99.2	98.6	98.8	97.8	98.6
4	2	20	10,000	99.2	98.2	99.4	99.4	99.2	98.8
4	1	22	2,500	98.0	99.2	98.4	98.0	98.4	98.2
4	1	22	5,000	98.8	98.4	98.6	98.2	98.2	98.8
4	1	22	7,500	99.6	99.4	98.8	98.2	98.8	98.6
4	1	22	10,000	99.2	98.8	99.4	99.4	99.0	97.8

4.3.2 Power

To illustrate the effect of mean and covariance associations upon power, simulations were performed to record the frequency by which the null hypothesis was rejected when the alternative hypotheses were true. The example of a two component multivariate response was used, with the first component binary, and the second continuous. With the same data-set as that used for confidence regions, rs375027891 served as the SNP. Four principal components were included as covariates, with 2,500, 5,000, 7,500, and 10,000 observations.

Under the null, the SNP effect was set to 0 for both components, alongside the coefficients for the principal components, for a 6 by 2 coefficient matrix of:

$$\boldsymbol{B}_0 = \begin{pmatrix} -1.0 & 0.5 \\ \mathbf{0} & \mathbf{0} \\ -0 & 0 \end{pmatrix}$$

Under the null, the covariance matrices were set to be mutually equal, and equal to the identity matrix:

$$\boldsymbol{\Sigma}_0 = \boldsymbol{T}_0 = \boldsymbol{\Omega}_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Under the various alternative hypotheses, the coefficients corresponding to the SNPs, as well as the covariance matrices, were incremented by 0.025 each time. For coefficients, this was controlled by $\boldsymbol{\zeta} = \{\zeta_1, \zeta_2\}$:

$$\boldsymbol{B}_{\boldsymbol{\zeta}} = \begin{pmatrix} -1.0 & 0.5 \\ \mathbf{0} & \mathbf{0} \\ \zeta_1 & \zeta_2 \end{pmatrix}$$

For covariances, this was controlled by parameter ξ. The covariance matrix for individuals of 0 alternative allele copies always remained the same, at the identity matrix:

$$\boldsymbol{\Sigma}_{\xi} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

However, individuals with one copy of the alternative allele would then have a covariance matrix of:

$$\boldsymbol{T}_{\xi} = \begin{pmatrix} 1 & \xi \\ \xi & 1+\xi \end{pmatrix}$$

And subsequently, individuals with two copies:

$$\boldsymbol{\Omega}_{\xi} = \begin{pmatrix} 1 & 2\xi \\ 2\xi & 1+2\xi \end{pmatrix}$$

The alternative hypotheses were expressed through five scenarios. In each scenario, different combinations of ζ_1, ζ_2, and ξ were incremented, for 10 total increments in each scenario. Under the first scenario, only ζ_1 was incremented up by 0.025 each time; ζ_2 and ξ remained at zero. Under the second, only ζ_2 was incremented. Both ζ_1 and ζ_2 were incremented for the third, with ξ remaining unchanged. For the fourth, only ξ was incremented; under the fifth, all three parameters were incremented. The results, with varying observation counts, are depicted in Figure 4.1.

As would be expected, for each scenario, larger effect sizes lead to greater rejection frequencies. The scenario where both ζ_1 and ζ_2 were incremented were rejected more frequently than the scenarios where only ζ_1 or ζ_2 were incremented. When all three of ζ_1, ζ_2, and ξ were incremented, this yielded the highest rejection frequency. Larger numbers of observations allowed for greater power.

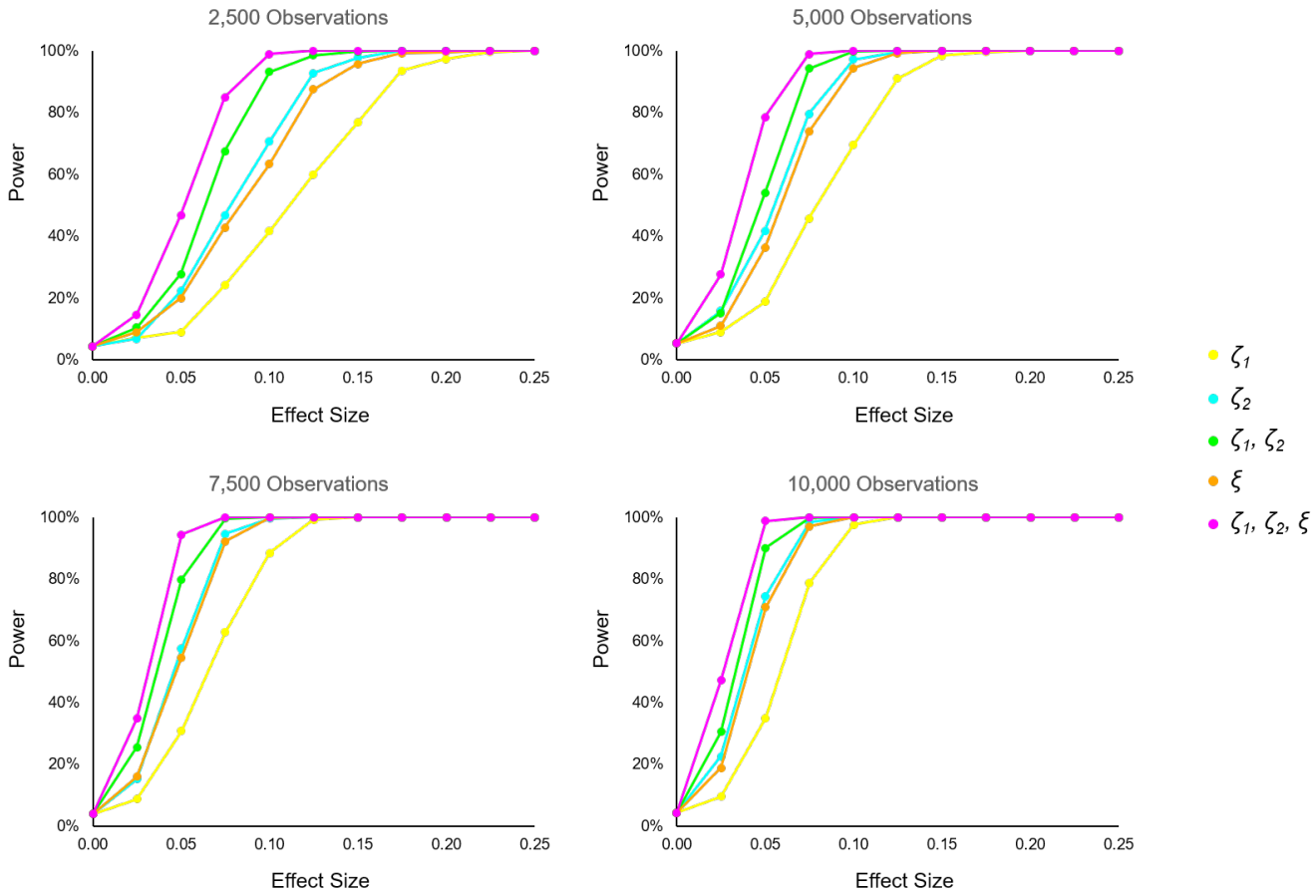

Figure 4.1: Power curves under the five scenarios. 500 replications were used for each scenario and increment, with 2,500, 5,000, 7,500, and 10,000 observations.

4.4 Discussion

The simulation results illustrate the favorable performance characteristics of the proposed method, and provide support for the Two-Stage Composite Likelihood as a methodology for testing mean and variance associations for multivariate heterogeneous traits. Coverage proportions for coefficients and covariance parameters generally hover around their ascribed intervals. χ^2 coverage proportions for the Wald statistic tend to be a bit more erratic, particularly for lower sample-sizes, but also realize more stable performance with greater sample sizes. There are some interesting areas for future analyses and investigation.

4.4.1 Hypothesis Testing

Our current implementation uses the Wald Test for hypothesis testing. The Wald Test need not be the only option for simultaneous-testing of mean changes and heteroskedasticity via the Two-Stage Composite Likelihood; ex-ante, the Likelihood Ratio Test and Score test would appear to be options, as well. However, we found the Wald Test more promising for our approach.

Nonetheless, future work could involve taking a look at the Score Test and/or the Likelihood Ratio Test, as potential alternatives to the Wald Test. The LRT and Score Test can sometimes have more favorable statistical properties (Harrell, 2001; Cattelan and Sartori, 2016). The test-statistics and parameter estimates may have more well-behaved coverage proportions, particularly for smaller sample sizes. For the Wald Test, any subtle issues in the estimating the covariance matrix can be magnified in the test-statistic (Harrell, 2001). The Wald Test also lacks invariance with respect to re-parameterization (Cattelan and Sartori, 2016; Harrell, 2001).

In general, the Wald Test requires only one set of parameter estimations in significance testing (Harrell, 2001), although it does require full estimation of the covariance matrix. For the Score Test and LRT, re-estimation of parameters under the null with restrictions would be needed (such as coefficients are zero and/or that variances are homoskedastic across allele

counts). Further complications and complexities can arise when working with two-stage estimation and composite likelihood. The Score Test tends to run into more computational issues, and for the composite likelihood at least, the LRT has an "awkward" asymptotic distribution (Cattelan and Sartori, 2016). The LRT for the composite likelihood has an asymptotic distribution that is a weighted sum of P χ^2 random variables (corresponding to P free parameters), and would need some adjustments to recover an approximately χ^2 asymptotic distribution (Cattelan and Sartori, 2016).

4.4.2 Ordinal Components

Incorporating handling of ordinal components would potentially be interesting, alongside binary and continuous components. Conceptually-speaking, an ordinal univariate response lies in between a binary and continuous response, and thus the addition of ordinal components would more fully cover the range of possible response types.

For example, consider a univariate binary observation y_i with latent observation y_i^* that can take-on a 0 or 1 depending if standard normal y_i^* exceeds $c_i = \boldsymbol{x_i}b$ or not. We can generalize this for a greater number of responses beyond 2. Following (Sajaia, 2009), a univariate ordinal observation y_i with latent variable y_i^* would correspondingly take on values of $\{1, 2, ..., J\}$:

$$y_i = \begin{cases} 1 & y_i^* \leq c_1 \\ 2 & c_1 < y_i^* \leq c_2 \\ ... & \\ J & y_i^* > c_{J-1} \end{cases}$$

Sajaia (2009) discusses maximum likelihood estimation of bivariate ordinal responses. However, for the Two-Stage Composite Likelihood—although ordinal components fit in the framework from a conceptual standpoint—it would remain to be seen how a binary-ordinal or ordinal-continuous response would be estimated. The multiple cut-points from a ordinal component pose a new source of challenge, as a binary component only has one such cut-point.

A priori, it seems associated likelihoods in the Two-Stage Composite Likelihood estimations involving ordinal components could be materially more difficult than either binary components and/or continuous components.

4.4.3 Other Aspects of Genomics

Another area that could be interesting is consideration of possible tie-ins with gene-gene and/or gene-environment interactions. Detecting SNPs associated with variance could be indicative of the presence of interaction effects, as discussed by Young et al. (2018) with the example of identifying associations with body mass index using UK Biobank data. Interactions between loci can be quite difficult to model, as the number of potential interactions of interest tends to increase much faster than the number of loci.

Being a genomics-focused approach (although easily applicable to other domains), we are primarily concerned with biallelic loci, where an individual's genotype at a particular locus would assume a value of 0, 1, or 2, for three groups in total. Although our approach for heteroskedasticity, conceptually-speaking, readily extends beyond three groups, more investigation in this realm would certainly be accretive. While most loci are biallelic, modern-day data-sets, such as 1000 Genomes, often do contain data on multiallelic loci (Auton, A. and Abecasis, G. and Altshuler, D. et al., 2015). However, multiallelic loci would likely put additional pressure on sample size considerations, as each group would need a sufficient sample size to reliably estimate their covariance matrices.

4.5 Supplementary Tables

Table 4.4: Coverage percentages using rs375027891 with 2 components, of which 1 was binary, for 6 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	σ_{12}	σ_{22}	τ_{12}	τ_{22}	ω_{12}	ω_{22}
2,500	90	90.2	88.2	90.0	90.6	90.4	89.6	91.0	91.4	90.6
2,500	95	93.8	93.2	94.0	95.2	95.4	94.4	95.2	95.8	96.0
2,500	99	98.2	97.4	99.4	99.0	99.0	98.6	98.8	98.4	99.0
5,000	90	90.6	91.6	89.4	88.8	90.8	89.6	93.6	89.6	88.6
5,000	95	95.8	96.2	94.6	94.4	95.6	94.4	96.8	95.8	94.4
5,000	99	99.0	98.6	98.6	98.8	99.2	98.4	99.0	99.4	98.2
7,500	90	91.0	92.2	84.8	88.4	90.4	91.0	89.2	92.0	90.2
7,500	95	96.4	95.6	92.8	93.6	96.8	95.2	94.0	96.8	96.2
7,500	99	99.4	99.0	98.0	98.4	99.8	99.2	99.0	99.6	99.4
10,000	90	90.2	92.2	90.0	90.4	89.2	91.0	88.4	90.0	88.4
10,000	95	95.4	95.6	94.6	95.2	95.0	95.8	93.6	95.8	94.0
10,000	99	99.0	99.2	98.8	99.4	98.8	98.8	99.2	99.0	98.8

Table 4.5: Coverage percentages using rs375027891 with 3 components, of which 1 was binary, for 13 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	σ_{12}	σ_{13}	σ_{22}	σ_{23}	σ_{33}	τ_{12}	τ_{13}	τ_{22}	τ_{23}	τ_{33}	ω_{12}	ω_{13}	ω_{22}	ω_{23}	ω_{33}
2,500	90	88.0	89.0	90.8	93.0	88.0	90.4	89.6	91.4	88.8	90.4	89.4	88.2	89.0	89.4	90.2	89.4	88.2	87.6	91.0
2,500	95	93.6	94.2	95.4	96.0	94.8	94.6	95.4	96.6	93.8	96.2	94.4	92.8	95.0	93.6	94.4	94.8	92.4	92.8	95.0
2,500	99	98.6	99.2	99.2	99.2	99.0	98.2	98.2	99.2	99.0	99.2	97.8	99.0	99.4	98.4	98.6	98.6	98.6	98.0	98.6
5,000	90	88.0	87.6	90.8	89.0	89.2	88.4	89.8	90.2	89.2	91.6	90.6	90.8	90.2	90.0	91.4	92.6	89.4	90.2	91.2
5,000	95	95.0	93.0	95.6	94.0	94.0	93.0	94.4	95.8	95.4	95.8	96.0	95.4	96.0	96.2	94.2	96.2	95.2	95.0	95.0
5,000	99	99.2	98.8	99.4	98.2	98.2	98.4	98.8	99.2	98.8	99.0	99.6	99.0	99.0	99.2	98.6	98.6	99.8	98.6	99.4
7,500	90	90.2	90.8	90.4	88.4	92.4	91.2	88.8	90.4	90.2	89.6	90.8	90.0	89.6	88.4	91.2	88.8	89.6	90.8	90.8
7,500	95	96.0	94.6	94.0	95.4	97.0	96.0	93.4	95.4	94.4	94.4	95.4	95.2	95.2	93.6	95.4	95.2	95.0	94.8	94.4
7,500	99	99.4	99.0	98.8	99.0	99.0	99.2	98.6	99.2	98.6	98.8	98.6	99.0	98.4	98.6	98.8	98.6	99.0	99.2	100.0
10,000	90	92.2	91.8	90.2	88.8	89.8	91.6	91.6	89.0	89.6	87.6	88.6	90.2	90.0	90.6	89.4	90.6	89.4	91.8	91.0
10,000	95	96.6	95.6	95.2	94.8	95.8	95.6	95.8	94.4	95.8	93.0	93.6	95.2	95.0	95.6	95.4	95.4	94.4	97.0	95.4
10,000	99	99.6	99.2	98.6	99.4	98.8	99.0	99.4	99.0	99.0	99.2	98.4	99.0	98.8	98.4	99.0	99.2	99.8	99.0	98.0

Table 4.6: Coverage percentages using rs375027891 with 3 components, of which 2 were binary, for 11 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	σ_{12}	σ_{13}	σ_{23}	σ_{33}	τ_{12}	τ_{13}	τ_{23}	τ_{33}	ω_{12}	ω_{13}	ω_{23}	ω_{33}
2,500	90	88.2	88.6	88.8	91.4	88.4	88.8	92.2	92.4	88.8	89.2	88.4	89.4	90.0	89.8	90.4	89.2
2,500	95	93.0	95.8	93.6	94.2	94.4	94.8	95.8	96.2	93.4	94.6	94.8	94.6	93.2	94.6	94.2	96.2
2,500	99	98.6	99.4	99.2	99.2	98.8	98.6	99.4	99.2	98.4	99.0	99.2	98.6	98.0	98.6	99.4	99.2
5,000	90	86.0	89.8	88.0	91.2	87.6	91.4	90.6	89.6	89.6	90.4	90.8	88.2	88.2	90.6	87.6	92.0
5,000	95	93.0	94.6	94.2	95.0	95.0	96.8	95.6	94.8	95.2	94.2	96.0	92.8	94.2	95.4	94.2	95.4
5,000	99	98.8	99.2	99.2	99.2	98.6	99.4	99.2	98.4	100.0	98.6	98.6	97.6	99.0	99.2	99.0	99.2
7,500	90	91.6	89.8	89.2	90.6	88.6	92.4	88.4	87.6	92.2	89.6	90.6	90.4	89.0	90.8	90.0	89.0
7,500	95	94.4	96.0	93.6	94.6	92.8	95.2	94.6	93.4	95.4	93.8	96.0	95.4	93.8	94.4	95.8	93.0
7,500	99	99.2	99.8	98.8	98.6	99.0	98.6	99.4	98.6	99.0	98.8	99.4	99.8	99.0	99.2	99.4	98.8
10,000	90	88.4	87.4	89.8	90.6	91.4	90.6	91.4	88.2	89.6	90.0	89.6	90.8	89.4	88.6	87.0	88.0
10,000	95	94.0	92.2	94.2	95.6	97.0	95.6	95.8	94.2	95.2	94.0	95.0	94.8	95.4	93.2	92.0	93.0
10,000	99	98.2	98.6	99.4	99.0	98.8	98.8	99.6	98.6	99.6	98.6	98.6	99.2	99.4	98.0	98.0	99.0

Table 4.7: Coverage percentages using rs375027891 with 4 components, of which 1 was binary, for 22 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	β_{46}	σ_{12}	σ_{13}	σ_{14}	σ_{22}	σ_{23}	σ_{24}	σ_{33}	σ_{34}	σ_{44}
2,500	90	84.4	89.8	89.8	91.4	88.4	88.8	88.6	89.2	91.2	87.6	88.4	89.0	90.8	88.8
2,500	95	92.4	94.6	95.4	96.0	94.8	93.8	94.2	94.4	94.0	94.6	93.6	93.2	94.0	94.0
2,500	99	98.0	99.2	98.8	99.0	99.6	98.6	99.0	99.0	98.8	98.6	98.4	98.2	99.0	99.0
5,000	90	89.4	90.8	90.8	90.2	89.8	88.8	91.4	88.0	89.4	90.4	88.8	90.2	91.2	89.6
5,000	95	94.8	96.0	96.0	95.6	94.2	95.4	96.4	93.4	94.8	95.6	94.2	96.2	95.8	94.8
5,000	99	98.8	99.8	99.4	99.6	98.4	98.6	99.4	98.6	98.6	99.2	98.8	99.0	99.4	99.0
7,500	90	90.4	88.0	88.2	90.2	90.2	91.8	88.4	90.2	86.6	88.4	87.2	90.8	89.8	91.6
7,500	95	95.6	93.6	94.0	95.0	95.6	94.8	94.8	94.8	93.8	95.0	93.8	95.2	94.6	95.2
7,500	99	99.6	99.0	99.2	99.0	98.2	99.0	99.0	99.0	98.4	99.6	98.2	98.6	98.8	98.6
10,000	90	88.2	89.6	89.8	88.4	87.0	88.0	90.6	88.4	90.6	91.0	88.4	90.0	87.2	87.8
10,000	95	93.0	94.4	96.0	94.6	94.4	95.4	95.4	94.0	95.6	96.0	93.6	95.8	93.2	93.2
10,000	99	99.2	98.2	99.6	98.6	98.6	98.8	99.0	98.8	99.4	99.6	98.2	98.4	99.6	98.6

Obs.	CI	τ_{12}	τ_{13}	τ_{14}	τ_{22}	τ_{23}	τ_{24}	τ_{33}	τ_{34}	τ_{44}	ω_{12}	ω_{13}	ω_{14}	ω_{22}	ω_{23}	ω_{24}	ω_{33}	ω_{34}	ω_{44}
2,500	90	92.0	90.6	89.0	89.4	89.6	88.8	89.4	91.0	90.2	86.2	90.4	89.2	89.6	89.6	91.2	88.8	91.0	89.0
2,500	95	95.8	96.0	94.8	94.4	95.4	94.4	95.0	95.6	95.6	92.2	94.2	94.6	93.8	93.8	96.2	94.2	97.0	94.2
2,500	99	99.4	98.4	98.6	98.4	98.4	98.6	98.4	99.4	99.0	97.4	98.2	99.0	98.6	98.2	99.6	98.2	99.0	98.6
5,000	90	90.2	89.6	89.4	92.4	89.0	89.8	90.0	88.4	92.0	89.0	91.6	89.8	91.8	91.8	90.8	91.2	88.4	92.2
5,000	95	95.0	94.4	95.2	96.0	95.4	95.2	96.2	96.4	96.4	94.4	95.8	95.0	95.4	96.2	95.2	96.0	94.0	96.8
5,000	99	99.0	98.8	99.4	99.2	99.0	99.2	99.0	99.2	99.4	99.4	98.6	99.4	99.2	99.0	99.0	99.4	98.8	98.8
7,500	90	88.6	88.8	90.4	89.6	91.6	91.8	90.4	88.8	92.4	91.8	90.4	89.0	90.4	90.8	94.0	85.8	88.2	88.2
7,500	95	94.6	95.0	95.6	95.0	95.2	95.4	96.2	95.6	96.6	96.0	96.4	95.0	94.6	95.0	97.4	93.4	93.8	93.6
7,500	99	98.4	99.4	99.4	99.0	99.0	99.2	99.6	99.0	99.6	99.0	99.8	99.4	99.0	98.6	99.8	97.4	98.6	99.2
10,000	90	93.0	91.0	89.6	90.0	90.2	88.6	88.6	88.6	88.8	87.6	91.0	89.8	88.6	90.6	87.6	89.6	90.8	91.0
10,000	95	96.2	95.2	94.6	95.8	94.6	93.6	94.2	94.2	92.2	94.8	94.8	96.0	94.8	96.2	94.4	94.6	95.8	95.6
10,000	99	99.6	98.6	98.6	99.4	99.2	99.2	99.0	99.0	98.4	99.0	99.4	98.8	99.6	99.4	99.2	98.8	99.2	99.4

Table 4.8: Coverage percentages using rs375027891 with 4 components, of which 2 were binary, for 20 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	β_{46}	σ_{12}	σ_{13}	σ_{14}	σ_{23}	σ_{24}	σ_{33}	σ_{34}	σ_{44}
2,500	90	86.8	91.8	88.0	91.2	91.8	90.4	90.2	91.2	86.2	91.6	89.0	90.4	88.8
2,500	95	92.8	96.0	94.4	95.6	96.0	94.6	95.4	98.0	93.6	95.2	94.2	94.4	93.8
2,500	99	98.0	99.0	98.4	99.2	99.4	99.6	99.2	99.8	98.4	98.8	99.0	98.6	98.0
5,000	90	87.8	91.2	91.0	89.8	88.4	89.0	89.4	88.8	90.4	89.8	90.2	92.4	86.2
5,000	95	93.4	94.0	95.0	94.4	93.6	94.4	93.2	94.2	94.8	94.4	94.6	96.0	93.2
5,000	99	98.8	98.2	98.4	99.4	99.0	98.8	98.2	98.8	100.0	99.2	99.4	98.6	98.0
7,500	90	89.6	90.2	88.6	87.8	89.8	90.0	88.4	92.2	87.8	90.6	88.2	90.2	89.6
7,500	95	93.6	95.2	94.2	93.8	94.8	94.8	94.2	96.8	93.2	94.4	94.2	95.4	96.0
7,500	99	98.4	99.2	99.0	99.0	98.6	99.2	98.2	99.4	99.0	99.0	98.6	99.2	99.2
10,000	90	88.8	91.6	92.4	90.2	89.8	87.8	89.0	90.4	89.4	89.4	91.0	88.6	90.6
10,000	95	95.6	96.2	96.8	94.6	95.6	93.2	94.4	96.2	94.4	94.0	95.2	94.0	94.4
10,000	99	99.2	99.2	99.2	98.6	99.6	98.6	99.2	99.8	99.4	98.4	98.8	98.8	98.4

Obs.	CI	τ_{12}	τ_{13}	τ_{14}	τ_{23}	τ_{24}	τ_{33}	τ_{34}	τ_{44}	ω_{12}	ω_{13}	ω_{14}	ω_{23}	ω_{24}	ω_{33}	ω_{34}	ω_{44}
2,500	90	87.8	89.2	88.2	90.4	93.0	89.2	87.8	88.8	89.0	89.4	90.6	89.8	88.8	90.4	87.2	88.2
2,500	95	93.8	94.8	95.2	94.8	96.4	95.0	92.6	93.6	93.6	95.6	94.8	95.2	92.6	95.4	94.4	95.8
2,500	99	99.4	99.2	99.6	98.8	99.6	99.2	97.6	98.8	98.8	99.2	98.6	98.8	98.2	98.2	98.4	99.4
5,000	90	92.6	90.4	88.8	91.6	89.0	89.4	88.4	92.6	91.2	89.8	92.4	88.6	91.0	90.6	90.0	89.8
5,000	95	96.0	95.2	94.2	97.0	95.0	94.6	93.6	96.4	95.4	94.4	95.6	94.2	95.2	96.0	95.0	94.0
5,000	99	97.8	99.4	98.6	99.4	98.6	97.8	98.6	99.6	99.2	98.2	98.8	98.8	99.2	98.0	98.6	98.8
7,500	90	90.4	89.4	89.0	92.0	91.2	89.0	91.6	92.0	91.0	89.6	93.4	89.0	89.8	88.4	91.6	89.6
7,500	95	95.4	94.8	93.8	96.0	95.4	95.8	95.2	96.0	95.6	94.2	96.0	95.4	95.6	93.8	97.0	94.0
7,500	99	99.0	99.0	97.8	99.8	99.4	99.4	98.8	99.2	98.6	99.4	99.8	99.0	99.8	98.8	99.6	99.0
10,000	90	91.6	91.6	89.0	90.8	87.6	90.6	89.6	89.2	91.0	90.0	89.8	89.0	88.0	90.4	90.4	88.8
10,000	95	96.6	96.0	92.8	95.0	93.4	94.8	95.0	93.6	96.4	94.4	95.4	96.4	95.6	93.8	94.8	93.6
10,000	99	99.6	99.4	98.6	99.0	98.2	98.6	98.2	98.6	99.4	98.4	98.8	99.4	99.0	98.4	98.8	97.8

Table 4.9: Coverage percentages using rs375027891 with 4 components, of which 3 were binary, for 18 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	β_{46}	σ_{12}	σ_{13}	σ_{14}	σ_{23}	σ_{24}	σ_{34}	σ_{44}
2,500	90	87.8	90.0	87.4	90.6	89.2	88.2	89.4	90.0	87.0	90.8	90.2	91.2
2,500	95	94.2	93.6	93.8	95.4	95.0	95.0	93.8	95.2	93.2	95.0	94.6	95.4
2,500	99	98.4	98.4	98.4	99.0	99.2	99.0	98.4	98.8	98.8	99.2	99.2	98.0
5,000	90	90.8	90.8	87.8	88.4	94.2	90.8	91.8	91.4	90.8	90.6	90.2	90.2
5,000	95	95.6	94.8	95.0	93.4	96.2	95.2	96.4	96.2	95.2	95.4	94.6	95.8
5,000	99	99.6	98.6	99.0	99.0	98.4	99.2	98.8	99.8	99.0	99.4	99.2	98.6
7,500	90	90.6	90.0	90.6	92.2	90.8	88.8	92.0	90.2	91.4	89.2	92.0	91.4
7,500	95	96.2	95.2	96.0	96.2	95.2	95.0	95.8	94.4	95.6	94.0	96.2	95.2
7,500	99	99.2	99.6	99.6	99.8	99.2	99.0	98.8	99.2	99.2	99.2	99.4	98.8
10,000	90	90.4	88.8	90.8	90.0	90.2	89.0	92.4	89.0	91.4	88.4	90.4	90.6
10,000	95	96.2	93.6	95.0	95.6	95.4	94.4	96.0	95.2	95.0	93.6	95.6	95.2
10,000	99	99.4	98.0	99.4	99.8	99.4	99.0	100.0	99.2	99.6	99.0	98.4	99.8

Obs.	CI	τ_{12}	τ_{13}	τ_{14}	τ_{23}	τ_{24}	τ_{34}	τ_{44}	ω_{12}	ω_{13}	ω_{14}	ω_{23}	ω_{24}	ω_{34}	ω_{44}
2,500	90	90.6	90.8	91.0	89.4	88.4	92.4	91.0	89.8	88.4	88.8	90.0	91.8	90.8	89.8
2,500	95	94.0	95.4	96.0	95.2	94.2	96.0	95.6	93.8	94.0	93.6	95.8	94.6	96.0	94.8
2,500	99	98.6	99.4	98.8	99.6	99.0	99.4	98.4	99.6	98.6	98.6	99.8	98.6	98.4	98.8
5,000	90	88.4	88.8	91.2	91.8	89.8	91.6	91.0	91.2	89.6	90.8	90.4	89.8	89.6	91.2
5,000	95	94.0	94.2	95.4	95.2	94.8	95.0	96.6	96.0	94.4	94.2	96.4	96.2	95.4	96.6
5,000	99	97.6	98.8	99.6	98.8	99.0	99.2	99.0	99.0	98.8	99.0	99.6	99.6	99.8	99.2
7,500	90	91.2	88.8	88.2	89.2	89.4	92.4	89.2	89.4	91.6	91.8	91.4	90.6	88.4	90.6
7,500	95	95.8	95.0	92.6	94.6	94.6	95.8	95.0	95.2	95.2	95.2	96.4	95.2	94.8	96.6
7,500	99	99.4	99.0	98.0	98.6	99.2	99.0	99.0	99.2	99.2	98.8	99.6	98.8	99.0	99.4
10,000	90	87.4	90.2	88.2	87.6	90.6	91.0	88.0	88.4	90.4	91.0	86.4	88.4	90.2	90.0
10,000	95	93.8	95.2	95.0	94.8	95.8	96.2	94.0	94.6	95.0	96.2	92.6	95.6	95.0	95.0
10,000	99	99.2	98.8	99.0	99.2	99.4	99.4	98.4	98.2	98.8	99.2	98.8	98.8	99.2	99.2

Table 4.10: Coverage percentages using rs536422311 with 2 components, of which 1 was binary, for 6 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	σ_{12}	σ_{22}	τ_{12}	τ_{22}	ω_{12}	ω_{22}
2,500	90	91.0	87.2	90.4	91.8	88.6	91.2	91.0	92.2	90.6
2,500	95	95.4	93.6	95.2	97.0	94.8	95.4	94.4	96.6	94.6
2,500	99	99.0	98.6	99.6	99.4	99.0	99.2	99.0	99.2	99.2
5,000	90	88.0	89.2	88.0	90.6	90.4	88.2	89.4	90.6	88.2
5,000	95	94.4	94.4	94.2	95.2	95.8	94.4	94.2	94.4	93.6
5,000	99	98.4	99.4	98.8	99.2	99.2	98.0	98.6	99.2	97.6
7,500	90	92.4	89.2	90.2	90.6	93.0	88.8	90.8	89.2	89.8
7,500	95	96.2	95.8	96.0	94.8	96.2	93.0	95.0	94.6	94.4
7,500	99	99.4	99.4	99.0	99.0	99.4	97.8	98.8	99.2	98.0
10,000	90	89.8	88.8	92.2	88.0	87.2	90.4	88.4	89.8	86.0
10,000	95	95.4	95.4	95.4	94.6	93.0	94.8	93.2	95.8	93.4
10,000	99	98.2	99.4	99.0	98.6	98.0	99.8	99.4	98.8	98.8

Table 4.11: Coverage percentages using rs536422311 with 3 components, of which 1 was binary, for 13 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	σ_{12}	σ_{13}	σ_{22}	σ_{23}	σ_{33}	τ_{12}	τ_{13}	τ_{22}	τ_{23}	τ_{33}	ω_{12}	ω_{13}	ω_{22}	ω_{23}	ω_{33}
2,500	90	89.6	89.8	92.8	89.6	92.2	89.4	89.0	90.2	88.0	90.0	90.2	87.8	90.4	88.6	89.6	88.4	89.6	88.6	91.0
2,500	95	93.8	95.2	96.2	94.2	95.0	93.8	93.2	96.0	93.6	95.2	95.6	93.2	94.0	94.4	94.2	93.0	94.4	94.0	94.8
2,500	99	97.8	98.6	99.2	99.4	99.8	98.8	98.0	99.4	99.2	98.8	99.6	99.0	98.8	99.2	98.4	98.6	98.4	98.2	99.4
5,000	90	89.6	90.6	88.6	89.6	90.4	90.2	91.6	90.2	90.6	91.2	88.2	89.4	91.8	89.2	87.0	91.6	91.8	90.0	91.0
5,000	95	94.4	94.4	93.8	93.4	95.4	93.8	95.4	95.4	94.4	96.8	94.4	95.4	95.8	94.8	93.6	96.6	95.6	94.0	95.0
5,000	99	98.8	99.2	99.2	98.6	98.6	97.8	99.4	99.6	98.2	99.4	98.6	99.0	99.8	98.6	99.4	99.6	99.2	98.2	98.6
7,500	90	90.2	90.8	92.4	89.6	89.6	87.8	88.8	90.6	89.4	88.6	92.2	89.8	89.8	91.6	89.6	88.4	89.8	90.0	90.4
7,500	95	94.6	94.6	95.6	94.4	94.6	94.8	94.4	95.6	94.2	94.2	96.2	95.2	96.0	95.4	94.6	94.2	94.6	94.4	95.2
7,500	99	98.8	98.6	99.2	99.0	98.4	98.8	99.2	99.0	99.4	98.6	99.8	99.4	99.4	98.8	99.0	98.8	98.0	99.0	98.6
10,000	90	90.4	89.0	91.0	91.8	91.0	89.8	89.6	88.2	88.6	90.2	91.0	88.8	88.4	91.6	88.6	89.2	90.4	88.8	91.8
10,000	95	95.8	94.2	95.0	95.4	95.6	94.8	95.6	93.4	94.0	95.0	95.6	94.8	93.6	94.4	94.2	93.2	94.6	95.8	95.0
10,000	99	98.8	99.6	99.4	99.4	98.8	99.4	97.8	98.2	98.8	98.8	99.0	99.2	99.0	98.6	98.8	98.8	99.0	98.4	98.6

Table 4.12: Coverage percentages using rs536422311 with 3 components, of which 2 were binary, for 11 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	σ_{12}	σ_{13}	σ_{23}	σ_{33}	τ_{12}	τ_{13}	τ_{23}	τ_{33}	ω_{12}	ω_{13}	ω_{23}	ω_{33}
2,500	90	88.6	90.2	92.6	92.6	90.2	90.4	87.4	91.2	90.8	91.2	91.4	86.4	91.2	88.2	90.4	90.6
2,500	95	95.0	95.6	96.2	96.2	94.4	95.2	95.2	96.6	95.6	94.2	95.6	94.4	95.2	94.6	94.8	95.8
2,500	99	98.6	98.6	99.2	99.6	98.8	99.0	99.0	99.2	98.8	99.2	98.8	99.6	98.8	99.0	98.8	98.8
5,000	90	89.4	90.2	89.6	89.2	89.6	91.4	91.8	90.2	90.2	88.8	91.2	92.2	87.4	88.8	89.8	90.2
5,000	95	94.4	94.0	94.6	94.8	94.4	95.4	96.4	94.6	95.4	95.0	95.0	95.6	94.2	93.8	94.8	94.0
5,000	99	99.0	99.4	98.6	99.4	99.2	99.4	99.2	99.4	99.4	99.0	99.0	98.4	98.2	98.4	98.2	98.8
7,500	90	87.6	92.0	92.6	89.6	92.0	90.0	90.2	89.8	86.6	87.4	90.2	90.8	90.2	90.2	88.0	90.4
7,500	95	93.4	96.4	96.0	94.2	95.8	94.4	96.4	94.8	93.4	94.2	95.4	94.0	95.0	95.4	93.4	93.8
7,500	99	99.0	99.0	98.8	99.2	99.6	99.2	99.0	98.6	99.0	98.4	98.8	98.6	99.2	99.6	98.8	98.4
10,000	90	89.0	90.6	88.6	89.2	88.6	87.8	89.8	90.4	90.0	89.2	90.6	91.8	88.8	87.6	90.6	90.2
10,000	95	93.6	95.2	93.4	94.2	94.6	92.8	94.2	94.2	95.0	93.2	96.2	94.8	93.2	92.6	94.8	95.0
10,000	99	98.4	99.2	98.4	98.4	99.8	98.4	98.6	98.8	99.2	98.2	99.8	98.8	98.8	98.4	99.2	98.6

Table 4.13: Coverage percentages using rs536422311 with 4 components, of which 1 was binary, for 22 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	β_{46}	σ_{12}	σ_{13}	σ_{14}	σ_{22}	σ_{23}	σ_{24}	σ_{33}	σ_{34}	σ_{44}
2,500	90	87.6	89.8	90.2	91.4	92.4	90.8	89.0	88.8	91.4	91.0	88.6	91.6	90.4	89.4
2,500	95	93.6	95.2	94.4	95.4	95.8	93.8	94.4	94.2	95.2	94.8	94.4	94.6	94.8	94.4
2,500	99	99.2	99.2	99.2	98.8	99.2	99.2	98.4	98.8	98.8	98.8	98.6	99.0	98.2	99.0
5,000	90	90.2	90.6	91.0	89.0	90.0	89.0	89.6	87.6	90.6	88.6	89.8	89.6	91.6	90.2
5,000	95	94.4	94.4	95.0	93.4	95.4	95.6	95.6	92.0	95.0	92.6	93.6	96.0	95.6	95.0
5,000	99	98.4	99.8	99.0	98.6	99.0	98.2	98.6	98.6	99.4	98.4	98.8	99.4	99.4	99.0
7,500	90	89.2	90.6	88.8	89.4	90.4	88.4	91.2	90.0	91.0	90.2	87.0	90.6	88.4	89.6
7,500	95	94.4	96.2	93.4	94.8	95.2	95.6	94.4	95.0	95.2	94.6	93.4	95.4	93.8	95.0
7,500	99	99.4	99.4	99.2	99.2	98.6	99.0	99.0	99.2	99.0	99.0	98.6	98.2	99.2	99.4
10,000	90	90.0	92.2	92.0	89.6	90.6	88.2	92.0	90.4	91.8	88.2	91.4	88.2	88.0	90.0
10,000	95	95.0	96.0	96.2	95.2	95.2	95.4	95.2	94.4	96.4	94.2	95.4	94.2	94.8	95.2
10,000	99	98.8	99.6	100.0	99.4	98.8	99.2	98.6	99.0	99.4	99.6	99.0	98.8	98.8	99.4

Obs.	CI	τ_{12}	τ_{13}	τ_{14}	τ_{22}	τ_{23}	τ_{24}	τ_{33}	τ_{34}	τ_{44}	ω_{12}	ω_{13}	ω_{14}	ω_{22}	ω_{23}	ω_{24}	ω_{33}	ω_{34}	ω_{44}
2,500	90	89.4	90.6	89.0	90.0	88.2	89.2	91.2	90.2	91.2	86.8	89.4	91.0	90.6	92.2	90.4	92.8	90.6	86.2
2,500	95	94.4	95.0	94.2	95.6	94.4	95.8	95.0	95.6	95.2	92.6	94.2	94.4	94.8	96.6	96.0	96.8	94.4	93.8
2,500	99	98.4	98.0	98.2	99.4	98.6	98.8	99.0	98.8	99.2	97.8	99.0	98.8	98.6	99.2	99.0	99.2	98.6	98.8
5,000	90	87.6	90.8	89.8	88.4	91.8	88.6	91.8	91.2	89.4	88.4	88.4	87.8	91.2	93.6	88.2	91.4	89.2	87.8
5,000	95	94.2	94.6	93.8	95.2	96.0	94.0	96.0	95.4	95.2	94.0	93.2	92.8	95.0	96.0	93.8	95.2	93.6	93.0
5,000	99	99.4	99.4	98.6	99.6	98.8	98.8	99.2	99.2	99.0	98.6	98.2	98.6	99.0	98.8	98.8	98.4	99.0	97.2
7,500	90	89.2	88.2	91.2	90.2	92.6	91.4	91.2	90.8	90.2	88.4	90.0	92.4	91.6	89.4	91.2	88.0	92.6	89.4
7,500	95	93.8	95.2	95.2	94.2	97.0	95.4	95.2	95.2	94.8	94.2	93.6	96.6	95.4	94.4	95.0	93.6	97.0	94.8
7,500	99	98.4	99.4	99.0	98.6	99.6	99.0	99.2	98.8	99.0	99.2	99.0	99.0	99.2	99.2	99.0	99.0	99.4	99.0
10,000	90	91.8	88.0	87.8	90.2	89.8	89.0	92.6	89.8	90.2	90.2	91.0	90.4	91.2	91.2	90.2	89.4	92.4	90.8
10,000	95	94.6	93.8	93.8	95.4	94.6	94.6	96.0	94.6	96.0	95.6	95.4	95.6	96.2	95.4	95.4	94.4	97.0	94.0
10,000	99	99.2	98.8	99.0	99.2	98.8	98.8	99.2	98.2	99.0	99.4	99.0	98.4	99.0	99.4	99.2	99.0	99.6	98.8

Table 4.14: Coverage percentages using rs536422311 with 4 components, of which 2 were binary, for 20 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	β_{46}	σ_{12}	σ_{13}	σ_{14}	σ_{23}	σ_{24}	σ_{33}	σ_{34}	σ_{44}
2,500	90	85.4	91.2	92.0	89.0	89.0	87.0	90.6	89.8	91.6	88.0	89.8	89.4	88.8
2,500	95	91.8	96.2	96.0	94.0	94.6	94.0	95.2	95.2	96.4	93.0	94.8	95.2	93.0
2,500	99	97.8	99.4	99.6	97.6	99.4	99.0	99.4	98.8	99.2	98.6	98.2	99.6	98.0
5,000	90	88.0	89.0	89.6	90.4	90.0	87.6	90.2	90.0	86.4	89.2	89.2	90.2	91.0
5,000	95	95.0	95.2	95.0	95.8	95.2	92.8	94.0	94.8	91.8	95.0	95.6	96.0	94.8
5,000	99	98.8	98.2	98.8	98.8	99.0	98.4	98.6	98.8	98.6	99.0	98.2	99.0	99.6
7,500	90	89.2	88.2	90.6	89.6	90.0	88.8	88.2	91.6	88.8	89.8	89.4	89.2	92.8
7,500	95	94.4	92.4	94.6	94.2	95.0	95.4	93.8	96.6	94.0	94.8	94.8	95.0	96.2
7,500	99	99.2	98.8	99.2	98.2	99.6	99.0	98.4	99.6	100.0	99.0	99.0	99.4	98.6
10,000	90	89.6	89.2	88.6	91.4	88.4	91.2	89.2	88.0	90.2	91.0	86.6	93.2	91.2
10,000	95	93.6	94.4	93.8	95.4	93.6	95.8	94.6	94.6	94.8	96.2	93.6	97.4	95.2
10,000	99	98.2	98.8	97.6	99.2	98.0	99.0	98.6	99.2	99.0	99.2	99.0	99.2	98.6

Obs.	CI	τ_{12}	τ_{13}	τ_{14}	τ_{23}	τ_{24}	τ_{33}	τ_{34}	τ_{44}	ω_{12}	ω_{13}	ω_{14}	ω_{23}	ω_{24}	ω_{33}	ω_{34}	ω_{44}
2,500	90	90.2	90.4	89.4	89.6	87.2	89.8	89.6	90.0	88.2	88.4	90.4	88.6	89.0	90.4	88.4	88.6
2,500	95	94.4	94.4	95.4	94.4	94.2	95.2	94.2	94.8	92.8	93.6	95.0	93.6	94.2	94.8	94.0	93.6
2,500	99	99.2	99.0	99.0	99.4	99.2	99.6	98.0	99.2	98.2	98.2	98.8	97.6	99.2	98.4	98.2	98.2
5,000	90	90.8	89.2	89.2	87.8	91.6	90.0	90.0	90.2	87.6	89.6	90.2	88.2	90.0	89.2	90.6	88.4
5,000	95	95.6	94.8	93.4	93.2	96.8	95.2	95.4	94.2	93.6	93.4	95.4	94.2	95.2	94.4	93.8	92.4
5,000	99	99.6	98.8	99.2	98.2	99.4	99.2	98.8	99.0	98.2	98.6	99.2	98.6	99.4	98.4	98.6	98.6
7,500	90	91.8	92.4	89.8	89.4	90.6	90.8	88.4	88.4	89.4	89.8	89.2	90.8	88.0	91.0	91.0	90.4
7,500	95	96.4	96.0	94.0	96.0	95.8	94.8	94.6	93.2	94.0	96.2	94.2	94.8	95.0	95.8	95.6	94.8
7,500	99	99.4	99.2	99.4	99.4	99.2	99.2	98.2	98.8	98.8	99.2	99.0	99.8	99.0	98.8	99.2	98.4
10,000	90	89.6	88.8	89.0	92.6	93.8	89.0	92.4	91.4	90.4	89.4	92.2	90.8	88.8	89.6	90.6	91.8
10,000	95	94.6	96.0	94.0	96.6	97.0	95.4	96.8	95.8	95.2	94.8	97.0	95.6	94.8	93.4	94.8	95.0
10,000	99	98.6	99.4	99.0	99.6	99.4	99.2	99.6	99.4	99.2	99.0	99.4	98.6	99.4	97.4	98.4	98.8

Table 4.15: Coverage percentages using rs536422311 with 4 components, of which 3 were binary, for 18 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	β_{46}	σ_{12}	σ_{13}	σ_{14}	σ_{23}	σ_{24}	σ_{34}	σ_{44}
2,500	90	85.4	91.0	90.2	89.0	92.4	92.6	91.8	87.8	88.4	89.2	89.4	90.4
2,500	95	93.6	96.0	94.6	95.0	97.0	96.2	96.0	93.4	93.8	94.6	93.6	95.2
2,500	99	98.2	99.8	98.0	98.4	100.0	100.0	98.6	98.8	98.4	99.2	99.2	98.8
5,000	90	88.8	89.6	91.2	91.6	89.8	88.8	86.8	89.8	88.0	87.6	88.2	89.6
5,000	95	92.6	95.6	96.4	95.8	93.8	95.2	92.8	95.4	94.6	92.4	94.2	94.8
5,000	99	98.4	99.6	99.2	99.4	99.2	98.8	98.6	99.0	98.6	99.0	98.8	99.2
7,500	90	89.0	88.8	90.0	91.0	91.8	89.0	91.0	91.2	90.6	89.4	88.8	88.8
7,500	95	95.6	94.6	94.4	95.4	95.6	94.8	95.6	95.6	95.6	94.4	94.0	95.2
7,500	99	99.0	97.6	98.8	99.2	99.0	99.2	98.8	99.2	99.2	99.0	99.6	98.8
10,000	90	89.6	90.6	90.4	88.2	88.2	91.8	89.2	89.2	89.6	87.2	89.2	89.2
10,000	95	94.6	96.2	95.4	93.6	95.0	96.2	94.6	94.4	95.0	92.8	93.2	95.0
10,000	99	99.0	99.2	99.0	98.0	98.2	98.8	99.0	99.4	99.2	97.6	98.2	99.2

Obs.	CI	τ_{12}	τ_{13}	τ_{14}	τ_{23}	τ_{24}	τ_{34}	τ_{44}	ω_{12}	ω_{13}	ω_{14}	ω_{23}	ω_{24}	ω_{34}	ω_{44}
2,500	90	91.2	91.4	89.4	90.8	89.2	89.8	89.8	89.2	88.8	90.2	87.4	86.6	88.0	91.2
2,500	95	96.0	95.6	93.4	95.4	93.8	94.0	95.2	94.0	93.6	94.8	92.2	93.4	94.0	95.8
2,500	99	99.6	98.8	98.8	99.0	98.8	99.0	99.4	98.2	98.2	99.2	98.6	98.6	98.6	98.6
5,000	90	88.2	91.2	89.6	90.2	89.2	88.4	90.8	88.8	91.0	90.2	89.2	90.0	89.8	87.0
5,000	95	94.6	96.2	93.8	95.8	95.0	93.4	95.2	94.2	95.8	95.6	95.0	94.6	95.2	93.2
5,000	99	99.0	99.6	98.8	98.6	98.8	99.0	99.0	99.4	99.4	99.2	99.0	99.0	98.6	98.8
7,500	90	90.6	89.8	92.6	88.2	87.6	90.4	90.2	88.0	90.8	90.8	90.4	92.0	91.2	89.2
7,500	95	95.4	95.0	96.4	93.6	93.6	95.6	95.8	94.0	96.8	95.4	95.4	95.8	95.0	94.2
7,500	99	98.8	99.6	100.0	99.4	98.4	99.0	99.8	98.2	99.2	99.4	98.6	99.2	99.2	98.0
10,000	90	89.0	90.8	89.0	89.6	90.8	88.8	89.8	88.4	90.6	88.4	90.2	91.2	92.4	88.2
10,000	95	94.2	94.8	93.8	93.8	94.6	94.6	95.0	93.8	94.8	94.6	95.0	95.8	95.4	93.2
10,000	99	98.8	99.0	98.6	99.0	98.4	98.8	99.8	99.0	99.0	99.0	99.0	99.0	99.4	98.6

Table 4.16: Coverage percentages using rs540691877 with 2 components, of which 1 was binary, for 6 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	σ_{12}	σ_{22}	τ_{12}	τ_{22}	ω_{12}	ω_{22}
2,500	90	88.0	89.4	90.4	87.6	91.4	89.8	90.4	91.2	87.8
2,500	95	94.4	95.2	95.6	93.6	95.6	94.0	94.6	95.2	93.0
2,500	99	99.0	98.8	99.2	98.2	99.2	99.0	98.4	98.2	98.4
5,000	90	89.8	89.0	91.0	88.4	91.0	91.6	93.2	90.8	91.4
5,000	95	93.8	93.0	96.0	94.0	94.6	94.4	96.2	95.2	94.6
5,000	99	98.8	98.6	99.4	98.2	98.6	99.4	99.6	98.8	98.8
7,500	90	88.6	88.2	88.6	88.6	89.4	91.0	89.8	90.4	89.2
7,500	95	93.6	93.2	94.4	94.6	94.6	95.0	94.6	95.6	93.2
7,500	99	99.0	99.2	98.8	98.6	98.8	98.8	99.0	98.6	98.2
10,000	90	89.2	89.4	90.8	89.2	88.6	90.8	89.0	90.2	88.0
10,000	95	95.0	92.2	95.8	94.0	93.8	95.8	93.6	94.2	93.6
10,000	99	98.8	98.8	99.0	98.8	99.2	99.0	98.6	98.0	97.8

Table 4.17: Coverage percentages using rs540691877 with 3 components, of which 1 was binary, for 13 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	σ_{12}	σ_{13}	σ_{22}	σ_{23}	σ_{33}	τ_{12}	τ_{13}	τ_{22}	τ_{23}	τ_{33}	ω_{12}	ω_{13}	ω_{22}	ω_{23}	ω_{33}
2,500	90	87.6	89.4	92.2	90.6	91.2	90.0	91.8	90.2	89.0	87.2	90.0	91.6	89.6	88.8	89.6	91.0	89.6	88.8	90.0
2,500	95	94.4	94.0	96.2	94.4	94.6	95.2	95.4	94.4	95.0	94.6	94.0	95.6	95.0	94.2	93.8	96.0	94.0	93.6	94.8
2,500	99	98.2	98.2	99.6	98.6	97.8	99.2	97.6	98.6	98.0	98.6	99.2	99.0	99.0	98.6	98.4	98.8	98.8	99.0	99.6
5,000	90	88.6	89.0	90.6	88.8	87.6	90.6	90.4	89.6	89.2	90.8	89.6	88.8	89.2	91.0	88.8	89.8	91.2	91.2	89.0
5,000	95	95.6	94.8	95.8	94.6	93.4	94.8	95.4	94.0	95.4	95.2	95.8	93.6	95.2	97.4	93.6	95.6	96.2	95.6	95.0
5,000	99	99.4	99.6	99.2	98.8	99.0	99.4	98.2	98.8	98.6	99.0	99.6	98.4	98.8	99.8	99.0	98.8	99.4	99.6	98.8
7,500	90	86.2	91.0	90.4	87.6	91.4	91.2	88.2	89.0	90.2	90.2	91.4	91.2	90.0	87.2	87.2	88.2	91.4	88.2	89.4
7,500	95	93.2	94.8	95.6	92.6	95.6	95.2	94.0	93.8	95.6	95.8	95.6	95.4	95.8	93.6	94.0	94.8	95.0	93.6	94.8
7,500	99	99.0	99.6	99.2	98.6	98.8	99.0	98.6	98.6	98.4	98.6	99.6	99.2	99.2	98.8	98.8	99.2	99.6	98.4	98.2
10,000	90	90.8	91.2	89.2	90.2	91.2	89.6	93.0	90.6	91.6	91.0	90.2	89.0	89.8	92.0	91.0	90.4	91.2	88.6	89.0
10,000	95	96.4	96.4	94.4	95.4	96.0	95.6	96.6	95.6	96.0	94.2	95.0	94.0	94.4	97.8	95.0	94.8	95.6	95.2	96.4
10,000	99	99.4	99.6	99.2	98.8	99.2	98.8	99.6	98.8	99.4	98.6	99.0	99.0	98.8	99.6	99.4	99.4	98.8	99.4	99.0

Table 4.18: Coverage percentages using rs540691877 with 3 components, of which 2 were binary, for 11 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	σ_{12}	σ_{13}	σ_{23}	σ_{33}	τ_{12}	τ_{13}	τ_{23}	τ_{33}	ω_{12}	ω_{13}	ω_{23}	ω_{33}
2,500	90	88.4	90.4	88.8	89.0	88.2	90.0	87.4	90.2	89.4	89.4	92.2	90.2	91.2	91.0	91.4	92.2
2,500	95	94.6	95.0	94.8	94.0	93.6	93.8	93.4	94.2	93.2	94.6	95.6	93.8	95.8	95.2	95.6	94.6
2,500	99	98.2	98.6	98.6	98.4	98.8	98.4	98.0	98.6	99.0	98.8	98.6	98.6	99.2	99.0	98.8	98.6
5,000	90	90.2	89.0	90.0	89.2	92.4	88.8	87.2	89.6	90.8	90.2	89.2	90.4	90.8	90.0	91.6	89.8
5,000	95	94.4	95.4	94.8	95.4	96.8	94.6	94.8	95.0	94.2	93.8	93.6	94.4	95.0	96.6	95.8	94.6
5,000	99	99.0	99.0	99.2	99.6	99.4	99.4	98.6	98.0	98.8	99.2	98.6	98.8	99.4	99.2	98.4	98.6
7,500	90	89.8	91.6	91.0	91.6	88.4	91.4	91.0	92.4	88.8	87.4	88.6	90.0	89.2	88.0	87.6	90.4
7,500	95	95.8	96.2	95.4	95.4	95.2	96.4	94.0	96.0	93.4	93.6	95.6	96.4	93.6	94.2	93.4	95.6
7,500	99	99.4	99.0	99.2	99.4	99.0	98.8	99.0	98.6	98.6	97.4	99.2	99.6	99.2	98.6	99.4	98.8
10,000	90	89.6	89.0	90.8	88.8	90.6	88.8	88.4	90.0	89.2	88.2	90.4	90.0	91.2	89.6	93.2	88.8
10,000	95	95.0	93.8	95.6	94.2	94.4	93.4	95.0	96.2	94.4	94.4	94.6	96.0	95.8	93.8	96.8	95.8
10,000	99	99.6	99.8	98.8	98.4	98.4	98.4	98.2	99.4	99.6	98.6	98.8	99.2	98.8	99.6	100.0	99.0

Table 4.19: Coverage percentages using rs540691877 with 4 components, of which 1 was binary, for 22 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	β_{46}	σ_{12}	σ_{13}	σ_{14}	σ_{22}	σ_{23}	σ_{24}	σ_{33}	σ_{34}	σ_{44}
2,500	90	87.4	90.0	91.4	89.2	91.2	90.2	89.0	89.6	90.4	90.0	90.6	89.4	89.4	89.4
2,500	95	92.6	94.0	95.6	94.8	96.4	95.0	94.0	95.2	94.8	95.0	94.8	95.4	94.4	95.0
2,500	99	98.4	99.0	98.4	99.0	99.4	99.0	98.6	99.0	99.0	99.6	99.0	98.0	98.4	98.2
5,000	90	88.2	90.8	89.8	87.2	89.0	89.2	89.4	89.0	91.0	87.6	92.2	90.2	89.2	89.6
5,000	95	93.0	95.4	95.8	92.8	95.6	93.4	93.6	94.8	96.6	93.2	97.0	95.4	95.0	94.8
5,000	99	98.6	99.4	99.6	99.2	99.2	99.2	98.2	98.8	99.2	98.2	99.4	99.0	99.2	98.8
7,500	90	86.6	88.8	88.8	89.0	89.2	89.2	89.0	89.4	92.8	88.0	87.6	90.0	88.6	88.2
7,500	95	92.0	93.0	95.0	93.0	94.8	94.0	95.8	95.4	96.8	93.4	93.4	95.2	93.4	93.6
7,500	99	98.8	97.8	99.4	98.0	98.8	99.0	99.8	99.0	99.4	98.8	98.2	98.8	98.8	99.2
10,000	90	88.4	90.6	89.6	87.6	90.0	91.8	92.2	91.4	91.0	91.8	89.2	89.0	90.2	90.4
10,000	95	94.8	95.8	94.6	92.8	95.2	95.4	96.0	95.6	97.2	95.8	94.8	93.6	95.6	95.4
10,000	99	99.4	99.0	98.4	98.4	98.6	98.6	99.6	99.0	99.2	99.8	98.6	98.6	99.0	99.8

Obs.	CI	τ_{12}	τ_{13}	τ_{14}	τ_{22}	τ_{23}	τ_{24}	τ_{33}	τ_{34}	τ_{44}	ω_{12}	ω_{13}	ω_{14}	ω_{22}	ω_{23}	ω_{24}	ω_{33}	ω_{34}	ω_{44}
2,500	90	89.4	90.8	89.2	89.2	89.6	90.2	92.0	87.2	91.4	91.4	89.6	88.8	87.8	88.6	90.4	88.6	91.2	91.4
2,500	95	95.8	95.0	94.4	94.0	95.8	95.6	95.4	95.2	96.6	96.8	94.6	94.0	93.4	94.8	94.8	94.4	95.6	95.6
2,500	99	99.4	98.4	98.0	98.6	99.8	99.6	99.4	99.2	99.6	99.0	98.6	98.6	98.8	99.2	98.4	98.8	99.0	99.0
5,000	90	90.8	89.8	91.8	88.6	89.4	90.8	90.6	91.2	86.4	90.6	91.4	89.4	89.6	88.2	89.6	90.4	88.2	88.4
5,000	95	95.2	95.2	96.6	95.4	94.8	95.4	95.0	95.6	93.4	96.0	96.2	94.4	95.6	94.8	94.2	95.4	95.0	94.4
5,000	99	99.0	99.8	99.6	98.2	98.6	99.6	99.4	98.8	98.6	99.4	99.4	98.8	98.4	99.6	98.6	99.2	99.0	98.4
7,500	90	90.4	90.0	90.0	90.6	89.8	90.2	87.6	87.4	91.8	89.4	90.4	90.8	91.0	88.2	87.0	90.2	89.0	90.0
7,500	95	95.4	93.6	95.0	95.8	94.4	95.0	92.8	93.8	97.4	94.6	95.2	94.8	95.2	94.6	94.0	95.4	95.6	94.8
7,500	99	98.4	98.4	99.4	98.6	99.2	99.4	98.0	98.2	99.4	99.6	98.2	98.4	98.8	98.6	98.6	99.0	99.4	98.6
10,000	90	90.6	90.0	91.0	90.2	89.8	89.0	89.4	92.4	88.4	87.6	88.0	91.0	91.0	91.6	88.2	90.2	88.6	88.4
10,000	95	96.6	95.4	94.6	96.4	94.2	93.2	94.4	95.8	93.8	94.6	94.6	95.8	95.6	96.2	94.4	95.4	94.4	92.6
10,000	99	99.0	98.8	98.8	99.4	99.0	98.2	98.8	99.4	98.4	99.0	98.4	99.4	98.8	99.2	99.2	99.8	99.6	98.6

Table 4.20: Coverage percentages using rs540691877 with 4 components, of which 2 were binary, for 20 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	β_{46}	σ_{12}	σ_{13}	σ_{14}	σ_{23}	σ_{24}	σ_{33}	σ_{34}	σ_{44}
2,500	90	86.2	92.8	88.4	87.4	90.2	89.0	91.4	91.2	87.8	91.0	89.0	91.8	87.2
2,500	95	93.4	96.0	92.6	93.4	94.8	93.2	96.2	95.6	93.2	94.4	94.0	95.6	93.4
2,500	99	98.4	98.2	99.0	99.2	99.6	98.6	98.8	98.8	97.8	99.0	98.0	98.8	98.6
5,000	90	90.8	88.4	90.0	91.4	88.6	89.6	90.2	88.6	91.8	90.8	90.2	90.6	91.8
5,000	95	95.2	94.4	96.4	96.2	94.4	94.0	93.8	95.0	96.0	94.8	97.0	94.6	95.6
5,000	99	98.6	99.4	99.2	98.8	98.4	99.6	98.0	99.2	98.6	98.0	99.4	99.0	98.8
7,500	90	86.4	90.2	89.2	91.6	90.2	90.2	89.4	88.2	90.2	90.0	91.4	90.0	92.0
7,500	95	92.8	94.4	94.2	97.0	94.6	94.6	94.4	94.6	95.2	94.8	95.2	94.6	96.2
7,500	99	98.6	98.8	99.0	99.0	99.4	99.0	99.2	98.0	98.6	98.8	98.0	97.6	99.0
10,000	90	87.8	88.6	89.4	89.6	90.8	90.8	89.4	89.4	90.2	90.6	89.6	90.6	89.6
10,000	95	94.8	95.2	95.2	94.6	95.8	94.2	95.0	94.4	94.4	96.0	95.0	95.0	95.4
10,000	99	99.4	98.6	99.0	98.6	99.2	99.0	98.4	99.4	98.6	99.0	98.4	99.2	98.4

Obs.	CI	τ_{12}	τ_{13}	τ_{14}	τ_{23}	τ_{24}	τ_{33}	τ_{34}	τ_{44}	ω_{12}	ω_{13}	ω_{14}	ω_{23}	ω_{24}	ω_{33}	ω_{34}	ω_{44}
2,500	90	88.0	91.4	90.2	89.0	90.2	89.0	90.0	89.2	91.2	89.2	88.6	89.6	89.8	91.4	89.4	88.6
2,500	95	95.0	95.8	95.4	93.8	95.2	93.8	94.6	94.2	94.8	95.0	94.2	93.6	94.0	95.4	94.0	93.2
2,500	99	98.2	99.0	97.8	98.8	98.6	98.6	99.2	97.6	99.2	99.2	98.4	99.2	99.2	99.8	99.2	97.6
5,000	90	87.4	89.2	87.2	92.2	86.6	88.4	89.2	90.6	91.4	88.6	89.4	89.2	87.6	91.2	90.4	90.4
5,000	95	94.0	95.2	93.8	95.0	95.2	95.2	94.8	95.2	95.4	95.6	94.4	94.2	93.6	96.0	94.2	94.0
5,000	99	99.8	99.0	99.2	98.8	99.0	98.6	99.6	99.0	99.0	99.0	99.2	99.0	98.8	98.8	99.2	98.8
7,500	90	89.8	88.6	89.6	90.2	88.8	89.6	89.4	89.8	87.8	90.2	89.0	91.6	89.4	90.2	90.8	90.6
7,500	95	93.6	93.6	95.6	94.6	92.8	94.2	95.4	96.0	93.4	94.6	94.6	96.2	95.6	94.6	94.4	95.6
7,500	99	99.2	99.2	99.2	98.6	98.2	97.4	99.2	98.8	98.2	99.4	98.8	99.2	99.2	98.8	97.8	99.6
10,000	90	91.4	88.8	90.4	91.2	87.8	89.2	88.6	90.6	88.8	90.2	88.2	89.2	89.0	90.8	89.6	91.4
10,000	95	94.8	94.6	95.2	96.6	94.4	93.4	94.6	94.6	94.4	95.0	93.4	95.6	93.8	95.2	95.0	96.6
10,000	99	98.8	99.0	98.8	99.0	99.0	99.0	99.6	99.8	99.2	99.0	98.8	99.4	99.4	99.0	98.6	98.6

Table 4.21: Coverage percentages using rs540691877 with 4 components, of which 3 were binary, for 18 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	β_{46}	σ_{12}	σ_{13}	σ_{14}	σ_{23}	σ_{24}	σ_{34}	σ_{44}
2,500	90	87.8	89.8	88.4	91.4	90.2	89.6	89.8	92.2	89.6	90.8	87.6	91.4
2,500	95	92.8	95.6	94.2	96.0	94.2	94.6	95.4	96.8	93.6	94.6	93.2	95.4
2,500	99	98.2	99.0	99.2	99.2	98.6	98.8	99.0	99.2	98.4	98.4	98.2	98.8
5,000	90	89.4	89.6	91.2	89.2	92.8	89.8	89.2	90.0	88.2	91.6	89.0	89.8
5,000	95	93.6	95.0	95.6	95.4	97.2	94.6	94.4	94.8	92.8	95.6	94.6	95.6
5,000	99	97.6	99.2	99.6	99.4	99.2	99.4	99.0	99.2	98.2	99.2	99.6	99.2
7,500	90	90.2	91.2	92.2	90.2	92.0	89.8	90.0	90.2	90.8	90.6	91.0	87.6
7,500	95	95.2	95.8	96.4	93.8	95.4	94.8	94.0	95.2	95.4	95.2	95.0	96.0
7,500	99	99.2	100.0	99.2	99.4	98.8	98.6	97.4	99.4	98.8	99.6	98.6	99.6
10,000	90	89.6	90.4	86.8	90.4	89.2	89.2	90.4	89.4	92.4	88.6	88.8	90.0
10,000	95	95.8	95.6	93.2	95.6	95.8	94.8	95.6	95.0	96.2	94.6	94.4	96.0
10,000	99	99.0	99.0	98.8	98.2	99.4	98.4	99.4	99.4	99.0	99.4	99.6	99.4

Obs.	CI	τ_{12}	τ_{13}	τ_{14}	τ_{23}	τ_{24}	τ_{34}	τ_{44}	ω_{12}	ω_{13}	ω_{14}	ω_{23}	ω_{24}	ω_{34}	ω_{44}
2,500	90	90.4	88.6	92.4	90.4	89.4	88.2	88.2	90.4	90.4	91.4	89.2	91.8	90.6	91.8
2,500	95	94.4	94.6	97.2	95.4	95.4	93.6	94.0	95.0	95.6	95.8	94.8	96.0	95.2	96.2
2,500	99	98.4	98.6	99.4	99.0	99.2	98.6	98.6	98.6	99.0	99.4	98.0	99.2	99.6	99.4
5,000	90	91.0	89.4	89.4	88.4	90.4	91.4	91.0	91.0	88.4	86.0	88.4	89.6	91.8	90.2
5,000	95	95.2	93.6	95.0	93.0	96.2	94.0	95.0	94.8	93.6	94.0	93.8	94.4	96.8	95.0
5,000	99	99.2	99.0	99.0	98.6	99.6	99.4	98.8	99.6	99.4	98.4	98.0	99.4	99.4	99.0
7,500	90	92.2	90.0	88.4	88.0	89.6	91.4	91.0	90.2	91.4	90.4	90.8	91.6	90.8	89.6
7,500	95	96.8	94.8	95.0	95.2	94.4	96.0	96.2	95.2	95.2	95.6	94.2	97.2	96.6	94.0
7,500	99	99.8	98.8	99.0	99.2	98.8	99.6	99.2	99.0	99.4	99.4	98.4	98.8	99.4	99.2
10,000	90	91.0	89.6	89.8	90.8	90.4	89.0	91.0	89.0	90.8	89.4	91.6	90.6	89.2	89.2
10,000	95	95.8	94.6	93.4	94.6	95.4	94.4	93.8	93.6	94.2	93.8	97.0	94.6	94.8	94.8
10,000	99	99.2	99.8	98.8	99.4	99.6	99.2	98.8	98.6	98.8	98.8	99.2	98.6	98.8	98.8

Table 4.22: Coverage percentages using rs577080509 with 2 components, of which 1 was binary, for 6 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	σ_{12}	σ_{22}	τ_{12}	τ_{22}	ω_{12}	ω_{22}
2,500	90	91.0	89.8	91.2	90.0	91.4	91.2	89.6	91.8	89.6
2,500	95	95.4	94.0	94.8	94.2	95.2	95.4	95.6	96.2	94.2
2,500	99	98.8	98.6	97.6	99.2	98.6	98.8	98.6	99.4	99.4
5,000	90	88.8	90.2	91.2	89.6	90.6	89.8	89.6	88.8	89.4
5,000	95	93.4	95.4	95.4	95.2	94.2	94.6	94.6	93.8	95.6
5,000	99	98.8	99.6	99.0	98.8	99.2	98.8	98.2	99.0	98.8
7,500	90	89.2	88.6	89.8	90.0	89.0	88.8	92.0	91.2	91.6
7,500	95	96.0	94.4	95.4	96.2	94.0	93.6	96.0	94.6	96.6
7,500	99	98.6	98.8	99.4	98.6	98.0	98.8	99.4	98.8	99.2
10,000	90	89.0	89.0	90.4	89.6	92.0	88.8	88.4	89.2	89.8
10,000	95	95.0	94.0	94.6	94.6	96.0	93.8	93.8	94.0	94.4
10,000	99	99.2	98.4	99.2	98.6	99.4	99.4	98.8	98.0	99.0

Table 4.23: Coverage percentages using rs577080509 with 3 components, of which 1 was binary, for 13 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	σ_{12}	σ_{13}	σ_{22}	σ_{23}	σ_{33}	τ_{12}	τ_{13}	τ_{22}	τ_{23}	τ_{33}	ω_{12}	ω_{13}	ω_{22}	ω_{23}	ω_{33}
2,500	90	87.4	89.4	90.2	91.4	90.2	88.6	88.6	88.6	89.0	90.0	86.8	90.6	89.4	90.2	87.4	89.4	91.2	88.8	92.0
2,500	95	93.2	96.6	93.0	95.0	93.4	92.4	94.2	93.4	95.0	94.4	94.2	94.8	94.2	93.4	94.2	95.0	95.2	94.4	95.4
2,500	99	97.6	99.4	99.0	99.2	98.4	98.6	98.0	97.8	98.0	99.2	98.2	98.8	99.2	98.4	99.0	98.8	99.0	98.8	99.2
5,000	90	89.2	87.0	89.0	88.8	91.6	89.0	90.8	90.6	88.6	90.2	89.8	90.4	87.2	92.8	88.8	90.2	91.8	90.4	90.6
5,000	95	94.0	93.4	93.8	93.6	95.2	94.2	95.4	94.2	95.4	94.8	94.6	95.6	93.6	96.2	93.4	94.6	94.8	94.8	96.0
5,000	99	98.8	99.0	99.0	99.6	98.8	98.8	98.6	98.6	99.4	98.4	98.8	99.2	99.2	99.8	98.0	98.6	99.6	98.2	99.4
7,500	90	87.6	89.8	87.6	90.4	88.0	92.4	89.8	91.4	89.6	90.6	89.8	90.6	87.8	88.4	91.4	88.4	90.0	91.2	88.4
7,500	95	95.2	96.4	95.2	95.2	94.4	95.0	93.6	95.4	95.6	95.0	95.4	95.6	92.8	95.6	94.8	95.8	96.0	95.4	93.4
7,500	99	99.2	99.6	99.6	98.4	98.8	98.2	98.2	99.2	98.8	98.6	99.2	99.0	98.0	98.2	98.4	98.8	99.2	99.2	98.0
10,000	90	89.8	88.2	88.6	89.2	90.0	90.6	90.2	89.8	91.2	88.0	88.0	88.6	90.4	89.6	89.2	90.4	89.4	89.2	89.8
10,000	95	95.8	93.6	93.8	94.2	95.8	95.6	93.4	94.2	95.6	93.2	93.4	95.6	96.0	94.8	95.8	94.4	94.6	94.4	94.8
10,000	99	99.6	98.8	98.6	99.0	99.0	99.4	98.8	99.2	98.8	99.0	99.4	99.0	99.4	99.0	99.2	99.4	99.4	98.6	99.8

Table 4.24: Coverage percentages using rs577080509 with 3 components, of which 2 were binary, for 11 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	σ_{12}	σ_{13}	σ_{23}	σ_{33}	τ_{12}	τ_{13}	τ_{23}	τ_{33}	ω_{12}	ω_{13}	ω_{23}	ω_{33}
2,500	90	87.0	92.2	90.0	88.4	88.4	91.8	91.4	89.4	88.2	87.8	87.6	90.0	91.4	86.8	89.2	88.8
2,500	95	91.6	96.6	95.2	94.6	95.4	96.8	94.6	94.2	94.0	94.2	93.6	95.4	95.2	93.2	95.2	94.4
2,500	99	97.6	99.4	99.0	98.8	99.2	98.6	97.2	98.8	98.8	98.8	98.8	99.2	99.4	99.2	99.0	98.6
5,000	90	91.6	89.6	89.4	90.6	89.2	87.2	89.8	89.4	90.2	89.6	91.2	89.4	91.8	91.6	89.0	91.0
5,000	95	94.2	94.6	94.8	94.2	95.0	93.6	94.2	94.2	95.0	95.4	96.0	95.6	95.6	96.2	93.6	95.6
5,000	99	98.6	99.4	99.6	99.4	98.2	97.4	98.4	98.8	99.6	99.8	99.4	98.2	98.8	98.6	99.2	99.2
7,500	90	90.8	90.8	89.6	89.2	88.8	87.2	93.0	89.6	88.0	90.6	87.2	87.4	90.4	91.0	89.0	90.0
7,500	95	95.4	95.4	95.2	95.0	93.0	92.8	97.2	95.2	93.6	94.4	93.6	94.4	94.8	95.4	93.6	95.8
7,500	99	98.4	98.8	99.2	99.0	98.2	98.8	99.8	99.4	99.2	98.6	98.6	99.2	99.0	99.0	98.4	99.4
10,000	90	89.2	89.2	90.6	88.6	90.4	87.8	88.0	89.8	92.0	90.6	92.4	91.2	89.2	93.0	90.6	88.0
10,000	95	93.6	95.0	96.0	95.4	96.2	95.0	93.8	94.8	95.6	94.6	95.8	95.2	94.0	96.4	95.2	94.2
10,000	99	99.4	99.0	99.2	98.8	98.8	98.8	98.8	98.6	99.0	99.2	99.0	99.6	99.2	99.2	99.0	99.2

Table 4.25: Coverage percentages using rs577080509 with 4 components, of which 1 was binary, for 22 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	β_{46}	σ_{12}	σ_{13}	σ_{14}	σ_{22}	σ_{23}	σ_{24}	σ_{33}	σ_{34}	σ_{44}
2,500	90	84.4	90.2	88.8	91.4	89.6	88.6	89.4	90.6	88.6	91.0	89.4	90.6	90.0	89.8
2,500	95	90.4	94.8	93.6	96.2	95.0	95.0	93.8	95.8	94.4	95.2	94.2	94.2	94.0	94.6
2,500	99	98.0	99.2	98.4	99.4	99.2	99.2	99.0	99.8	98.6	98.6	98.6	99.2	98.8	97.8
5,000	90	84.8	90.0	88.0	90.6	89.6	89.6	87.4	91.8	90.0	90.8	90.2	89.8	90.8	87.8
5,000	95	92.0	94.6	94.4	94.2	95.4	93.8	93.6	96.4	94.0	94.6	95.0	95.8	94.8	92.6
5,000	99	98.2	98.0	98.2	97.8	98.8	98.0	98.2	99.0	98.4	98.6	98.2	99.0	99.2	97.6
7,500	90	90.4	92.8	89.2	89.8	87.8	90.8	90.4	90.0	89.8	91.2	88.8	91.2	91.8	89.8
7,500	95	94.0	96.4	94.8	95.2	93.2	94.8	95.6	95.2	95.2	97.0	94.2	95.2	95.6	95.0
7,500	99	98.2	99.4	99.0	99.2	99.2	98.4	99.2	99.0	99.2	99.4	98.8	99.0	98.6	98.4
10,000	90	90.2	89.4	90.2	91.0	91.2	91.0	92.8	90.6	92.6	86.8	90.2	89.6	90.8	89.2
10,000	95	95.6	94.8	95.6	95.4	95.4	95.4	97.4	95.4	96.4	94.2	94.8	95.0	96.0	94.8
10,000	99	99.4	98.8	98.8	99.2	99.2	99.4	99.6	98.4	99.2	99.0	99.4	99.6	99.4	99.4

Obs.	CI	τ_{12}	τ_{13}	τ_{14}	τ_{22}	τ_{23}	τ_{24}	τ_{33}	τ_{34}	τ_{44}	ω_{12}	ω_{13}	ω_{14}	ω_{22}	ω_{23}	ω_{24}	ω_{33}	ω_{34}	ω_{44}
2,500	90	89.0	91.8	90.0	92.2	89.6	88.2	90.4	90.6	87.0	90.0	90.4	88.6	88.4	90.2	89.2	92.0	88.4	90.8
2,500	95	94.8	96.8	94.8	95.6	94.8	93.2	93.4	94.4	95.4	94.4	95.0	93.6	93.4	94.4	93.4	95.2	94.0	94.2
2,500	99	98.6	99.2	99.2	98.6	99.4	99.4	98.6	99.0	98.8	98.4	99.0	99.4	98.6	98.6	97.6	99.8	98.8	98.8
5,000	90	92.0	88.4	89.4	88.6	89.8	89.4	89.6	90.4	90.4	89.2	88.6	90.6	90.6	88.8	88.4	90.6	92.8	90.4
5,000	95	96.2	92.4	95.2	92.4	94.0	95.0	93.6	94.2	94.6	95.2	93.4	95.0	93.8	94.6	95.2	95.4	96.2	95.0
5,000	99	99.4	98.6	99.0	98.6	99.2	99.2	98.4	99.4	98.4	99.6	99.0	99.2	99.0	99.0	99.0	99.0	99.6	98.6
7,500	90	90.2	89.6	91.6	88.4	89.6	89.8	88.6	91.6	89.8	91.2	90.8	89.6	91.4	90.2	90.6	91.8	91.4	90.0
7,500	95	94.6	96.0	95.6	93.0	95.4	95.0	93.4	95.0	95.2	95.6	95.6	94.4	96.0	95.0	95.8	95.6	95.0	94.4
7,500	99	99.0	98.6	99.6	98.2	99.4	99.4	99.0	99.4	99.2	99.0	99.2	98.6	99.2	99.0	99.2	99.2	98.8	98.6
10,000	90	90.2	89.0	92.0	87.8	90.4	91.2	90.0	89.4	88.6	88.8	87.6	92.0	89.0	90.0	90.4	90.8	91.6	90.8
10,000	95	93.8	95.4	96.2	95.2	96.0	95.8	96.2	93.8	94.4	94.8	94.2	95.8	95.2	95.0	95.6	95.8	95.6	94.2
10,000	99	98.8	98.8	99.2	98.4	98.6	99.0	99.2	98.2	98.4	98.4	99.2	98.2	98.8	99.6	98.8	99.2	99.4	98.0

Table 4.26: Coverage percentages using rs577080509 with 4 components, of which 2 were binary, for 20 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	β_{46}	σ_{12}	σ_{13}	σ_{14}	σ_{23}	σ_{24}	σ_{33}	σ_{34}	σ_{44}
2,500	90	88.2	90.2	90.6	92.4	91.6	89.4	89.2	91.0	90.8	88.8	92.6	88.8	89.2
2,500	95	93.0	95.2	95.8	96.2	96.4	94.2	93.0	94.8	96.0	93.8	96.4	94.8	93.6
2,500	99	98.2	99.4	99.2	99.2	99.4	99.0	98.0	98.4	99.6	99.0	98.4	99.0	98.4
5,000	90	88.8	89.8	92.0	88.8	91.6	89.2	92.2	90.2	89.8	89.0	89.0	87.8	91.0
5,000	95	92.8	97.0	96.2	93.8	95.8	94.4	95.4	94.4	94.2	95.0	93.0	93.8	96.6
5,000	99	98.2	99.0	99.2	98.4	99.2	99.4	99.4	99.4	98.0	98.8	98.2	99.0	99.4
7,500	90	88.4	91.4	88.6	89.2	89.2	91.4	89.8	90.6	90.4	88.6	90.0	90.8	88.0
7,500	95	94.4	96.4	93.8	94.6	95.6	95.6	95.4	96.4	94.8	93.4	94.2	95.0	94.4
7,500	99	98.8	99.2	99.0	98.8	99.2	99.8	99.2	99.2	99.0	98.4	98.6	98.6	99.0
10,000	90	88.4	90.2	91.6	90.2	92.0	90.6	90.2	89.4	93.0	90.8	88.6	89.0	90.4
10,000	95	95.0	95.4	94.8	94.8	95.8	96.0	95.0	95.6	96.4	95.0	93.4	93.2	95.0
10,000	99	99.4	99.6	99.0	99.4	99.2	99.0	99.2	99.6	99.6	99.2	98.4	99.2	99.2

Obs.	CI	τ_{12}	τ_{13}	τ_{14}	τ_{23}	τ_{24}	τ_{33}	τ_{34}	τ_{44}	ω_{12}	ω_{13}	ω_{14}	ω_{23}	ω_{24}	ω_{33}	ω_{34}	ω_{44}
2,500	90	91.8	88.2	89.4	90.6	89.8	89.4	93.4	90.0	90.2	89.8	90.6	88.6	90.0	92.0	88.8	91.6
2,500	95	95.0	93.4	95.2	95.6	95.2	95.0	97.2	95.2	95.0	95.8	95.6	93.6	96.2	95.8	94.6	96.0
2,500	99	99.2	99.0	98.8	99.6	99.2	99.0	99.2	98.8	99.2	98.0	99.0	98.8	99.2	99.0	98.6	99.0
5,000	90	90.6	87.6	90.8	87.8	90.6	91.2	90.2	89.8	90.2	89.8	87.8	89.2	89.4	89.2	88.8	89.2
5,000	95	95.6	94.2	95.8	94.2	95.0	96.4	94.6	94.4	95.8	95.2	93.2	94.2	94.6	93.6	93.6	93.8
5,000	99	99.2	98.0	99.0	99.2	98.8	99.2	99.2	98.4	99.6	100.0	97.4	99.4	98.4	98.6	98.6	99.0
7,500	90	89.6	90.4	90.6	88.8	89.0	89.4	88.2	91.4	88.6	92.0	89.0	90.4	90.4	89.8	91.2	91.0
7,500	95	95.0	95.4	94.0	94.0	93.0	94.8	93.6	94.8	94.2	94.8	93.4	94.8	95.6	95.8	95.4	96.0
7,500	99	99.4	99.0	98.6	98.6	99.0	99.4	99.4	99.2	98.0	98.6	98.4	98.8	99.4	97.8	99.0	99.4
10,000	90	89.4	89.6	88.6	88.8	91.2	90.8	89.4	87.8	90.4	91.6	90.8	89.4	90.6	89.2	89.6	90.2
10,000	95	95.2	94.6	94.2	93.8	95.4	96.2	95.8	93.0	95.2	97.4	95.6	94.8	96.4	95.6	95.4	95.8
10,000	99	98.4	98.6	99.0	98.8	99.2	99.2	98.4	98.6	98.8	99.2	99.4	99.0	99.0	99.4	99.0	99.4

Table 4.27: Coverage percentages using rs577080509 with 4 components, of which 3 were binary, for 18 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	β_{46}	σ_{12}	σ_{13}	σ_{14}	σ_{23}	σ_{24}	σ_{34}	σ_{44}
2,500	90	88.0	92.0	89.4	89.8	88.6	92.0	86.2	91.0	90.8	89.4	87.8	92.4
2,500	95	93.6	96.6	95.0	95.0	94.0	95.4	93.4	95.6	95.6	95.0	93.4	96.4
2,500	99	97.8	99.2	99.6	99.0	98.0	99.0	98.6	98.8	99.4	98.2	98.6	99.4
5,000	90	88.6	89.6	91.4	89.0	91.6	90.4	87.0	90.6	87.2	92.6	92.2	90.2
5,000	95	94.4	95.6	95.4	93.8	96.2	95.8	93.0	94.2	93.0	95.6	96.4	95.0
5,000	99	98.4	99.6	99.2	98.2	99.2	99.4	98.6	98.4	98.2	99.2	99.8	99.0
7,500	90	88.8	88.4	88.2	89.6	92.6	89.8	86.6	91.0	90.0	91.4	86.4	91.0
7,500	95	94.2	94.4	92.8	95.4	96.2	94.4	93.0	95.6	95.6	95.4	94.6	95.6
7,500	99	98.2	99.4	98.2	98.2	99.2	98.4	98.6	99.2	99.0	99.4	99.2	99.2
10,000	90	91.8	88.8	90.0	90.2	89.2	90.2	90.8	93.2	89.8	91.8	90.8	89.6
10,000	95	94.8	94.6	95.0	93.6	94.6	96.2	95.8	96.4	94.2	95.8	96.0	94.8
10,000	99	99.6	99.0	99.2	99.0	99.4	99.4	98.4	99.0	99.2	99.0	99.4	98.0

Obs.	CI	τ_{12}	τ_{13}	τ_{14}	τ_{23}	τ_{24}	τ_{34}	τ_{44}	ω_{12}	ω_{13}	ω_{14}	ω_{23}	ω_{24}	ω_{34}	ω_{44}
2,500	90	89.6	89.6	89.8	88.0	90.6	88.8	90.2	89.6	91.2	92.2	89.6	89.8	87.4	90.8
2,500	95	94.4	94.2	95.0	94.6	96.6	94.0	95.2	94.8	95.6	95.8	94.4	95.8	93.0	96.6
2,500	99	99.0	98.6	98.8	98.8	98.8	98.4	99.2	98.8	99.2	98.8	98.6	99.0	99.2	99.8
5,000	90	87.2	91.2	91.2	91.2	91.0	91.2	86.4	90.4	91.0	92.4	89.6	90.8	91.8	91.6
5,000	95	93.8	95.4	94.6	95.8	95.6	96.0	92.8	95.4	96.2	96.6	94.4	95.8	95.0	95.4
5,000	99	99.0	97.8	98.6	100.0	99.0	99.2	97.6	99.0	99.8	99.4	98.4	99.6	98.8	99.8
7,500	90	88.4	89.2	90.8	90.2	89.4	91.6	89.2	89.2	90.2	90.2	87.6	87.2	91.8	91.6
7,500	95	94.8	94.2	95.2	94.6	94.6	95.6	94.4	95.6	95.6	96.6	92.6	95.4	96.2	95.6
7,500	99	98.6	98.8	99.2	99.2	98.6	99.0	99.6	99.0	99.2	99.2	99.0	98.6	99.2	99.2
10,000	90	87.0	91.8	88.6	88.2	89.6	92.0	91.6	90.2	89.4	89.8	88.8	89.8	92.8	91.2
10,000	95	92.8	95.6	94.4	93.6	95.2	96.0	96.2	95.6	93.2	94.4	96.2	93.4	97.0	95.0
10,000	99	97.8	99.4	98.6	98.4	99.4	99.2	99.4	99.6	98.4	98.4	99.4	98.4	99.6	99.6

Table 4.28: Coverage percentages using rs71243915 with 2 components, of which 1 was binary, for 6 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	σ_{12}	σ_{22}	τ_{12}	τ_{22}	ω_{12}	ω_{22}
2,500	90	89.8	91.4	90.8	90.2	88.0	90.4	88.2	89.6	90.8
2,500	95	93.6	96.0	94.4	94.8	93.8	93.6	94.2	93.6	96.0
2,500	99	98.6	99.4	99.8	98.8	98.2	98.2	99.2	98.8	99.0
5,000	90	89.2	90.2	90.0	91.0	90.0	92.8	92.4	89.6	90.8
5,000	95	96.0	93.6	94.8	97.2	95.0	95.4	96.2	94.0	95.2
5,000	99	99.2	98.6	99.2	100.0	98.0	98.6	99.2	99.6	99.4
7,500	90	88.6	90.6	90.2	90.8	89.2	88.6	89.0	91.6	91.2
7,500	95	94.8	95.6	95.0	96.0	94.2	92.4	94.6	95.6	95.0
7,500	99	98.6	99.0	99.0	99.6	97.8	97.4	99.4	98.8	99.4
10,000	90	90.0	90.8	90.0	89.6	89.8	88.4	87.4	88.4	88.8
10,000	95	96.0	95.6	94.8	94.6	95.8	93.2	93.0	95.0	95.2
10,000	99	99.0	99.2	99.2	98.6	99.0	98.8	98.8	99.6	98.8

Table 4.29: Coverage percentages using rs71243915 with 3 components, of which 1 was binary, for 13 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	σ_{12}	σ_{13}	σ_{22}	σ_{23}	σ_{33}	τ_{12}	τ_{13}	τ_{22}	τ_{23}	τ_{33}	ω_{12}	ω_{13}	ω_{22}	ω_{23}	ω_{33}
2,500	90	88.6	87.0	91.0	89.0	89.2	90.0	89.8	90.6	90.8	91.0	89.8	90.6	90.0	92.0	88.8	90.2	89.2	90.4	90.8
2,500	95	93.6	92.8	95.2	95.6	93.8	92.8	94.4	94.8	94.4	95.2	95.6	95.0	95.0	95.2	94.0	94.2	95.0	96.4	94.8
2,500	99	98.8	99.0	99.2	99.4	98.8	98.0	98.2	99.0	99.2	99.0	99.0	98.8	99.2	99.2	99.4	98.6	98.6	99.8	98.8
5,000	90	88.8	89.2	89.4	89.8	90.0	89.4	89.8	89.4	90.2	89.8	92.0	86.0	89.8	90.0	91.2	90.8	89.0	91.2	87.2
5,000	95	94.2	95.8	95.8	94.8	95.6	94.2	94.4	94.2	95.8	94.6	96.6	94.4	94.2	94.8	95.8	96.2	94.0	95.4	93.4
5,000	99	98.8	99.2	98.4	98.4	98.4	98.4	99.8	98.8	99.8	99.2	99.2	98.6	99.6	99.0	99.6	99.4	98.6	99.2	98.2
7,500	90	89.8	92.0	89.0	91.2	90.8	89.0	87.6	90.4	92.0	91.0	88.8	90.2	91.8	94.0	89.2	91.4	90.2	87.8	91.6
7,500	95	94.4	96.4	94.2	96.0	95.0	94.8	93.4	95.4	95.0	96.2	94.2	95.0	97.0	96.6	95.6	97.0	94.8	93.6	94.8
7,500	99	98.8	99.2	98.0	99.4	98.0	99.0	99.0	98.6	99.6	99.8	98.8	98.8	99.0	99.4	98.8	99.4	99.0	98.2	98.6
10,000	90	89.4	89.6	89.6	91.4	91.4	90.8	93.0	88.0	89.2	90.8	89.2	88.4	91.2	88.8	90.6	87.4	91.8	89.2	90.6
10,000	95	95.0	95.2	95.6	95.2	95.6	95.4	96.4	94.2	93.6	95.8	94.4	93.4	96.0	94.2	94.4	93.0	94.6	93.4	95.2
10,000	99	99.0	99.0	99.2	99.0	99.4	99.0	99.0	98.2	98.4	99.0	99.4	98.6	99.8	99.2	98.8	99.0	99.0	98.8	98.6

Table 4.30: Coverage percentages using rs71243915 with 3 components, of which 2 were binary, for 11 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	σ_{12}	σ_{13}	σ_{23}	σ_{33}	τ_{12}	τ_{13}	τ_{23}	τ_{33}	ω_{12}	ω_{13}	ω_{23}	ω_{33}
2,500	90	87.6	89.2	86.8	89.6	88.0	90.2	90.2	89.0	91.8	90.6	89.2	90.4	89.4	89.4	92.4	88.8
2,500	95	93.8	94.4	94.2	93.6	93.6	95.4	94.6	93.2	95.8	96.2	94.8	95.6	95.0	94.8	96.2	94.2
2,500	99	99.0	99.0	98.0	97.8	98.0	98.8	98.8	98.0	98.6	99.6	99.2	99.0	98.4	99.2	98.2	98.6
5,000	90	89.6	89.2	90.6	88.8	90.8	91.8	90.2	89.0	90.8	88.6	89.4	92.4	91.4	89.2	88.6	89.0
5,000	95	95.6	94.6	96.0	93.8	95.4	95.4	95.8	94.6	97.0	94.6	95.0	95.2	95.2	94.8	93.4	93.2
5,000	99	99.4	99.6	99.4	99.0	98.8	98.8	99.4	98.8	99.2	98.6	99.4	98.6	99.4	98.6	98.8	98.0
7,500	90	90.0	87.6	90.0	89.4	92.2	89.6	87.8	89.8	90.2	92.4	91.8	89.2	89.0	90.4	91.6	90.2
7,500	95	94.6	91.6	95.6	93.4	96.2	95.4	93.4	95.2	96.2	96.0	96.4	95.6	93.6	94.0	94.8	95.2
7,500	99	99.0	98.8	99.0	99.0	99.2	99.0	98.6	99.2	98.4	98.8	99.2	99.0	98.4	98.8	98.8	99.0
10,000	90	88.8	89.4	91.0	88.8	89.6	90.2	88.2	92.2	89.0	92.0	90.0	89.8	90.0	88.6	87.4	90.6
10,000	95	95.4	96.0	95.2	94.6	94.8	95.6	94.8	95.4	94.6	95.4	94.8	95.2	94.6	95.4	94.6	94.2
10,000	99	99.2	99.2	99.0	98.8	99.4	98.2	98.8	99.0	99.6	98.8	99.2	98.2	99.2	99.2	97.6	98.4

Table 4.31: Coverage percentages using rs71243915 with 4 components, of which 1 was binary, for 22 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	β_{46}	σ_{12}	σ_{13}	σ_{14}	σ_{22}	σ_{23}	σ_{24}	σ_{33}	σ_{34}	σ_{44}
2,500	90	83.4	91.4	91.4	87.8	88.4	88.8	87.4	90.6	89.6	90.4	88.2	90.8	89.8	90.0
2,500	95	92.6	95.6	95.8	92.2	95.0	93.0	92.6	95.8	95.2	94.8	93.8	95.6	94.4	94.4
2,500	99	98.4	99.4	99.2	98.8	99.2	98.4	99.2	98.8	99.2	98.6	98.0	99.0	98.2	98.8
5,000	90	85.0	89.4	90.0	88.4	88.2	90.6	90.0	89.4	89.4	90.8	89.2	90.0	91.0	88.8
5,000	95	92.4	95.2	95.2	94.8	93.0	93.8	94.8	94.6	94.4	95.6	94.8	96.0	97.0	94.8
5,000	99	98.2	99.6	99.2	98.6	98.8	98.4	99.2	99.6	98.4	99.2	99.0	98.6	99.6	99.6
7,500	90	89.4	90.8	90.2	90.0	88.4	89.8	89.0	88.8	90.4	89.4	86.2	89.8	90.0	89.0
7,500	95	93.8	94.8	94.6	95.4	94.6	95.4	94.8	94.0	94.6	93.8	93.6	93.6	95.2	95.4
7,500	99	98.8	99.6	99.0	99.0	99.2	98.8	98.2	99.4	99.6	98.2	99.0	99.2	99.0	98.6
10,000	90	85.2	88.6	89.6	90.0	89.0	89.0	89.6	89.8	89.4	89.2	89.2	89.2	90.6	86.4
10,000	95	93.2	95.6	95.4	96.2	95.0	94.0	94.8	94.6	94.8	95.2	94.0	94.0	94.2	92.8
10,000	99	99.0	99.6	99.6	99.6	98.8	99.2	98.8	99.0	98.6	98.8	98.6	98.6	98.2	98.6

Obs.	CI	τ_{12}	τ_{13}	τ_{14}	τ_{22}	τ_{23}	τ_{24}	τ_{33}	τ_{34}	τ_{44}	ω_{12}	ω_{13}	ω_{14}	ω_{22}	ω_{23}	ω_{24}	ω_{33}	ω_{34}	ω_{44}
2,500	90	89.6	87.6	90.6	90.0	88.8	92.0	88.2	89.2	92.6	91.0	90.8	87.2	88.2	88.4	89.4	87.8	90.4	89.0
2,500	95	94.2	93.2	96.2	95.0	93.6	95.0	95.2	94.8	95.8	94.0	94.4	93.2	94.4	92.0	94.6	94.4	96.0	94.0
2,500	99	99.0	97.0	99.0	99.2	99.0	99.2	98.8	98.8	98.8	98.0	98.4	98.2	98.2	97.2	99.2	98.6	98.8	98.8
5,000	90	89.6	90.2	91.2	88.4	90.4	89.2	88.4	91.2	88.6	90.0	90.4	89.8	91.6	91.2	91.0	90.6	88.4	90.8
5,000	95	94.4	93.8	94.4	95.4	95.0	95.2	94.0	95.6	94.6	94.4	96.0	94.4	96.6	97.0	95.0	95.4	95.4	94.2
5,000	99	98.4	99.0	98.4	98.8	99.4	99.2	99.0	99.2	97.8	97.8	99.4	99.2	99.0	99.4	99.4	98.8	99.2	99.2
7,500	90	88.6	89.8	90.0	91.0	90.4	88.6	90.4	88.4	90.0	89.0	90.8	89.6	88.2	90.4	92.2	87.2	90.4	89.8
7,500	95	95.0	94.8	95.2	95.0	95.0	95.8	96.2	94.2	94.8	95.6	95.0	95.6	94.8	94.0	97.0	94.0	94.8	94.8
7,500	99	99.0	99.0	98.4	98.8	99.2	99.0	98.6	99.2	99.4	99.2	99.0	99.2	98.2	98.4	99.6	99.0	98.6	98.4
10,000	90	89.2	89.6	90.0	93.2	89.4	90.0	90.4	90.8	89.8	90.6	89.6	89.0	88.6	88.4	87.2	88.0	90.0	87.4
10,000	95	94.8	95.0	95.4	96.0	93.6	94.6	95.6	96.4	94.8	94.4	96.2	94.8	94.4	94.0	94.0	93.2	95.8	95.2
10,000	99	98.8	98.8	98.8	99.2	98.8	99.0	99.0	98.8	99.2	98.6	99.2	99.0	99.0	98.6	99.0	97.4	98.8	99.0

Table 4.32: Coverage percentages using rs71243915 with 4 components, of which 2 were binary, for 20 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	β_{46}	σ_{12}	σ_{13}	σ_{14}	σ_{23}	σ_{24}	σ_{33}	σ_{34}	σ_{44}
2,500	90	86.0	89.8	90.6	89.0	88.6	90.8	87.4	89.6	87.0	90.0	90.2	88.8	89.8
2,500	95	91.0	94.6	95.0	95.4	94.0	94.8	93.2	96.0	93.4	94.0	94.4	94.2	95.0
2,500	99	98.4	99.2	98.4	98.8	98.0	99.4	98.4	99.2	98.4	98.8	98.0	99.4	99.2
5,000	90	88.6	89.2	92.4	90.8	90.4	90.4	87.6	87.8	88.8	90.8	91.6	90.0	91.4
5,000	95	93.8	94.2	96.0	94.4	95.4	95.8	93.4	93.8	95.0	95.2	95.2	94.0	95.4
5,000	99	98.6	98.6	99.0	98.4	99.6	98.8	98.4	97.8	98.4	99.0	99.0	99.4	99.0
7,500	90	88.4	91.0	89.0	85.8	89.8	92.2	88.8	90.0	90.2	90.4	87.4	87.4	89.0
7,500	95	93.4	95.6	95.0	92.2	93.6	96.2	95.2	94.6	94.4	95.4	92.2	94.0	93.2
7,500	99	97.8	99.4	99.2	98.0	98.8	99.0	99.4	99.2	99.0	99.2	98.0	98.0	98.0
10,000	90	92.0	91.0	89.4	91.4	90.6	89.4	90.4	90.6	87.8	90.0	89.2	89.4	91.0
10,000	95	96.0	96.8	93.6	95.6	94.6	95.8	93.6	95.0	94.0	94.8	95.6	95.8	95.0
10,000	99	99.2	99.4	98.8	99.4	99.0	99.0	98.8	98.8	98.8	99.6	98.8	99.2	98.6

Obs.	CI	τ_{12}	τ_{13}	τ_{14}	τ_{23}	τ_{24}	τ_{33}	τ_{34}	τ_{44}	ω_{12}	ω_{13}	ω_{14}	ω_{23}	ω_{24}	ω_{33}	ω_{34}	ω_{44}
2,500	90	88.6	91.2	90.6	91.2	91.0	89.0	86.8	88.8	90.6	90.4	88.0	91.4	87.6	89.6	88.6	88.4
2,500	95	94.6	94.8	95.6	95.6	95.0	94.8	93.8	95.4	95.4	95.8	93.8	96.0	94.8	95.2	94.4	94.4
2,500	99	99.0	98.8	99.0	99.2	99.4	99.2	97.8	98.8	99.0	99.6	99.2	99.2	99.0	99.2	99.2	99.2
5,000	90	91.6	89.8	87.2	90.4	87.6	91.4	89.8	90.4	90.4	91.4	90.4	89.0	91.4	87.4	90.6	88.4
5,000	95	94.2	95.2	94.0	95.4	94.0	96.4	95.2	95.2	95.0	96.4	96.0	94.2	94.6	95.4	94.2	94.2
5,000	99	99.4	99.0	99.2	99.6	98.4	99.6	98.4	99.2	98.2	98.8	99.0	99.6	99.2	98.8	98.4	98.8
7,500	90	90.0	91.8	92.2	90.8	89.4	90.4	91.8	89.4	90.4	89.8	90.6	91.0	89.2	90.0	89.8	89.2
7,500	95	95.4	96.6	96.4	95.8	95.2	94.4	96.4	94.8	95.2	95.8	95.2	94.8	95.0	95.0	94.0	96.0
7,500	99	99.2	99.4	99.0	99.6	98.6	99.4	98.8	98.6	99.6	99.8	99.4	99.2	99.0	97.6	99.0	99.2
10,000	90	89.6	90.6	88.6	88.8	91.0	90.8	93.6	92.0	88.2	90.0	89.8	91.6	91.2	90.0	89.8	90.8
10,000	95	94.6	97.0	95.0	94.8	96.4	95.8	97.0	96.6	93.8	95.8	94.4	96.6	96.4	94.0	93.8	95.8
10,000	99	98.2	99.8	99.2	98.4	99.6	98.8	99.6	99.0	99.4	99.0	99.0	99.2	99.6	98.8	98.8	99.6

Table 4.33: Coverage percentages using rs71243915 with 4 components, of which 3 were binary, for 18 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	β_{46}	σ_{12}	σ_{13}	σ_{14}	σ_{23}	σ_{24}	σ_{34}	σ_{44}
2,500	90	87.2	91.0	90.4	90.0	91.2	89.8	90.0	91.6	88.4	89.2	89.8	90.8
2,500	95	93.8	95.6	94.6	95.8	95.8	94.2	93.8	97.2	94.0	94.0	95.0	95.2
2,500	99	99.0	99.4	99.2	99.2	99.2	99.0	99.2	99.2	99.2	98.4	99.4	99.8
5,000	90	89.4	90.0	88.2	88.2	90.6	89.0	88.8	90.6	89.6	88.8	90.4	91.0
5,000	95	93.6	95.8	94.8	95.0	94.8	95.0	93.6	95.4	95.4	94.8	95.4	95.8
5,000	99	98.2	99.2	98.4	99.2	99.2	99.4	98.8	99.2	98.6	99.0	98.6	98.6
7,500	90	88.8	87.4	90.0	89.4	90.4	91.8	90.6	89.6	86.6	91.4	87.8	89.4
7,500	95	94.8	93.4	95.2	95.0	96.2	95.4	93.8	94.6	93.2	97.0	94.4	95.0
7,500	99	97.8	98.0	99.0	99.8	99.2	98.8	99.0	97.6	98.4	99.4	98.8	99.2
10,000	90	89.4	90.6	90.4	91.6	92.8	91.0	87.8	87.4	89.6	90.6	90.0	92.0
10,000	95	95.2	94.0	95.6	95.0	96.2	95.8	94.8	94.8	94.0	95.2	95.4	96.4
10,000	99	98.6	99.2	99.2	99.4	99.4	98.8	98.6	99.4	98.4	99.0	98.8	99.4

Obs.	CI	τ_{12}	τ_{13}	τ_{14}	τ_{23}	τ_{24}	τ_{34}	τ_{44}	ω_{12}	ω_{13}	ω_{14}	ω_{23}	ω_{24}	ω_{34}	ω_{44}
2,500	90	88.8	88.4	89.8	90.0	89.8	88.6	92.2	90.4	89.8	89.4	89.0	86.0	91.6	90.2
2,500	95	93.4	95.4	94.8	95.8	95.4	92.8	97.6	95.6	95.0	94.0	94.0	94.4	95.6	94.2
2,500	99	99.4	99.2	98.6	99.8	98.8	98.6	99.4	99.2	98.8	99.0	98.8	98.2	98.8	98.6
5,000	90	88.6	87.6	91.6	90.4	91.8	90.6	89.0	91.6	88.2	91.2	89.8	91.6	91.4	92.2
5,000	95	96.2	92.6	97.0	95.0	95.0	94.2	94.4	96.2	94.6	95.0	94.2	94.6	95.0	95.8
5,000	99	98.6	97.8	99.6	98.2	99.2	99.4	99.4	99.4	98.6	98.6	98.8	98.6	99.8	98.8
7,500	90	91.0	90.0	91.4	89.6	90.2	92.8	90.8	91.4	92.4	91.6	90.8	89.8	90.8	91.0
7,500	95	95.2	96.2	95.8	95.2	96.4	97.2	95.0	93.8	96.6	95.2	95.8	95.0	96.4	95.2
7,500	99	99.2	99.4	98.6	99.4	99.4	99.2	98.8	98.8	99.8	99.2	98.4	99.6	99.2	98.6
10,000	90	89.4	90.8	90.0	89.0	89.2	90.6	89.2	90.2	90.6	89.0	90.4	91.0	92.2	90.6
10,000	95	95.0	94.8	94.8	94.8	94.0	95.2	94.4	94.4	96.4	93.8	95.8	95.4	97.6	96.0
10,000	99	99.2	98.2	99.4	98.2	98.4	99.0	98.8	98.4	99.4	99.0	99.0	98.4	99.2	99.0

Table 4.34: Coverage percentages using rs77267992 with 2 components, of which 1 was binary, for 6 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	σ_{12}	σ_{22}	τ_{12}	τ_{22}	ω_{12}	ω_{22}
2,500	90	87.8	89.8	89.4	87.8	90.8	88.0	91.2	90.4	87.8
2,500	95	93.4	96.4	95.0	94.0	93.6	94.8	96.0	95.6	93.6
2,500	99	98.8	99.6	98.8	99.0	98.2	99.6	98.8	99.2	98.2
5,000	90	90.4	92.0	88.6	89.4	88.2	88.8	88.6	89.6	90.0
5,000	95	94.2	97.0	93.0	95.6	93.8	95.0	94.8	95.2	95.8
5,000	99	99.4	99.8	98.4	98.8	99.0	98.6	98.8	99.8	99.0
7,500	90	91.6	91.4	90.2	89.8	92.2	89.4	88.6	90.4	89.8
7,500	95	94.4	94.0	96.0	94.4	96.0	95.4	94.0	95.6	95.4
7,500	99	99.6	99.2	99.2	99.4	98.8	99.2	99.0	99.0	99.4
10,000	90	91.6	91.0	90.6	90.0	90.0	87.6	89.8	89.4	90.4
10,000	95	95.0	97.2	95.4	94.0	94.6	93.8	95.2	94.4	94.0
10,000	99	98.6	99.0	98.6	98.4	99.4	99.0	99.0	99.4	98.4

Table 4.35: Coverage percentages using rs77267992 with 3 components, of which 1 was binary, for 13 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	σ_{12}	σ_{13}	σ_{22}	σ_{23}	σ_{33}	τ_{12}	τ_{13}	τ_{22}	τ_{23}	τ_{33}	ω_{12}	ω_{13}	ω_{22}	ω_{23}	ω_{33}
2,500	90	88.2	89.8	91.6	91.0	87.2	90.6	91.0	90.0	89.4	91.2	87.4	89.4	90.4	89.8	91.0	90.0	90.0	88.2	90.0
2,500	95	94.2	94.4	94.6	95.0	93.4	95.2	95.4	95.2	94.2	95.8	93.6	95.0	95.8	93.6	96.0	94.4	95.2	94.6	95.0
2,500	99	98.6	98.6	99.4	99.4	98.6	98.4	99.0	98.2	98.6	98.8	98.6	98.0	99.6	98.2	98.2	98.4	98.8	99.6	98.8
5,000	90	89.4	89.8	88.6	90.6	92.0	90.6	89.6	90.4	91.0	91.4	89.0	91.2	89.8	92.6	87.6	90.6	88.8	88.4	92.8
5,000	95	94.6	94.6	94.6	95.6	95.2	94.8	95.6	94.6	95.4	95.6	94.4	95.2	95.0	96.6	93.4	95.4	94.8	94.2	96.0
5,000	99	99.4	99.0	98.8	99.2	98.6	99.0	100.0	98.8	99.0	98.2	99.0	99.2	98.8	99.6	97.8	99.0	99.0	98.6	99.4
7,500	90	87.2	90.4	91.0	88.6	89.8	90.2	89.8	90.8	91.2	89.2	91.2	92.8	91.0	92.0	92.6	89.8	91.0	89.6	90.8
7,500	95	94.0	95.6	95.6	94.6	95.2	95.0	94.0	95.6	95.2	94.0	95.8	96.8	94.8	96.4	97.2	94.2	94.8	95.0	94.8
7,500	99	99.0	99.2	99.4	98.8	98.8	98.2	98.4	99.0	99.0	98.6	99.0	99.8	99.2	99.2	99.4	98.8	98.6	98.0	99.4
10,000	90	91.4	89.8	89.0	89.2	87.4	90.6	90.6	90.6	90.2	90.8	89.4	91.8	92.4	87.6	90.8	90.0	89.6	90.2	90.4
10,000	95	95.4	95.0	95.0	94.0	93.4	95.0	96.0	94.6	95.6	95.0	95.2	96.8	95.6	95.8	95.8	94.4	94.4	94.6	93.8
10,000	99	99.0	98.8	98.4	98.8	98.0	98.4	98.6	99.2	99.0	98.0	99.8	99.8	98.6	99.2	99.0	98.4	99.4	99.2	98.8

Table 4.36: Coverage percentages using rs77267992 with 3 components, of which 2 were binary, for 11 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	σ_{12}	σ_{13}	σ_{23}	σ_{33}	τ_{12}	τ_{13}	τ_{23}	τ_{33}	ω_{12}	ω_{13}	ω_{23}	ω_{33}
2,500	90	88.4	90.8	89.6	88.2	86.6	88.0	89.4	90.4	92.4	88.2	89.4	91.2	89.0	89.6	89.4	87.2
2,500	95	94.4	95.8	94.4	94.2	92.8	93.8	94.2	94.4	96.2	93.6	96.2	96.2	94.2	94.2	93.6	92.0
2,500	99	98.6	99.6	99.0	98.6	99.0	98.2	99.4	98.8	99.2	99.2	99.6	98.8	99.2	99.0	98.2	98.0
5,000	90	86.6	88.6	90.6	90.6	89.4	86.2	91.4	90.8	89.2	88.2	90.8	92.4	90.0	90.0	89.4	92.0
5,000	95	93.2	95.8	95.6	94.8	94.4	92.8	95.0	95.0	93.0	92.6	94.6	96.4	96.0	95.2	95.6	96.0
5,000	99	98.2	98.8	99.2	98.8	99.2	98.2	99.2	99.0	99.2	98.8	99.4	99.6	99.4	99.4	98.6	99.2
7,500	90	90.4	89.8	89.4	92.4	90.2	91.4	88.4	91.4	89.0	91.4	86.6	92.0	90.0	87.4	89.4	90.4
7,500	95	95.6	95.6	94.0	95.8	94.6	94.8	92.4	96.0	94.2	95.2	93.8	96.2	94.4	93.6	95.0	94.8
7,500	99	98.8	98.8	98.8	98.8	99.2	99.2	97.8	99.4	99.0	98.6	98.8	99.0	99.0	98.8	99.4	98.8
10,000	90	90.0	89.8	90.4	90.6	89.4	88.2	87.6	90.2	91.2	89.2	90.8	90.6	89.8	91.6	90.2	89.8
10,000	95	94.8	94.8	94.8	94.8	95.8	95.2	94.4	96.2	95.0	94.2	95.8	95.0	94.8	94.6	94.6	93.8
10,000	99	99.0	99.2	99.0	99.2	99.2	99.2	98.6	99.4	98.8	98.8	99.4	99.2	98.2	99.0	99.0	98.4

Table 4.37: Coverage percentages using rs77267992 with 4 components, of which 1 was binary, for 22 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	β_{46}	σ_{12}	σ_{13}	σ_{14}	σ_{22}	σ_{23}	σ_{24}	σ_{33}	σ_{34}	σ_{44}
2,500	90	86.6	90.4	90.0	90.2	90.4	89.8	91.2	90.4	91.4	90.4	88.2	92.6	89.8	91.4
2,500	95	92.4	94.8	95.8	94.4	93.8	94.4	94.4	94.8	96.6	94.8	92.4	96.2	94.6	96.6
2,500	99	98.2	99.0	99.2	99.2	98.8	99.4	99.2	99.0	99.0	98.6	97.4	99.4	98.6	99.0
5,000	90	86.8	92.0	88.2	90.4	90.2	88.2	87.0	90.0	92.8	89.2	88.2	91.0	89.6	90.2
5,000	95	93.6	96.6	93.2	93.4	94.4	95.2	93.6	94.8	97.0	93.4	93.6	96.0	94.0	95.2
5,000	99	98.8	99.8	98.8	98.0	99.6	98.2	98.8	99.0	99.6	98.2	98.8	99.4	99.0	98.4
7,500	90	88.6	91.8	91.2	88.2	92.0	88.4	88.8	90.2	88.8	89.4	88.2	89.0	92.8	89.4
7,500	95	94.2	95.4	95.2	94.0	95.6	94.8	93.6	94.4	95.8	94.6	94.6	93.8	97.0	94.4
7,500	99	98.6	98.6	99.2	98.8	99.0	98.0	98.8	98.4	99.4	99.2	99.0	98.6	99.6	98.6
10,000	90	91.4	92.8	90.0	92.8	90.2	91.2	90.8	90.6	89.2	88.8	87.4	91.2	89.6	90.4
10,000	95	95.2	95.6	94.4	95.8	94.4	97.0	96.4	94.2	94.4	94.2	93.6	95.6	95.2	95.0
10,000	99	97.8	99.4	99.4	98.8	99.0	99.6	99.2	98.4	98.4	99.0	99.0	99.0	98.4	98.0

Obs.	CI	τ_{12}	τ_{13}	τ_{14}	τ_{22}	τ_{23}	τ_{24}	τ_{33}	τ_{34}	τ_{44}	ω_{12}	ω_{13}	ω_{14}	ω_{22}	ω_{23}	ω_{24}	ω_{33}	ω_{34}	ω_{44}
2,500	90	86.6	92.0	88.2	90.6	90.2	90.4	88.2	89.8	89.0	93.0	90.4	89.4	90.6	87.2	89.4	89.2	89.6	87.8
2,500	95	93.0	96.4	93.2	95.0	94.4	96.6	94.0	94.2	93.8	97.6	95.4	93.4	95.0	94.0	95.4	94.8	94.2	92.2
2,500	99	98.8	99.6	98.6	98.4	99.0	99.4	97.8	99.2	98.2	99.8	98.8	97.8	98.6	99.0	98.8	98.6	98.8	99.0
5,000	90	90.0	91.8	89.2	88.6	87.6	91.2	87.2	87.6	91.4	87.2	89.0	90.4	91.8	90.2	90.0	88.4	88.0	90.6
5,000	95	94.8	95.4	94.6	94.0	93.0	95.0	92.8	93.4	95.8	92.8	93.2	95.6	95.6	96.2	94.2	94.0	93.2	95.6
5,000	99	99.8	99.4	99.2	99.0	97.4	98.6	98.2	98.8	99.2	98.8	98.0	99.0	99.0	99.2	98.4	98.6	98.0	99.8
7,500	90	89.6	91.4	89.8	93.8	92.6	90.2	89.4	91.4	91.6	90.0	88.8	89.2	91.8	89.0	92.8	89.4	90.8	89.6
7,500	95	94.4	95.2	94.0	97.4	95.6	94.2	96.4	95.6	95.8	94.8	94.2	94.6	96.0	94.8	96.6	94.2	95.4	95.4
7,500	99	98.6	98.4	98.6	99.6	99.4	98.6	99.4	98.6	99.6	99.0	97.8	99.0	99.2	98.8	99.0	98.4	99.0	99.6
10,000	90	87.8	89.2	89.8	94.0	88.4	90.8	88.6	87.8	91.2	92.2	90.2	89.6	92.0	91.6	91.4	90.2	90.8	90.0
10,000	95	93.8	94.6	95.6	97.0	94.8	95.4	94.0	94.8	95.2	97.0	95.2	94.0	97.0	96.2	94.8	95.0	95.2	96.0
10,000	99	98.0	99.4	99.4	99.6	99.2	98.2	98.0	99.0	99.0	99.6	98.6	99.2	99.4	99.0	99.2	99.6	99.0	99.4

Table 4.38: Coverage percentages using rs77267992 with 4 components, of which 2 were binary, for 20 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	β_{46}	σ_{12}	σ_{13}	σ_{14}	σ_{23}	σ_{24}	σ_{33}	σ_{34}	σ_{44}
2,500	90	87.8	91.2	89.8	92.0	84.6	88.8	89.2	90.8	90.0	89.2	89.4	91.8	88.4
2,500	95	94.2	95.0	94.4	96.4	91.8	92.8	95.6	96.4	94.8	93.6	93.2	95.4	95.2
2,500	99	98.8	99.0	99.6	100.0	98.8	98.6	98.6	99.2	99.2	98.0	98.2	98.8	99.2
5,000	90	88.4	90.0	89.6	87.8	88.8	90.6	89.4	89.8	89.4	90.6	90.0	91.2	90.2
5,000	95	95.0	94.6	95.2	92.4	94.0	94.6	94.0	94.6	94.8	95.6	94.8	96.2	95.2
5,000	99	98.8	98.4	99.0	98.0	99.4	99.0	98.0	99.6	98.6	99.2	99.0	99.0	99.0
7,500	90	91.6	88.6	90.6	90.2	89.4	90.0	88.8	89.6	90.6	89.6	88.4	94.2	90.8
7,500	95	95.6	95.0	95.0	94.2	94.6	95.2	94.0	95.4	95.6	95.2	94.4	97.4	95.4
7,500	99	98.6	99.4	99.6	98.6	97.8	99.2	98.8	99.2	98.8	98.8	99.0	99.6	98.4
10,000	90	90.6	90.4	91.6	88.6	90.8	89.2	90.0	90.6	89.8	88.4	90.2	89.8	91.0
10,000	95	95.4	95.0	96.0	94.6	94.6	95.0	95.6	95.2	94.4	94.0	94.4	94.2	94.8
10,000	99	98.8	99.2	99.6	99.4	98.8	97.8	98.8	99.0	98.4	98.6	98.0	98.8	99.0

Obs.	CI	τ_{12}	τ_{13}	τ_{14}	τ_{23}	τ_{24}	τ_{33}	τ_{34}	τ_{44}	ω_{12}	ω_{13}	ω_{14}	ω_{23}	ω_{24}	ω_{33}	ω_{34}	ω_{44}
2,500	90	89.6	90.6	90.6	89.8	85.4	89.8	88.4	88.4	89.2	93.8	91.4	90.2	89.0	90.0	90.8	89.6
2,500	95	95.2	95.4	94.8	95.0	92.0	96.8	93.6	94.2	95.6	96.2	96.2	94.8	95.6	95.0	96.0	95.0
2,500	99	99.4	99.0	98.4	99.2	98.8	99.8	98.8	98.4	98.8	99.6	98.6	99.4	99.6	99.2	99.2	99.0
5,000	90	91.2	91.0	88.4	90.0	92.6	88.6	91.2	88.2	89.6	90.2	90.0	91.4	86.6	92.2	86.8	91.8
5,000	95	95.2	95.0	93.6	95.6	97.2	94.2	96.2	95.0	95.6	94.6	94.0	95.8	94.4	96.8	93.0	96.6
5,000	99	99.2	98.6	98.4	99.2	99.6	99.2	98.8	99.2	99.4	99.2	98.6	99.8	99.0	99.4	98.0	99.4
7,500	90	88.8	91.2	87.8	92.0	88.4	91.0	89.8	88.2	88.8	89.4	89.8	92.0	91.2	90.6	89.0	91.2
7,500	95	96.0	95.8	95.2	96.0	93.8	95.8	95.2	94.0	95.0	93.4	95.4	96.4	95.8	94.8	94.6	95.2
7,500	99	98.6	98.6	99.6	99.0	98.8	99.2	99.2	98.8	99.0	98.2	99.4	99.2	99.0	98.8	99.4	99.2
10,000	90	89.2	91.0	90.6	88.8	90.4	91.2	91.6	90.2	88.0	90.2	90.4	91.2	92.0	91.0	89.8	89.8
10,000	95	94.0	95.6	95.0	93.4	96.2	96.8	96.4	94.8	94.4	94.6	96.0	94.8	96.4	95.6	95.2	94.8
10,000	99	98.4	99.2	98.8	98.2	99.2	99.4	98.8	98.4	99.4	98.6	99.4	99.0	99.6	98.8	99.0	99.2

Table 4.39: Coverage percentages using rs77267992 with 4 components, of which 3 were binary, for 18 degrees of freedom. 500 replications were performed to generate the figures. W denotes the Wald test-statistic. β_{j6} denotes the SNP coefficient for the j'th component. σ_{jk}, τ_{jk}, and ω_{jk} denote the heteroskedastic covariance parameters for components j and k.

Obs.	CI	W	β_{16}	β_{26}	β_{36}	β_{46}	σ_{12}	σ_{13}	σ_{14}	σ_{23}	σ_{24}	σ_{34}	σ_{44}
2,500	90	85.4	90.8	88.6	89.6	90.4	89.4	88.2	91.2	91.8	91.4	92.2	90.4
2,500	95	93.6	96.6	94.4	93.6	95.2	93.2	93.2	96.4	93.8	95.2	95.8	96.0
2,500	99	98.4	99.2	98.6	99.2	98.8	99.0	99.2	99.0	98.0	99.0	99.4	99.2
5,000	90	88.6	90.8	90.8	89.6	89.2	90.4	87.4	89.6	89.2	91.2	89.6	88.2
5,000	95	93.2	95.6	95.6	96.4	94.4	94.4	94.6	94.4	94.2	96.8	95.6	93.8
5,000	99	99.4	99.8	99.0	99.8	99.0	98.8	98.8	98.8	98.8	98.8	98.8	98.6
7,500	90	91.6	91.4	91.2	90.0	89.6	91.8	88.6	90.8	88.2	88.4	89.6	90.0
7,500	95	95.4	95.8	95.6	95.2	94.4	95.0	95.6	95.6	94.0	94.4	94.2	95.4
7,500	99	98.6	99.2	99.2	99.4	99.0	98.8	99.0	99.8	99.6	99.0	98.6	98.4
10,000	90	90.2	90.0	90.2	91.6	91.6	88.8	92.4	91.4	89.6	90.0	89.8	88.6
10,000	95	95.8	95.0	95.4	96.4	96.2	95.0	96.8	95.2	94.0	95.0	94.8	93.8
10,000	99	98.8	99.2	99.0	99.0	99.4	98.8	99.4	99.4	99.2	98.8	98.6	97.4

Obs.	CI	τ_{12}	τ_{13}	τ_{14}	τ_{23}	τ_{24}	τ_{34}	τ_{44}	ω_{12}	ω_{13}	ω_{14}	ω_{23}	ω_{24}	ω_{34}	ω_{44}
2,500	90	89.4	89.4	90.2	87.8	90.2	90.8	86.2	91.6	92.2	91.0	90.2	88.8	91.6	88.8
2,500	95	93.6	94.0	94.0	94.0	95.2	96.0	92.8	96.6	96.2	95.4	96.2	93.4	95.6	93.8
2,500	99	98.4	99.6	98.6	98.6	98.6	98.6	98.4	98.6	99.0	99.4	99.6	98.4	99.4	98.4
5,000	90	89.0	88.2	89.2	89.6	89.8	90.8	91.6	88.4	89.2	89.6	89.4	90.4	90.0	88.6
5,000	95	95.0	92.6	94.6	93.4	93.8	95.6	95.8	94.0	93.4	94.0	94.2	96.2	96.0	93.8
5,000	99	99.4	98.8	98.0	97.6	98.6	99.0	99.2	99.6	98.6	98.8	99.6	99.0	99.2	98.8
7,500	90	89.6	87.4	91.2	88.8	91.6	88.6	88.4	88.2	91.2	91.4	88.6	90.6	91.0	88.8
7,500	95	94.8	93.4	95.4	94.6	96.4	93.4	95.0	94.4	95.4	96.0	94.6	95.0	94.8	93.8
7,500	99	98.8	99.2	99.4	99.2	98.6	98.2	99.2	98.8	99.2	98.8	98.4	98.8	100.0	99.0
10,000	90	89.0	91.6	88.8	94.0	88.2	90.6	89.8	90.8	91.0	90.8	91.2	91.0	90.6	90.0
10,000	95	95.4	95.0	93.2	97.2	92.2	95.2	94.8	94.6	95.6	95.6	95.6	94.6	95.4	95.0
10,000	99	99.6	99.2	98.8	99.2	97.4	99.2	98.8	99.2	99.2	99.4	98.8	98.8	99.4	98.6

References

G. Abraham and M. Inouye. Fast principal component analysis of large-scale genome-wide data. *PLOS ONE*, 9(4):1–5, 04 2014. doi: 10.1371/journal.pone.0093766. URL https://doi.org/10.1371/journal.pone.0093766.

D. Altshuler, P. Donnelly, and T. I. H. Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005. ISSN 1476-4687. doi: 10.1038/nature04226. URL https://doi.org/10.1038/nature04226.

P. L. Auer and G. Lettre. Rare variant association studies: considerations, challenges and opportunities. *Genome Medicine*, 7(1):16, 2015. ISSN 1756-994X. doi: 10.1186/s13073-015-0138-2. URL https://doi.org/10.1186/s13073-015-0138-2.

Auton, A. and Abecasis, G. and Altshuler, D. et al. A global reference for human genetic variation. *Nature*, 526:68–74, 2015. URL https://doi.org/10.1038/nature15393.

D. D. Boos and L. A. Stefanski. Likelihood Construction and Estimation. In *Essential Statistical Inference: Theory and Methods*, pages 27–124. Springer New York, New York, NY, 2013. ISBN 978-1-4614-4818-1. doi: 10.1007/978-1-4614-4818-1. URL https://doi.org/10.1007/978-1-4614-4818-1.

T. J. Bouchard. Genetic Influence on Human Psychological Traits: A Survey. *Current Directions in Psychological Science*, 13(4):148–151, Aug. 2004. ISSN 0963-7214. doi: 10.1111/j.0963-7214.2004.00295.x. URL https://doi.org/10.1111/j.0963-7214.2004.00295.x. Publisher: SAGE Publications Inc.

W. M. Brandler, D. Antaki, M. Gujral, M. L. Kleiber, J. Whitney, M. S. Maile, O. Hong, T. R. Chapman, S. Tan, P. Tandon, T. Pang, S. C. Tang, K. K. Vaux, Y. Yang, E. Harrington, S. Juul, D. J. Turner, B. Thiruvahindrapuram, G. Kaur, Z. Wang, S. F. Kingsmore, J. G. Gleeson, D. Bisson, B. Kakaradov, A. Telenti, J. C. Venter, R. Corominas, C. Toma,

B. Cormand, I. Rueda, S. Guijarro, K. S. Messer, C. M. Nievergelt, M. J. Arranz, E. Courchesne, K. Pierce, A. R. Muotri, L. M. Iakoucheva, A. Hervas, S. W. Scherer, C. Corsello, and J. Sebat. Paternally inherited cis-regulatory structural variants are associated with autism. *Science*, 360(6386):327–331, 2018. ISSN 0036-8075. doi: 10.1126/science.aan2261. URL https://science.sciencemag.org/content/360/6386/327. Publisher: American Association for the Advancement of Science _eprint: https://science.sciencemag.org/content/360/6386/327.full.pdf.

G. R. Brown, K. N. Laland, and M. B. Mulder. Bateman's principles and human sex roles. *Trends in Ecology & Evolution*, 24(6):297–304, Jun 2009. ISSN 0169-5347. doi: 10.1016/j.tree.2009.02.005. URL https://doi.org/10.1016/j.tree.2009.02.005.

S. R. Browning and B. L. Browning. Haplotype phasing: existing methods and new developments. *Nature reviews. Genetics*, 12(10):703–714, Sept. 2011. ISSN 1471-0064. doi: 10.1038/nrg3054. URL https://pubmed.ncbi.nlm.nih.gov/21921926.

A. Budu-Aggrey, J. Bowes, P. E. Stuart, M. Zawistowski, L. C. Tsoi, R. Nair, D. R. Jadon, N. McHugh, E. Korendowych, J. T. Elder, A. Barton, and S. Raychaudhuri. A rare coding allele in IFIH1 is protective for psoriatic arthritis. *Annals of the Rheumatic Diseases*, 76 (7):1321–1324, July 2017. ISSN 0003-4967. doi: 10.1136/annrheumdis-2016-210592. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5530346/.

W. S. Bush and J. H. Moore. Chapter 11: Genome-wide association studies. *PLOS Computational Biology*, 8(12):1–11, 12 2012. doi: 10.1371/journal.pcbi.1002822. URL https://doi.org/10.1371/journal.pcbi.1002822.

Y. Cao, P. Wei, M. Bailey, J. S. K. Kauwe, and T. J. Maxwell. A versatile omnibus test for detecting mean and variance heterogeneity. *Genetic Epidemiology*, 38(1):51–59, 2014. doi: 10.1002/gepi.21778. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.21778.

L. Cappellari and S. P. Jenkins. Multivariate probit regression using simulated maximum likelihood. *The Stata Journal*, 3(3):278–294, 2003. doi: 10.1177/1536867X0300300305. URL https://doi.org/10.1177/1536867X0300300305.

G. Casella and R. L. Berger. *Statistical inference*. Brooks/COle Cengage Learning, 2017.

M. Cattelan and N. Sartori. Empirical and simulated adjustments of composite likelihood ratio statistics. *Journal of Statistical Computation and Simulation*, 86(5):1056–1067, 2016. doi: 10.1080/00949655.2015.1053091. URL https://doi.org/10.1080/00949655.2015.1053091.

S. J. Chanock, T. Manolio, M. Boehnke, E. Boerwinkle, D. J. Hunter, G. Thomas, J. N. Hirschhorn, G. Abecasis, D. Altshuler, J. E. Bailey-Wilson, L. D. Brooks, L. R. Cardon, M. Daly, P. Donnelly, J. F. Fraumeni, N. B. Freimer, D. S. Gerhard, C. Gunter, A. E. Guttmacher, M. S. Guyer, E. L. Harris, J. Hoh, R. Hoover, C. A. Kong, K. R. Merikangas, C. C. Morton, L. J. Palmer, E. G. Phimister, J. P. Rice, J. Roberts, C. Rotimi, M. A. Tucker, K. J. Vogan, S. Wacholder, E. M. Wijsman, D. M. Winn, F. S. Collins, and N.-N. W. G. o. R. i. A. Studies. Replicating genotype-phenotype associations. *Nature*, 447(7145):655–660, 2007. ISSN 1476-4687. doi: 10.1038/447655a. URL https://doi.org/10.1038/447655a.

S. Chib and E. Greenberg. Analysis of multivariate probit models. *Biometrika*, 85(2):347–361, 1998. ISSN 00063444. URL http://www.jstor.org/stable/2337362.

C. Choirat, J. Honaker, K. Imai, G. King, and O. Lau. *Zelig: Everyone's Statistical Software*, 2018. URL http://zeligproject.org/. Version 5.1.6.1.

Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet (London, England)*, 381(9875):1371–1379, Apr. 2013. ISSN 1474-547X. doi: 10.1016/S0140-6736(12)62129-1.

R. B. D'Agostino and H. K. Russell. *Multivariate Bartlett Test*. American Cancer Society, 2005. ISBN 9780470011812. doi: 10.1002/0470011815.b2a13048. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/0470011815.b2a13048.

A. R. de Leon and B. Wu. Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine*, 30(2):175–185, 2011. doi: 10.1002/sim.4087. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4087.

U. Demkow and T. Wolańczyk. Genetic tests in major psychiatric disorders—integrating molecular medicine with clinical psychiatry—why is it so difficult? *Translational Psychiatry*, 7(6):e1151, June 2017. ISSN 2158-3188. doi: 10.1038/tp.2017.106. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5537634/.

J. C. Denny, M. D. Ritchie, M. A. Basford, J. M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D. R. Masys, D. M. Roden, and D. C. Crawford. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 26(9):1205–1210, May 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq126. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2859132/.

B. W. Domingue, J. Fletcher, D. Conley, and J. D. Boardman. Genetic and educational assortative mating among us adults. *Proceedings of the National Academy of Sciences*, 111(22):7996–8000, 2014. ISSN 0027-8424. doi: 10.1073/pnas.1321426111. URL https://www.pnas.org/content/111/22/7996.

D. Fanelli. Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences*, 115(11): 2628–2631, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1708272114. URL https://www.pnas.org/content/115/11/2628. Publisher: National Academy of Sciences _eprint: https://www.pnas.org/content/115/11/2628.full.pdf.

M.-L. Feddag. Composite likelihood estimation for multivariate probit latent traits models. *Communications in Statistics - Theory and Methods*, 42(14):2551–2566, 2013. doi: 10.1080/03610926.2010.538793. URL https://doi.org/10.1080/03610926.2010.538793.

F. Fidler and J. Wilcox. Reproducibility of scientific results. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2018 edition, 2018.

G. M. Fitzmaurice and N. M. Laird. Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association*, 90(431):845–852, 1995. doi: 10.1080/01621459.1995.10476583. URL https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1995.10476583.

T. J. Flavin and L. Sheenan. The role of U.S. subprime mortgage-backed assets in propagating the crisis: Contagion or interdependence? *The North American Journal of Economics and Finance*, 34:167 – 186, 2015. ISSN 1062-9408. doi: https://doi.org/10.1016/j.najef.2015.09.001. URL http://www.sciencedirect.com/science/article/pii/S1062940815000716.

A. for Healthcare Research and Quality. Medical expenditure panel survey (meps). *Agency for Healthcare Research and Quality, Rockville, MD*, 2018. URL https://www.ahrq.gov/data/meps.html.

O. Francois. Running structure-like population genetic analyses with r. 2016. R Tutorials in Population Genetics, U. Grenoble-Alpes, pages 1-9.

E. Frichot and O. Francois. LEA: an R package for Landscape and Ecological Association studies. 2014. https://www.bioconductor.org/packages/release/bioc/vignettes/LEA/inst/doc/LEA.pdf.

T. E. Galesloot, K. van Steen, L. A. L. M. Kiemeney, L. L. Janss, and S. H. Vermeulen. A Comparison of Multivariate Genome-Wide Association Methods. *PLOS ONE*, 9(4):1–8, 2014. doi: 10.1371/journal.pone.0095923. URL https://doi.org/10.1371/journal.pone.0095923. Publisher: Public Library of Science.

A. Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1(2):141–149, 1992. doi: 10.1080/10618600.1992.10477010. URL https://amstat.tandfonline.com/doi/abs/10.1080/10618600.1992.10477010.

A. Ghosh, F. A. Wright, and F. Zou. Unified analysis of secondary traits in case–control association studies. *Journal of the American Statistical Association*, 108(502):566–576, 2013. doi: 10.1080/01621459.2013.793121. URL https://doi.org/10.1080/01621459.2013.793121.

R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, L.-Y. Ch'ang, W. Huang, B. Liu, Y. Shen, P. K.-H. Tam, L.-C. Tsui, M. M. Y. Waye, J. T.-F. Wong, C. Zeng, Q. Zhang, M. S. Chee, L. M. Galver, S. Kruglyak, S. S. Murray, A. R. Oliphant, A. Montpetit, T. J. Hudson, F. Chagnon, V. Ferretti, M. Leboeuf, M. S. Phillips, A. Verner, P.-Y. Kwok, S. Duan,

D. L. Lind, R. D. Miller, J. P. Rice, N. L. Saccone, P. Taillon-Miller, M. Xiao, Y. Nakamura, A. Sekine, K. Sorimachi, T. Tanaka, Y. Tanaka, T. Tsunoda, E. Yoshino, D. R. Bentley, P. Deloukas, S. Hunt, D. Powell, D. Altshuler, S. B. Gabriel, H. Zhang, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, T. Aniagwu, P. A. Marshall, O. Matthew, C. Nkwodimmah, C. D. M. Royal, M. F. Leppert, M. Dixon, L. D. Stein, F. Cunningham, A. Kanani, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, P. Donnelly, J. Marchini, G. A. T. McVean, S. R. Myers, L. R. Cardon, G. R. Abecasis, A. Morris, B. S. Weir, J. C. Mullikin, S. T. Sherry, M. Feolo, M. J. Daly, S. F. Schaffner, R. Qiu, A. Kent, G. M. Dunston, K. Kato, N. Niikawa, B. M. Knoppers, M. W. Foster, E. W. Clayton, V. O. Wang, J. Watkin, E. Sodergren, G. M. Weinstock, R. K. Wilson, L. L. Fulton, J. Rogers, B. W. Birren, H. Han, H. Wang, M. Godbout, J. C. Wallenburg, P. L'Archevêque, G. Bellemare, K. Todani, T. Fujita, S. Tanaka, A. L. Holden, E. H. Lai, F. S. Collins, L. D. Brooks, J. E. McEwen, M. S. Guyer, E. Jordan, J. L. Peterson, J. Spiegel, L. M. Sung, L. F. Zacharia, K. Kennedy, M. G. Dunn, R. Seabrook, M. Shillito, B. Skene, J. G. Stewart, D. L. Valle (chair), E. W. Clayton (co chair), L. B. Jorde (co chair), M. K. Cho, T. Duster, M. Jasperse, J. Licinio, J. C. Long, P. N. Ossorio, P. Spallone, S. F. Terry, E. S. Lander (chair), E. H. Lai (co chair), D. A. Nickerson (co chair), M. Boehnke, J. A. Douglas, R. R. Hudson, L. Kruglyak, R. L. Nussbaum, T. I. H. Consortium, G. c. B. C. o. M. BioScience, ParAllele, C. H. Consortium, Illumina, M. U. Centre, G. Q. Innovation, U. o. C. a. S. F. University, Washington, U. o. T. RIKEN, , W. T. S. Institute, W. I. C. f. G. Research, C. e. c. Institute, sample-collection groups: Beijing Normal University, B. Genomics, E. E. I. Health Sciences University of Hokkaido, S. University, H. U. Ibadan, U. of, U. o. Utah, A. G. C. S. H. Laboratory, J. H. U. S. o. Medicine, U. o. Oxford, W. T. C. f. H. G. University of Oxford, U. N. I. o. Health, L. Ethical, S. I. C. A. of Social Sciences, G. I. Group, H. University, K. University, N. University, U. o. Montréal, U. o. Oklahoma, V. University, W. Trust, S. D. B. C. o. Medicine, W. University, S. M. C. A. o. Sciences, C. M. o. S. Technology, , G. Canada, G. Québec, C. S. S. Japanese Ministry of Education, Technology, T. S. Consortium, I. P. G. Populations, L. Ethical, S. I. Group, and M. Group. The international hapmap project. *Nature*, 426(6968):789–796, 2003. ISSN 1476-4687. doi: 10.1038/nature02168. URL https://doi.org/10.1038/nature02168.

J. Gratten and P. M. Visscher. Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Medicine*, 8(1):78, 2016. ISSN 1756-994X. doi: 10.1186/s13073-016-0332-x. URL https://doi.org/10.1186/s13073-016-0332-x.

A. J. Griffiths, J. H. Miller, D. T. Suzuki, R. C. Lewontin, and W. M. Gelbart. Sex chromosomes and sex-linked inheritance. *An Introduction to Genetic Analysis. 7th edition*, 2000. URL https://www.ncbi.nlm.nih.gov/books/NBK22079/. Publisher: W. H. Freeman.

Z. Griliches, M. D. Intriligator, R. F. Engle, D. McFadden, J. J. Heckman, and E. E. Leamer. *Handbook of econometrics*. North-Holland Pub. Co, 1983.

S. Hackinger and E. Zeggini. Statistical methods to detect pleiotropy in human complex traits. *Open Biology*, 7(11), Nov. 2017. ISSN 2046-2441. doi: 10.1098/rsob.170125. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5717338/.

J. W. Hardin. The robust variance estimator for two-stage models. *The Stata Journal*, 2(3):253–266, 2002. doi: 10.1177/1536867X0200200302. URL https://doi.org/10.1177/1536867X0200200302.

F. E. Harrell. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer, 2001.

T. Hastie, J. Friedman, and R. Tisbshirani. *The Elements of statistical learning: data mining, inference, and prediction*. Springer, 2017.

S. J. Hebbring. The challenges, advantages and future of phenome-wide association studies. *Immunology*, 141(2):157–165, Feb. 2014. ISSN 0019-2805. doi: 10.1111/imm.12195. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3904236/.

B. M. Henn, L. R. Botigué, C. D. Bustamante, A. G. Clark, and S. Gravel. Estimating the mutation load in human genomes. *Nature Reviews Genetics*, 16(6):333–343, 2015. ISSN 1471-0064. doi: 10.1038/nrg3931. URL https://doi.org/10.1038/nrg3931.

A. Henningsen. "mvprobit". *CRAN*, 2019. URL https://cran.r-project.org/web/packages/mvProbit/mvProbit.pdf.

B. Hirschler. Cashing in on dna: race on to unlock value in genetic data, Aug 2018. URL https://www.reuters.com/article/us-health-dna/cashing-in-on-dna-race-on-to-unlock-value-in-genetic-data-idUSKBN1KO0XC.

P. D. Hoff and X. Niu. A covariance regression model. *Statistica Sinica*, 22(2):729–753, 2012. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24310032.

B. Howie and J. Marchini. 1,000 genomes haplotypes – phase 3 integrated variant set release in ncbi build 37 (hg19) coordinates. 2015. https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html [Accessed 5-February-2020].

IGSR: The International Genome Sample Resource. 1000 genomes release: Phase 3. 2015. Index of /vol1/ftp/release/20130502/ ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/.

K. Imai, G. King, and O. Lau. Toward a common framework for statistical analysis and development. *Journal of Computational Graphics and Statistics*, 17(4):892–913, 2008. URL http://j.mp/msE15c.

Z. Jin. On some aspects of composite likelihood. *PhD dissertation, University of Toronto*, 2009.

B. Kenkel. Vectorized bivariate normal cdf, 2015. URL https://cran.r-project.org/web/packages/pbivnorm/pbivnorm.pdf.

W. Kent, C. Sugnet, T. Furey, K. Roskin, T. Pringle, A. Zahler, and D. Haussler. Ucsc genome browser: The human genome browser at ucsc, 2002.

P. P. Khil and R. D. Camerini-Otero. Genetic crossovers are predicted accurately by the computed human recombination map. *PLOS Genetics*, 6(1):1–11, 01 2010. doi: 10.1371/journal.pgen.1000831. URL https://doi.org/10.1371/journal.pgen.1000831.

L. Klei, D. Luca, B. Devlin, and K. Roeder. Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genetic Epidemiology*, 32(1):9–19, 2008. doi: 10.1002/gepi.20257. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.20257.

D. Labuda, J.-F. Lefebvre, P. Nadeau, and M.-H. Roy-Gagnon. Female-to-male breeding ratio in modern humans—an analysis based on historical recombinations. *The American Journal of Human Genetics*, 86(3):353–363, Mar 2010. ISSN 0002-9297. doi: 10.1016/j.ajhg.2010.01.029. URL https://doi.org/10.1016/j.ajhg.2010.01.029.

S. Lee, G. Abecasis, M. Boehnke, and X. Lin. Rare-variant association analysis: Study designs and statistical tests. *The American Journal of Human Genetics*, 95(1):5–23, Jul 2014. ISSN 0002-9297. doi: 10.1016/j.ajhg.2014.06.009. URL https://doi.org/10.1016/j.ajhg.2014.06.009.

C. Li and M. Li. Gwasimulator: a rapid whole-genome simulation program. *Bioinformatics*, 24(1):140–142, 2008. doi: 10.1093/bioinformatics/btm549. URL http://dx.doi.org/10.1093/bioinformatics/btm549.

B. Lindsay, G. Yi, and J. Sun. Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, 21, 01 2011.

K. E. Lohmueller, C. L. Pearce, M. Pike, E. S. Lander, and J. N. Hirschhorn. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics*, 33(2):177–182, 2003. ISSN 1546-1718. doi: 10.1038/ng1071. URL https://doi.org/10.1038/ng1071.

D. Lubinski. Exceptional cognitive ability: The phenotype. *Behavior Genetics*, 39 (4):350–358, 2009. ISSN 1573-3297. doi: 10.1007/s10519-009-9273-0. URL https://doi.org/10.1007/s10519-009-9273-0.

M. Luciano, V. Svinti, A. Campbell, R. E. Marioni, C. Hayward, A. F. Wright, M. S. Taylor, D. J. Porteous, P. Thomson, J. G. Prendergast, and et al. Exome sequencing to detect rare variants associated with general cognitive ability: A pilot study. *Twin Research and Human Genetics*, 18(2):117–125, 2015. doi: 10.1017/thg.2015.10.

S. Ly, K.-H. Pho, S. Ly, and W.-K. Wong. Determining distribution for the product of random variables by using copulas. *Risks*, 7(1), 2019. ISSN 2227-9091. doi: 10.3390/risks7010023. URL https://www.mdpi.com/2227-9091/7/1/23.

D. MacKenzie and T. Spears. 'the formula that killed wall street': The gaussian copula and modelling practices in investment banking. *Social Studies of Science*, 44(3):393–417, 2014. doi: 10.1177/0306312713517157. URL https://doi.org/10.1177/0306312713517157. PMID: 25051588.

B. E. Madsen and S. R. Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLOS Genetics*, 5(2):1–11, 02 2009. doi: 10.1371/journal.pgen.1000384. URL https://doi.org/10.1371/journal.pgen.1000384.

G. Marra and R. Radice. "gjrm". *CRAN*, 2019. URL https://cran.r-project.org/web/packages/GJRM/GJRM.pdf.

S. Menard. *Applied Logistic Regression Analysis*. Thousand Oaks, California, 2 edition, Apr. 2002. doi: 10.4135/9781412983433. URL https://methods.sagepub.com/book/applied-logistic-regression-analysis.

G. Moffa and J. Kuipers. Sequential monte carlo em for multivariate probit models. *Computational Statistics & Data Analysis*, 72:252 – 272, 2014. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2013.10.019. URL http://www.sciencedirect.com/science/article/pii/S0167947313003733.

G. Moser, S. H. Lee, B. J. Hayes, M. E. Goddard, N. R. Wray, and P. M. Visscher. Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. *PLOS Genetics*, 11(4):1–22, 2015. doi: 10.1371/journal.pgen.1004969. URL https://doi.org/10.1371/journal.pgen.1004969. Publisher: Public Library of Science.

P. Muir, S. Li, S. Lou, D. Wang, D. J. Spakowicz, L. Salichos, J. Zhang, G. M. Weinstock, F. Isaacs, J. Rozowsky, and M. Gerstein. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome biology*, 17:53–53, Mar. 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0917-0. URL https://pubmed.ncbi.nlm.nih.gov/27009100. Publisher: BioMed Central.

J. Mullahy. Estimation of multivariate probit models via bi-

variate probit. *The Stata Journal*, 16(1):37–51, 2016. URL https://EconPapers.repec.org/RePEc:tsj:stataj:v:16:y:2016:i:1:p:37-51.

K. M. Murphy and R. H. Topel. Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, 3(4):370–379, 1985. ISSN 07350015. URL http://www.jstor.org/stable/1391724.

J. November. More than moore's mores: Computers, genomics,and the embrace of innovation. *Journal of the History of Biology*, 51(4):807–840, 2018. ISSN 1573-0387. doi: 10.1007/s10739-018-9539-6. URL https://doi.org/10.1007/s10739-018-9539-6.

P. F. O'Reilly, C. J. Hoggart, Y. Pomyen, F. C. F. Calboli, P. Elliott, M.-R. Jarvelin, and L. J. M. Coin. MultiPhen: Joint Model of Multiple Phenotypes Can Increase Discovery in GWAS. *PLOS ONE*, 7(5):1–1, 2012. doi: 10.1371/journal.pone.0034861. URL https://doi.org/10.1371/journal.pone.0034861. Publisher: Public Library of Science.

S. H. Park, J. Y. Lee, and S. Kim. A methodology for multivariate phenotype-based genome-wide association studies to mine pleiotropic genes. *BMC Systems Biology*, 5(2):S13, Dec. 2011. ISSN 1752-0509. doi: 10.1186/1752-0509-5-S2-S13. URL https://doi.org/10.1186/1752-0509-5-S2-S13.

G. M. Peloso, D. J. Rader, S. Gabriel, S. Kathiresan, M. J. Daly, and B. M. Neale. Phenotypic extremes in rare variant study designs. *European Journal of Human Genetics*, 24(6):924–930, 2016. ISSN 1476-5438. doi: 10.1038/ejhg.2015.197. URL https://doi.org/10.1038/ejhg.2015.197.

B. Peng and C. I. Amos. Forward-time simulation of realistic samples for genome-wide association studies. *BMC Bioinformatics*, 11(1):442, Sep 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-442. URL https://doi.org/10.1186/1471-2105-11-442.

B. Peng and M. Kimmel. simupop: a forward-time population genetics simulation environment. *Bioinformatics*, 21(18):3686–3687, 2005. doi: 10.1093/bioinformatics/bti584. URL http://dx.doi.org/10.1093/bioinformatics/bti584.

B. Peng and M. Kimmel. Simulations provide support for the common disease-common variant hypothesis. *Genetics*, 175(2):763–776, Feb 2007. ISSN 0016-6731. doi: 10.1534/genetics.106.058164. URL https://pubmed.ncbi.nlm.nih.gov/17151262. 17151262[pmid].

E. Persyn, R. Redon, L. Bellanger, and C. Dina. The impact of a fine-scale population stratification on rare variant association test results. *PloS one*, 13(12):e0207677–e0207677, Dec 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0207677. URL https://pubmed.ncbi.nlm.nih.gov/30521541. 30521541[pmid].

L. Porras-Hurtado, Y. Ruiz, C. Santos, C. Phillips, n. Carracedo, and M. Lareu. An overview of structure: applications, parameter settings, and supporting software. *Frontiers in Genetics*, 4:98, 2013. ISSN 1664-8021. doi: 10.3389/fgene.2013.00098. URL https://www.frontiersin.org/article/10.3389/fgene.2013.00098.

H. F. Porter and P. F. O'Reilly. Multivariate simulation framework reveals performance of multi-trait gwas methods. *Scientific Reports*, 7(1):38837, 2017. ISSN 2045-2322. doi: 10.1038/srep38837. URL https://doi.org/10.1038/srep38837.

A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006. ISSN 1546-1718. doi: 10.1038/ng1847. URL https://doi.org/10.1038/ng1847.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155(2):945–959, 2000. ISSN 0016-6731. URL https://www.genetics.org/content/155/2/945. Publisher: Genetics _eprint: https://www.genetics.org/content/155/2/945.full.pdf.

A. Quintero and E. Lesaffre. Multilevel covariance regression with correlated random effects in the mean and variance structure. *Biometrical Journal*, 59(5):1047–1066, 2017. doi: 10.1002/bimj.201600193. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201600193.

D. J. Rashid, S. C. Chapman, H. C. Larsson, C. L. Organ, A.-G. Bebin, C. S. Merz-

dorf, R. Bradley, and J. R. Horner. From dinosaurs to birds: a tail of evolution. *EvoDevo*, 5(1):25, 2014. ISSN 2041-9139. doi: 10.1186/2041-9139-5-25. URL https://doi.org/10.1186/2041-9139-5-25.

C. Ren-Hua, T. Wei-Yun, H. Chang-Hsun, H. Kuan-Yi, H. C. A., and H. E. R. Seqsimla2: Simulating correlated quantitative traits accounting for shared environmental effects in user-specified pedigree structure. *Genetic Epidemiology*, 39(1):20–24, 2014. doi: 10.1002/gepi.21850. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.21850.

A. Retico, A. Giuliano, R. Tancredi, A. Cosenza, F. Apicella, A. Narzisi, L. Biagi, M. Tosetti, F. Muratori, and S. Calderoni. The effect of gender on the neuroanatomy of children with autism spectrum disorders: a support vector machine case-control study. *Molecular Autism*, 7(1):5, 2016. ISSN 2040-2392. doi: 10.1186/s13229-015-0067-3. URL https://doi.org/10.1186/s13229-015-0067-3.

M. R. Robinson, A. Kleinman, M. Graff, A. A. E. Vinkhuyzen, D. Couper, M. B. Miller, W. J. Peyrot, A. Abdellaoui, B. P. Zietsch, I. M. Nolte, J. V. van Vliet-Ostaptchouk, H. Snieder, B. Z. Alizadeh, H. M. Boezen, L. Franke, P. v. d. Harst, G. Navis, M. Rots, M. Swertz, B. H. R. Wolffenbuttel, C. Wijmenga, G. R. Abecasis, D. Absher, H. Alavere, E. Albrecht, H. L. Allen, P. Almgren, N. Amin, P. Amouyel, D. Anderson, A. M. Arnold, D. Arveiler, T. Aspelund, F. W. Asselbergs, T. L. Assimes, M. Atalay, A. P. Attwood, L. D. Atwood, S. J. L. Bakker, B. Balkau, A. J. Balmforth, C. Barlassina, I. Barroso, H. Basart, S. Bauer, J. S. Beckmann, J. P. Beilby, A. J. Bennett, Y. Ben-Shlomo, R. N. Bergman, S. Bergmann, S. I. Berndt, R. Biffar, A. M. Di Blasio, B. O. Boehm, M. Boehnke, H. Boeing, E. Boerwinkle, J. L. Bolton, A. Bonnefond, L. L. Bonnycastle, D. I. Boomsma, I. B. Borecki, S. R. Bornstein, N. Bouatia-Naji, G. Boucher, J. L. Bragg-Gresham, P. Brambilla, M. Bruinenberg, T. A. Buchanan, C. Buechler, G. Cadby, H. Campbell, M. J. Caulfield, C. Cavalcanti-Proença, G. Cesana, S. J. Chanock, D. I. Chasman, Y.-D. I. Chen, P. S. Chines, D. J. Clegg, L. Coin, F. S. Collins, J. M. Connell, W. Cookson, M. N. Cooper, D. C. Croteau-Chonka, L. A. Cupples, D. Cusi, F. R. Day, I. N. M. Day, G. V. Dedoussis, M. Dei, P. Deloukas, E. T. Dermitzakis, A. S. Dimas, M. Dimitriou, A. L. Dixon, M. Dörr, C. M. van Duijn, S. Ebrahim, S. Edkins, G. Eiriksdottir, K. Eisinger, N. Ek-

lund, P. Elliott, R. Erbel, J. Erdmann, M. R. Erdos, J. G. Eriksson, T. Esko, K. Estrada, D. M. Evans, U. de Faire, T. Fall, M. Farrall, M. F. Feitosa, M. M. Ferrario, T. Ferreira, J. Ferrières, K. Fischer, E. Fisher, G. Fowkes, C. S. Fox, P. W. Franks, R. M. Fraser, F. Frau, T. Frayling, N. B. Freimer, P. Froguel, M. Fu, S. Gaget, A. Ganna, P. V. Gejman, D. Gentilini, E. J. C. Geus, C. Gieger, B. Gigante, A. P. Gjesing, N. L. Glazer, M. E. Goddard, A. Goel, H. Grallert, J. Gräßler, H. Grönberg, L. C. Groop, C. J. Groves, V. Gudnason, C. Guiducci, S. Gustafsson, U. Gyllensten, A. S. Hall, P. Hall, G. Hallmans, A. Hamsten, T. Hansen, T. Haritunians, T. B. Harris, P. van der Harst, A.-L. Hartikainen, N. Hassanali, A. T. Hattersley, A. S. Havulinna, C. Hayward, N. L. Heard-Costa, A. C. Heath, J. Hebebrand, I. M. Heid, M. den Heijer, C. Hengstenberg, K.-H. Herzig, A. A. Hicks, A. Hingorani, A. Hinney, J. N. Hirschhorn, A. Hofman, C. C. Holmes, G. Homuth, J.-J. Hottenga, K. G. Hovingh, F. B. Hu, Y.-J. Hu, J. E. Huffman, J. Hui, H. Huikuri, S. E. Humphries, J. Hung, S. E. Hunt, D. Hunter, K. Hveem, E. Hyppönen, W. Igl, T. Illig, E. Ingelsson, C. Iribarren, B. Isomaa, A. U. Jackson, K. B. Jacobs, A. L. James, J.-O. Jansson, I. Jarick, M.-R. Jarvelin, K.-H. Jöckel, Å. Johansson, T. Johnson, J. Jolley, T. Jørgensen, P. Jousilahti, A. Jula, A. E. Justice, M. Kaakinen, M. Kähönen, E. Kajantie, S. Kanoni, W. H. L. Kao, L. M. Kaplan, R. C. Kaplan, J. Kaprio, K. Kapur, F. Karpe, S. Kathiresan, F. Kee, S. M. Keinanen-Kiukaanniemi, S. Ketkar, J. Kettunen, K.-T. Khaw, L. A. Kiemeney, T. O. Kilpeläinen, L. Kinnunen, M. Kivimaki, M. Kivmaki, M. M. Van der Klauw, M. E. Kleber, J. W. Knowles, W. Koenig, I. Kolcic, G. Kolovou, I. R. König, S. Koskinen, P. Kovacs, P. Kraft, A. T. Kraja, K. Kristiansson, K. Krjutskov, H. K. Kroemer, J. P. Krohn, V. Krzelj, D. Kuh, J. R. Kulzer, M. Kumari, Z. Kutalik, K. Kuulasmaa, J. Kuusisto, K. Kvaloy, M. Laakso, J. H. Laitinen, T. A. Lakka, C. Lamina, C. Langenberg, O. Lantieri, G. M. Lathrop, L. J. Launer, D. A. Lawlor, R. W. Lawrence, I. M. Leach, C. Lecoeur, S. H. Lee, T. Lehtimäki, M. F. Leitzmann, G. Lettre, D. F. Levinson, G. Li, S. Li, L. Liang, D.-Y. Lin, L. Lind, C. M. Lindgren, J. Lindström, J. Liu, A. Liuzzi, A. E. Locke, M.-L. Lokki, C. Loley, R. J. F. Loos, M. Lorentzon, J. Luan, R. N. Luben, B. Ludwig, P. A. Madden, R. Mägi, P. K. E. Magnusson, M. Mangino, P. Manunta, D. Marek, M. Marre, N. G. Martin, T. L. C. Study, and G. I. o. A. T. G. consortium. Genetic evidence of assortative mating in humans. *Nature Human Behaviour*, 1(1):0016, 2017. ISSN 2397-3374. doi: 10.1038/s41562-016-0016. URL `https://doi.org/10.1038/s41562-016-0016`.

L. Rönnegård and W. Valdar. Detecting major genetic loci controlling phenotypic variability in experimental crosses. *Genetics*, 188(2):435–447, 2011. ISSN 0016-6731. doi: 10.1534/genetics.111.127068. URL http://www.genetics.org/content/188/2/435.

J. G. Ruby, K. M. Wright, K. A. Rand, A. Kermany, K. Noto, D. Curtis, N. Varner, D. Garrigan, D. Slinkov, I. Dorfman, J. M. Granka, J. Byrnes, N. Myres, and C. Ball. Estimates of the heritability of human longevity are substantially inflated due to assortative mating. *Genetics*, 210(3):1109–1124, 2018. ISSN 0016-6731. doi: 10.1534/genetics.118.301613. URL https://www.genetics.org/content/210/3/1109.

Z. Sajaia. Maximum likelihood estimation of a bivariate ordered probit model: Implementation and monte carlo simulations. *The Stata Journal*, ii:1–18, 01 2009.

Y. D. Salinas, Z. Wang, and A. T. DeWan. Statistical Analysis of Multiple Phenotypes in Genetic Epidemiologic Studies: From Cross-Phenotype Associations to Pleiotropy. *American Journal of Epidemiology*, 187(4):855–863, Apr. 2018. ISSN 0002-9262. doi: 10.1093/aje/kwx296. URL https://academic.oup.com/aje/article/187/4/855/4081580. Publisher: Oxford Academic.

F. Salmon. The formula that killed wall street. *Significance*, 9 (1):16–20, 2012. doi: 10.1111/j.1740-9713.2012.00538.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1740-9713.2012.00538.x.

J. R. Schott. Some tests for the equality of covariance matrices. *Journal of Statistical Planning and Inference*, 94(1):25 – 36, 2001. ISSN 0378-3758. doi: https://doi.org/10.1016/S0378-3758(00)00209-3. URL http://www.sciencedirect.com/science/article/pii/S0378375800002093.

B. Servin and M. Stephens. Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits. *PLOS Genetics*, 3(7):1–13, 2007. doi: 10.1371/journal.pgen.0030114. URL https://doi.org/10.1371/journal.pgen.0030114. Publisher: Public Library of Science.

H. Shim, D. I. Chasman, J. D. Smith, S. Mora, P. M. Ridker, D. A. Nickerson,

R. M. Krauss, and M. Stephens. A Multivariate Genome-Wide Association Analysis of 10 LDL Subfractions, and Their Response to Statin Treatment, in 1868 Caucasians. *PLOS ONE*, 10(4):1–20, 2015. doi: 10.1371/journal.pone.0120758. URL https://doi.org/10.1371/journal.pone.0120758. Publisher: Public Library of Science.

D. H. Skuse. X-linked genes and mental functioning. *Human Molecular Genetics*, 14 (suppl_1):R27–R32, Apr. 2005. ISSN 0964-6906. doi: 10.1093/hmg/ddi112. URL https://academic.oup.com/hmg/article/14/suppl_1/R27/560887. Publisher: Oxford Academic.

G. W. Snedecor and W. G. Cochran. *Statistical methods*. Iowa State University Press, 1989.

M. Stephens. A Unified Framework for Association Analysis with Multiple Related Phenotypes. *PLOS ONE*, 8(7):1–19, 2013. doi: 10.1371/journal.pone.0065245. URL https://doi.org/10.1371/journal.pone.0065245. Publisher: Public Library of Science.

Z. Su, J. Marchini, and P. Donnelly. Hapgen2: simulation of multiple disease snps. *Bioinformatics*, 27(16):2304–2305, 2011. doi: 10.1093/bioinformatics/btr341. URL http://dx.doi.org/10.1093/bioinformatics/btr341.

C. S. Tang and M. A. R. Ferreira. A gene-based test of association using canonical correlation analysis. *Bioinformatics*, 28(6):845–850, 01 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts051. URL https://doi.org/10.1093/bioinformatics/bts051.

A. Teixeira-Pinto and S.-L. T. Normand. Correlated bivariate continuous and binary outcomes: Issues and applications. *Statistics in Medicine*, 28(13):1753–1773, 2009. doi: 10.1002/sim.3588. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3588.

S. Trent and W. Davies. The influence of sex-linked genetic mechanisms on attention and impulsivity. *Biological Psychology*, 89(1):1 – 13, 2012. ISSN 0301-0511. doi: https://doi.org/10.1016/j.biopsycho.2011.09.011. URL http://www.sciencedirect.com/science/article/pii/S0301051111002456.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood meth-

ods. *Statistica Sinica*, 21(1):5–42, 2011. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24309261.

S. Vattikuti, J. J. Lee, C. C. Chang, S. D. H. Hsu, and C. C. Chow. Applying compressed sensing to genome-wide association studies. *GigaScience*, 3(1), 06 2014. ISSN 2047-217X. doi: 10.1186/2047-217X-3-10. URL https://doi.org/10.1186/2047-217X-3-10. 2047-217X-3-10.

P. Visscher, M. Brown, M. McCarthy, and J. Yang. Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24, Jan 2012. ISSN 0002-9297. doi: 10.1016/j.ajhg.2011.11.029. URL https://doi.org/10.1016/j.ajhg.2011.11.029.

P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 10 years of gwas discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, Jul 2017. ISSN 0002-9297. doi: 10.1016/j.ajhg.2017.06.005. URL https://doi.org/10.1016/j.ajhg.2017.06.005.

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912526.

J. H. Wilson and T. Hunt. *Molecular biology of the cell, 4th edition: a problems approach*. Garland Science, 2002.

F. A. Wright, H. Huang, X. Guan, K. Gamiel, C. Jeffries, W. T. Barry, F. Pardo-Manuel de Villena, P. F. Sullivan, K. C. Wilhelmsen, and F. Zou. Simulating association studies: a data-based resampling method for candidate regions or whole genome scans. *Bioinformatics*, 23(19):2581–2588, 2007. doi: 10.1093/bioinformatics/btm386. URL http://dx.doi.org/10.1093/bioinformatics/btm386.

M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, Jul 2011. ISSN 0002-9297. doi: 10.1016/j.ajhg.2011.05.029. URL https://doi.org/10.1016/j.ajhg.2011.05.029.

J. XiaoCan, Y. YongLi, C. YuanCheng, C. ZhiWei, D. Yuhui, X. Zhenhua, Z. Weiping, X. Chao, Z. Qiang, X. Xin, D. HongWen, and S. XueZhong. Multivariate analysis of genome-wide data to identify potential pleiotropic genes for five major psychiatric disorders using MetaCCA. *Journal of affective disorders*, 242: 234–243, Jan. 2019. ISSN 0165-0327. doi: 10.1016/j.jad.2018.07.046. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6343670/.

Y. Xu, Y. Wu, C. Song, and H. Zhang. Simulating realistic genomic data with rare variants. *Genetic Epidemiology*, 37(2):163–172, 2013. doi: 10.1002/gepi.21696. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.21696.

J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher. Gcta: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76 – 82, 2011. ISSN 0002-9297. doi: https://doi.org/10.1016/j.ajhg.2010.11.011. URL http://www.sciencedirect.com/science/article/pii/S0002929710005987.

J. Yang, A. Bakshi, Z. Zhu, G. Hemani, A. A. E. Vinkhuyzen, S. H. Lee, M. R. Robinson, J. R. B. Perry, I. M. Nolte, J. V. van Vliet-Ostaptchouk, H. Snieder, T. Esko, L. Milani, R. Mägi, A. Metspalu, A. Hamsten, P. K. E. Magnusson, N. L. Pedersen, E. Ingelsson, N. Soranzo, M. C. Keller, N. R. Wray, M. E. Goddard, P. M. Visscher, and T. L. C. Study. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics*, 47(10):1114–1120, 2015. ISSN 1546-1718. doi: 10.1038/ng.3390. URL https://doi.org/10.1038/ng.3390.

Q. Yang and Y. Wang. Methods for Analyzing Multivariate Phenotypes in Genetic Association Studies, 2012. URL https://www.hindawi.com/journals/jps/2012/652569/. ISSN: 1687-952X Library Catalog: www.hindawi.com Pages: e652569 Publisher: Hindawi Volume: 2012.

C. Y.-J. Yim Seon-Hee. Reflections on the us fda's warning on direct-to-consumer genetic testing. *Genomics Inform*, 12(4):151–155, 2014. doi: 10.5808/GI.2014.12.4.151. URL http://genominfo.org/journal/view.php?number=100.

A. I. Young, F. L. Wauthier, and P. Donnelly. Identifying loci affecting trait variability and detecting interactions in genome-wide association studies. *Nature Genetics*, 50(11):1608–1614, 2018. ISSN 1546-1718. doi: 10.1038/s41588-018-0225-6. URL https://doi.org/10.1038/s41588-018-0225-6.

X. Yuan, D. J. Miller, J. Zhang, D. Herrington, and Y. Wang. An overview of population genetic data simulation. *Journal of Computational Biology*, 19(1):42–54, 2012. doi: 10.1089/cmb.2010.0188. URL https://doi.org/10.1089/cmb.2010.0188. PMID: 22149682.

F. Zhang, J. Liu, J. Chen, and H.-W. Deng. Hapsimu: a genetic simulation platform for population-based association studies. *BMC Bioinformatics*, 9(1):331, Aug 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-331. URL https://doi.org/10.1186/1471-2105-9-331.

Z. Zhang, X. Li, X. Ding, J. Li, and Q. Zhang. Gpopsim: a simulation tool for whole-genome genetic data. *BMC Genetics*, 16(1):10, Feb 2015. ISSN 1471-2156. doi: 10.1186/s12863-015-0173-4. URL https://doi.org/10.1186/s12863-015-0173-4.

Y. Zhao and H. Joe. Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics*, 33(3):335–356, 2005. doi: 10.1002/cjs.5540330303. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cjs.5540330303.

Y.-H. Zhou, J. S. Marron, and F. A. Wright. Computation of ancestry scores with mixed families and unrelated individuals. *Biometrics*, 74(1):155–164, 2017. doi: 10.1111/biom.12708. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12708.

L. Zhu, J. Lei, B. Devlin, and K. Roeder. Testing high-dimensional covariance matrices, with application to detecting schizophrenia risk genes. *Ann. Appl. Stat.*, 11(3):1810–1831, 09 2017. doi: 10.1214/17-AOAS1062. URL https://doi.org/10.1214/17-AOAS1062.

M. J. Zuidhof, B. L. Schneider, V. L. Carney, D. R. Korver, and F. E. Robinson. Growth, efficiency, and yield of commercial broilers from 1957, 1978, and 20051 1This is an Open Access article distributed under the terms of the Creative Commons Attribution-Noncommercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits noncommercial use, distribution, and reproduction in any medium,

provided the original work is properly cited. *Poultry Science*, 93(12):2970 – 2982, 2014. ISSN 0032-5791. doi: https://doi.org/10.3382/ps.2014-04291. URL http://www.sciencedirect.com/science/article/pii/S0032579119385505.