METHODS FOR MULTI-TRAIT POLYGENIC RISK SCORES

METHODS FOR MULTI-TRAIT POLYGENIC RISK SCORES

By YI WAN, B.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Master of Science

McMaster University

MASTER OF SCIENCE (2024)

Hamilton, Ontario, Canada (Department of Mathematics and Statistics)

TITLE:	Methods for Multi-Trait Polygenic Risk Scores
AUTHOR:	Yi Wan B.Sc. (Mathematics and Statistics), McMaster University, Hamilton, Canada
SUPERVISOR:	Dr. Angelo J. Canty
NUMBER OF PAGES:	xv, 83

Abstract

This thesis examines various methods for generating multi-trait polygenic risk scores (PRS). The primary objective is to see which multi-trait method performs best and are there any simpler methods that can perform as well. The thesis evaluates each method by comparing the weighted-average multi-trait PRS with true phenotype values (target traits), using the correlation coefficient (ρ) for continuous traits and the area under the receiver operating characteristic curve (AUC) for binary traits. It also investigates how different simulation parameters influence performance. Two additional novel multi-trait PRS methods are introduced in this work: **mt-lm** and **mt-CVb**. **mt-lm** is essentially a multiple linear regression for a continuous focal trait and logistic regression for a binary focal trait, while **mt-CVb** combines cross-validation and bagging techniques in a hybrid approach to improve model performance. The existing multi-trait method **wMT-SBLUP** consistently achieves the best performance, outperforming all other methods in most scenarios. While the two novel methods are not the top performers, they demonstrate better results compared to other methods (excluding **wMT-SBLUP**) for both continuous and binary focal traits across various parameter settings. Moreover, **mt-lm** offers the additional advantage of being faster than **wMT-SBLUP**.

Acknowledgements

I would first like to express my deepest gratitude to my supervisor, Dr. Angelo J. Canty, he introduced me to the interesting field of statistical genetics. I had the privilege of being his student during my undergraduate studies at McMaster University, Dr. Canty has always been an exceptionally patient, responsible, and dedicated teacher who genuinely hopes that students learn and benefit from his instruction. He gave me the opportunity to conduct undergraduate research and supervised my summer research before my master's program officially began. I am deeply appreciative of his guidance and support, he always encouraged me when I lacked confidence. He guided me in gradually learning how to read the literature, think critically, and solve problems independently. By providing hints rather than direct answers, which deepened my understanding of the concepts and helped me retain them better in my mind. When I shared my experiences with my mother, who is also a professor, she remarked on how lucky I was to have such an outstanding supervisor, and I totally agree.

I also appreciate Dr. Fred Hoppe, Dr. Shui Feng, Dr. Benjamin Bolker, Dr. Xinyu Zhao, and Dr. Katherine Davies. They answered numerous questions during both my undergraduate and graduate studies.

Then I would like to appreciate my parents, both of whom hold PhDs and have

been instrumental in nurturing my passion for research. From a young age, I was influenced by their work and dedication to their respective fields. Their unwavering support has been a constant throughout my academic journey, from elementary school to my master's degree. My father, whose research area is pure mathematics, often helped tutor me, while my mother provided invaluable insights into genetics from a biological perspective, directly relevant to my field of study. They've been like my personal teaching assistants from elementary school to my master's program, lol. And of course, a big thank you to my grandmother too!

Last I would like to thank my boyfriend, who has always been by my side. As a fellow master's student at this university, we attend classes together, work as TAs together. Although our research areas differ, we support each other by working side by side on our respective tasks. And all my best friends, who have been a tremendous help in both my studies and personal life, their encouragement and support have been invaluable.

愿同学们前程似锦！

Contents

List of Figures

List of Tables

Definitions and Abbreviations

Abbreviations

SNP	Single nucleotide polymorphism
MAF	Minor allele frequency
GWAS	Genome-wide association studies
PRS	Polygenic risk scores
ROC	Receiver operating characteristic
AUC	Area under the ROC curve

Definitions

DNA Deoxyribonucleic acid is composed of two long strands that form a double helix, with a backbone made of sugars and phosphate groups, and nucleotide bases (adenine (A), thymine (T), cytosine (C), guanine (G)) that pair between the two strands (A with T,

C with G). The genome is stored within the cell nucleus as DNA organized into chromosomes.

Chromosomes The human genome is divided into 23 pairs of chromosomes, each of which contains long sequences of DNA. Each chromosome contains hundreds to thousands of genes, chromosome 1 is the largest human chromosome.

Gene A gene is a sequence of DNA that contains the instructions for making a specific protein.

Genome The total complement of DNA within a single cell of an organism. Every living organism (from bacteria to humans) has a genome, though the size and complexity can vary significantly across species.

SNP Single nucleotide polymorphism (SNP) is the most common type of genetic variation among individuals. A SNP occurs when a single nucleotide (the basic building block of DNA) in the genome is altered.

Allele Alleles are different versions of a gene or genetic variant. For a specific genetic locus (position on a chromosome), an individual can have two alleles (one inherited from each parent).

Homozygous The individual has two identical alleles (e.g. AA or GG).

Heterozygous The individual has two different alleles (e.g. AG).

Minor allele Minor allele refers to the less frequent of the two alleles at a particular genetic locus in a given population.

MAF The minor allele frequency (MAF) is the proportion of the minor allele in a population.

Trait A specific characteristic or feature of an organism that can be influenced by genetics and environmental factors.

Genotype When talking about SNPs, the genotype indicates the particular alleles present at a given location in the genome.

Phenotype An observable trait.

GWAS A genome-wide association study is used to identify associations between genetic variants and traits or diseases across the genome, compare the genetic variants of individuals with a particular trait or condition to those without it, aiming to identify specific genetic loci that contribute to the trait or disease.

PRS A polygenic risk score is used to quantify an individual's genetic risk for a particular trait or disease based on the combined effect of many genetic variants, is a measure of an individual's genetic predisposition to a certain trait or disease. It is calculated using information from many SNPs across the genome.

Chapter 1

Introduction

1.1 Genetics

Single nucleotide polymorphisms (SNPs), are the most common genomic variant in humans. These are locations in the genome where there is variation in the nucleotide across individuals. Alleles are the possible nucleotides, and at any given genetic locus (position on a chromosome), an individual inherits two alleles (one from each parent). The minor allele frequency (MAF) is the proportion of the minor allele in a population. An individual's genotype at a given SNP tells which alleles they have at that location. A trait (phenotype) is a specific observable characteristic or feature of an organism that can be influenced by both genetics and environmental factors.

A genome-wide association study (GWAS) aims to identify associations between genetic variants and traits or diseases across the entire genome. For binary traits, it compares the genetic variants of individuals with a specific trait or condition to those without it, identifying genetic loci contributing to the trait or disease. For continuous traits, GWAS examines how genetic variants correlate with variation in the

trait across individuals. The input for GWAS always includes both genotype and phenotype data.

In genetics, the heritability of a phenotype is defined as:

$$H^2 = \frac{\text{Var(Genotype)}}{\text{Var(Phenotype)}}$$

which quantifies the proportion of phenotypic variation that is due to genetic variation. The genetic correlation matrix is used to show the genetic correlation between traits. Traits which share causal SNPs will have genetic correlation, and the correlation will increase if more causal SNPs are shared between the traits; additionally, altering the linkage disequilibrium (LD) structure can contribute to genetic correlation, as the same LD block influences both traits.

Polygenic risk scores (PRS) are used to quantify an individual's genetic risk for a particular trait or disease based on the combined effect of many genetic variants, and are calculated using information from many SNPs across the genome. A PRS is typically a weighted sum of genotypes at SNPs thought to be associated with the trait of interest.

1.2 Uni-trait PRS

The uni-trait PRS method estimates genetic risk for a single trait or disease by aggregating the effects of genetic variants associated with that trait. The utility of a PRS relies on the assumption that the trait's genetic architecture is polygenic, meaning it is influenced by numerous genetic variants, each with a small effect; it also assumes that these genetic effects are additive and that the associations identified in

GWAS are replicable and applicable to the target population. Typically, uni-trait PRS are constructed using summary statistics from a GWAS specific to the trait. A common method for performing a GWAS involves fitting a linear regression model for continuous traits, expressed as:

$$P = G \cdot \beta + \epsilon$$

where P represents the phenotype, β denotes the effect sizes, and G represents the genotype. For binary traits, logistic regression is used:

$$\log\left(\frac{\pi}{1-\pi}\right) = G \cdot \beta + \epsilon$$
$$\pi(G) = \frac{e^{G\cdot\beta+\epsilon}}{1+e^{G\cdot\beta+\epsilon}}$$

where $\pi(G)$ is the probability of the trait's occurrence for an individual with genotype vector G, the probability of the trait differs depending on the individual's genotypes. GWAS analysis provides outputs such as Z-scores, standard errors, effect size estimates, and p-values for each SNP and trait. Calculating the uni-trait PRS directly using the formula $G\cdot\hat{\beta}$, where the genotype G is multiplied by the effect size estimates $\hat{\beta}$, is the simplest method for performing uni-trait PRS.

We typically select "significant" SNPs from GWAS results for use in polygenic risk score (PRS) construction. The clumping and thresholding (C+T) method achieves this by first clumping to remove SNPs in high linkage disequilibrium (LD), retaining only the SNP with the lowest p-value within each LD block; next, thresholding is applied to select SNPs with p-values below a specified cutoff. This process ensures that the final set of SNPs is both statistically significant and largely independent.

In other words, the C+T method filters SNPs based on their statistical significance and LD structure to retain only independent significant SNPs. In this thesis, for simplicity, independent SNPs were generated in the simulation discussed in Chapter 3 and Chapter 4. However, in Chapter 5, correlated SNPs were simulated to better reflect real-world scenarios. We are conducting simulations instead of using real-life datasets because we aim to examine how the parameters influence performance.

There are alternative methods for performing uni-trait PRS besides using GWAS estimates directly, such as **snpboost** and **GraBLD**. **snpboost** (Klinkhammer et al., 2023) uses advanced machine learning techniques to improve PRS prediction by integrating various types of genetic data, enhancing the prediction of complex traits or diseases. **GraBLD** (Paré et al., 2017) combines gradient boosting with linkage disequilibrium adjustments to refine PRS and other predictive models. This method utilizes gradient boosted regression trees to optimize SNP weights, followed by a regional adjustment for linkage disequilibrium, leveraging the strengths of both techniques to improve predictive accuracy.

1.3 Multi-trait PRS

A multi-trait PRS method extends beyond focusing on a single trait by simultaneously incorporating multiple related traits. Similar to the uni-trait method, constructing a multi-trait PRS involves using data from GWAS; for multi-trait methods, GWAS data is required for every trait involved. However, multi-trait methods leverage additional traits to enhance the predictive performance of the final PRS. By using one or more additional traits to help predict the final PRS, these methods improve overall predictive power. In order for a multi-trait method to outperform uni-trait method,

the additional traits should have a significant genetic correlation with the focal trait, meaning they share a substantial portion of their genetic architecture.

There are several methods for performing multi-trait PRS, each combining uni-trait PRS for multiple traits to generate a final PRS. The method **MPS** (Krapohl et al., 2018) uses elastic net regularized regression to predict outcomes, selecting predictors and estimating their contributions. The **multi-PGS** method (Albiñana et al., 2023) uses standardized PRS to develop prediction models for a target outcome, applying both a linear model (lasso-penalized regression) and a non-linear model (boosted gradient trees, XGBoost). The **mtPGS** method from Xu et al. (2023) incorporates cross-validation to refine predictions. The **wMT-SBLUP** method, as described in Maier et al. (2018), combines the information from multiple traits by using best linear unbiased predictors (BLUPs) in the GWAS and deriving weights from the BLUP estimators. Finally, the **mtPRS-PCA** method from Zhai et al. (2023) uses the sum of the top eigenvectors from principal component analysis of the genetic correlation matrix.

In this thesis, we focus on comparing **mtPGS**, **wMT-SBLUP**, and **mtPRS-PCA** as described in Chapter 2. **MPS** is particularly suited for large datasets with many predictors. On the other hand, **multi-PGS** is designed to optimize a single target outcome, rather than integrating multiple traits, and uses 937 PRS to help improve prediction in their paper. For these reasons, we chose not to include these two methods in the simulation.

Additional methods, both uni-trait and multi-trait, are reviewed in Ma and Zhou (2021). The key figure in this paper provides an overview of PRS methods up to 2021, offering a comprehensive summary of PRS and their applications in predicting

complex traits.

A very interesting paper (Kachuri et al., 2024) explores the principles and methods of applying PRS across global populations. Genetic architectures can differ among populations due to varying allele frequencies, linkage disequilibrium patterns, and environmental interactions. These differences can impact the accuracy of PRS when applied to diverse groups. The paper addresses the challenges and strategies for transferring PRS across different populations, emphasizing the need for accurate and equitable risk assessments. Additionally, the application of PRS in pharmacogenomics (PGx) studies shows promising potential for enhancing patient stratification and drug response predictions; for example, the method **PRS-PGx-Bayesx** (Zhai et al., 2024) offers a new Bayesian approach to reduce Eurocentric or cross-population PRS prediction biases.

1.4 Overview

Chapter 2 presents three multi-trait methods from the literature that have been applied in this thesis. Chapter 3 describes the simulation setup and presents the results for scenarios involving both continuous and binary focal traits, with varied parameters. Chapter 4 introduces two novel multi-trait methods and compares their performance with the best of the methods discussed in Chapter 2. Chapter 5 describes a simulation in which there are correlated SNPs and discusses how the results differ from independent SNPs as simulated in Chapter 3 and Chapter 4. Chapter 6 summarizes the results and discusses the limitations of the work presented in this thesis.

Chapter 2

Multi-trait Polygenic Risk Score Methods from Literature

These existing multi-trait methods we chose and included in this chapter use genome-wide association studies (GWAS) as input to generate polygenic risk scores (PRS). Each method incorporates additional traits to improve the prediction of the focal (target) trait. The weights for each PRS are first estimated, allowing us to compute the final weighted average PRS by combining the corresponding individual PRS values. The input for GWAS consists of a genotype matrix and trait vectors, with GWAS analysis fitting models (e.g., a linear model or a generalized linear model) to produce outputs such as Z-scores, standard errors, effect size estimates, and p-values for each single nucleotide polymorphism (SNP) and trait.

Researchers typically use external GWAS data to ensure independence between datasets and because they may lack access to large sample datasets. Publicly available GWAS summary statistics, often derived from studies with very large sample sizes, are commonly used for this purpose.

2.1 mtPGS

The method **mtPGS** from Xu et al. (2023) calculates the single PRS by combing different PRS using:

$$\text{mtPGS} = \sum_{k=2}^{K} w_k \cdot \text{PRS}_k \tag{2.1.1}$$

where the weights $w = (w_2, ..., w_K)$ are obtained using weighted regression through cross-validation and $\text{PRS}_k = G \cdot \hat{\beta}_k$ is PRS for the k^{th} trait (G is the genotype vector for individual and $\hat{\beta}_k$ is the vector of estimates for the effect of the SNPs on the k^{th} trait). The PRS for the target trait is not used in the calculation in Equation 2.1.1 (assuming that the target trait is always labeled as the first trait). The goal of **mtPGS** is to develop a robust prediction model by combining genetic information from multiple traits that are likely to share genetic influences; this is why it does not use the PRS for the target trait, as using a separate set of traits helps capture the underlying genetic correlations among these traits and their contributions to the target outcome.

The cross-validation model is trained on $i-1$ of the i folds (the training set) and tested on the remaining fold (the validation set), where i is the number of equally sized (or roughly equal) subsets (or "folds") into which the dataset is randomly divided. This process is repeated i times, each time using a different fold as the validation set and the remaining $i-1$ folds for training.

The model is then fitted through multiple linear regression. 10-fold cross-validation is a commonly used default choice as it provides a good balance between bias and variance while remaining computationally feasible in most scenarios (Kuhn, 2013).

We use 10-fold cross-validation in this thesis. The estimates (weights) obtained from multiple linear regression, where the target trait is the outcome and the other PRS values for the other traits are the covariates in each fold, are used to calculate the weights. The final weights are the average of the estimates across all folds.

2.2 mtPRS-PCA

The method **mtPRS-PCA** from Zhai et al. (2023) calculates the weights for the combination of PRS from multiple traits using the principal component analysis (PCA) method and creates a composite single PRS, which has GWAS summary statistics for each trait and genetic correlation matrix as its inputs using formula:

$$\text{mtPRS-PCA} = \sum_{k=1}^{K} w_k \cdot \text{PRS}_k \tag{2.2.1}$$

where the weights $w = (w_1, ..., w_K)$ are created using genetic correlation matrix by PCA, calculated the cumulative sum of eigenvectors and stop at the sum of top eigenvectors that explained 80% of the genetic variability. In other words, the method for calculating weights involves first determining the number of components that explain at least 80% of the variance, then the loadings of the top principal components are extracted and the loadings across these components are summed to obtain the weights.

The inputs for this method are GWAS summary statistics and the genetic correlation matrix. For real data we can estimate the genetic correlation matrix using the linkage disequilibrium (LD) score regression approach (Bulik-Sullivan et al., 2015).

2.3 wMT-SBLUP

The method **wMT-SBLUP** from Maier et al. (2018) is designed to improve genetic prediction by leveraging genetic correlations among multiple traits. The full name of this method is "weighted index to generate approximate multi-trait summary statistics best linear unbiased predictor", calculates the multi-trait PRS using the formula:

$$\text{wMT-SBLUP}_{i,f} = \sum_{k=1}^{K} w_k \cdot \hat{g}_{i,k}^{\text{BLUP}} \tag{2.3.1}$$

Considered a general linear mixed model $P = G \cdot \beta + \epsilon$ for the GWAS (where β is random effect), with $\text{Var}(\beta) = B$ and $\text{Var}(\epsilon) = R$. For each trait (total k traits), P is a column vector of length $n \times 1$ and G has dimension $n \times p$. Assuming β is an $p \times 1$ true effect sizes vector for that trait with distribution $\beta \sim N(0, I_p \cdot \sigma_\beta^2)$, where I_p is an identity matrix of dimension p; ϵ is a column vector of independent residual effects, $\epsilon \sim N(0, I_n \cdot \sigma_\epsilon^2)$, where I_n is an identity matrix of dimension n.
We now have:

$$\begin{aligned}
B &= I_p \cdot \sigma_\beta^2 \\
R &= I_n \cdot \sigma_\epsilon^2 \\
\hat{\beta}_{\text{BLUP}} &= \left[G' R^{-1} G + B^{-1} \right]^{-1} G' R^{-1} P \\
\hat{g}_{i,k}^{\text{BLUP}} &= G \cdot \hat{\beta}_{\text{BLUP}_k}
\end{aligned}$$

the genetic prediction for each individual i and focal trait of interest f is obtained by combining the genetic predictions for each trait k, $\hat{g}_{i,k}^{\text{BLUP}}$, using the corresponding weights w_k.

Generally, GWAS are performed using a linear regression model, but this method differs from others by using a linear mixed-effects model, their estimates are $\hat{\beta}_{\text{BLUP}}$ (best linear unbiased prediction) instead of $\hat{\beta}_{\text{OLS}}$ (ordinary least squares), the former estimates the parameters of the error distribution and derives a linear unbiased predictor for the random effect sizes. One of the reference in Maier et al. (2018) describes the BLUP method for GWAS and its advantages (Goddard et al., 2009). In a typical GWAS, separate linear regression models are fitted for each SNP, resulting in as many models as there are SNPs; each model estimates the effect size of a single SNP while conditioning on covariates. In contrast, the BLUP approach fits a single mixed-effects model that accounts for the effects of all SNPs simultaneously. Instead of estimating individual SNP effects, the BLUP model estimates the distribution of SNP effect sizes. Furthermore, standard GWAS often focuses on explicitly estimating individual SNP effect sizes, which can lead to sparse models that retain only SNPs with significant p-values; on the other hand, the BLUP method efficiently incorporates information from all SNPs, including those with small or non-significant effects that may still contribute to the overall polygenic signal. That is, the BLUP approach helps mitigate overfitting and addresses the issue of multiple testing inherent in GWAS.

This method uses all the traits, including the focal trait, to generate the multi-trait PRS. The weights are calculated using the heritability and genetic correlation matrix, according to the formula:

$$
\begin{aligned}
w &= V^{-1}C \\
&= \begin{bmatrix} \mathrm{Var}(\hat{\beta}_{\mathrm{SBLUP}_1}) & \cdots & \mathrm{Cov}(\hat{\beta}_{\mathrm{SBLUP}_1}, \hat{\beta}_{\mathrm{SBLUP}_k}) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(\hat{\beta}_{\mathrm{SBLUP}_k}, \hat{\beta}_{\mathrm{SBLUP}_1}) & \cdots & \mathrm{Var}(\hat{\beta}_{\mathrm{SBLUP}_k}) \end{bmatrix}^{-1} C
\end{aligned} \tag{2.3.2}
$$

with

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_{\mathrm{SBLUP}_k}) &= R_k^2 \\
\mathrm{Cov}(\hat{\beta}_{\mathrm{SBLUP}_k}, \hat{\beta}_{\mathrm{SBLUP}_l}) &= \frac{r_G R_k^2 R_l^2}{\sqrt{h_k^2 h_l^2}}
\end{aligned}
$$

where h_k^2 is the heritability for trait k, r_G is the corresponding genetic correlation between focal trait and selected trait, and note that:

$$
\begin{aligned}
R_k^2 &= \frac{\phi + h_k^2 - \sqrt{(\phi + h_k^2)^2 - 4\phi h_k^4}}{2\phi} \\
\phi &= \frac{M_{\mathrm{eff}}}{N_k}
\end{aligned}
$$

where M_{eff} is the effective number of chromosome segments or the number of independent SNPs (for independent SNPs $M_{\mathrm{eff}} = p$ and for correlated SNPs in this thesis $M_{\mathrm{eff}} = \frac{p}{d}$, where p is number of SNPs and d is the block size), N_k is the sample size of trait k (sample size of the training set $N_k = 900$). The vector C represents the covariance vector, with the following formula, where the subscript f denotes the focal trait:

$$C = \begin{bmatrix} \text{Cov}(\beta_f, \hat{\beta}_{\text{SBLUP}_1}) \\ \vdots \\ \text{Cov}(\beta_f, \hat{\beta}_{\text{SBLUP}_k}) \end{bmatrix} \tag{2.3.3}$$

where

$$\text{Cov}(\beta_f, \hat{\beta}_{\text{SBLUP}_f}) = R_f^2$$
$$\text{Cov}(\beta_f, \hat{\beta}_{\text{SBLUP}_k}) = r_G \sqrt{\frac{h_f^2}{h_k^2}} R_k^2$$

In practice, many of the publicly available GWAS summary statistics use $\hat{\beta}_{\text{OLS}}$ instead of $\hat{\beta}_{\text{BLUP}}$; this makes it challenging to calculate PRS using $\hat{\beta}_{\text{BLUP}}$. Additionally, in this thesis, all SNPs were included in the simulation without applying any selection process, which reduces the advantage of BLUP over OLS. As a result, the overall difference between the two methods is minimal, and this thesis uses $\hat{\beta}_{\text{OLS}}$ in place of $\hat{\beta}_{\text{BLUP}}$.

Chapter 3

Comparison of Existing Methods

3.1 Simulation set-up

Three genotype matrices were generated for this simulation: the $n_g \times p$ genotype matrix for GWAS (G_g), the $n_t \times p$ genotype matrix for the training set (G_t), and the $n_v \times p$ genotype matrix for the validation set (G_v). In the $n \times p$ genotype matrix G:

$$G = \begin{bmatrix} \text{SNP}_1 & \text{SNP}_2 & \text{SNP}_3 & \dots & \text{SNP}_p \\ \vdots & \vdots & \vdots & \dots & \vdots \end{bmatrix}$$

where each row represents an individual and each column represents single nucleotide polymorphism (SNP), the values (0, 1, and 2 in each SNP represent the number of minor alleles at a given locus) in the matrix are the observed genotypes for an individual at a given SNP. There are a total of $n_g = 1800$ observations for G_g, $n_t = 900$ observations for G_t, and $n_v = 900$ observations for G_v; the total number of SNPs is $p = 5000$. In this simulation, the minor allele frequency (MAF) are different across SNPs, are generated from a Uniform$(0.05, 0.5)$ distribution.

In the $n \times k$ phenotype matrix P:

$$P = \begin{bmatrix} \text{Trait}_1 & \text{Trait}_2 & \text{Trait}_3 & \dots & \text{Trait}_k \\ \vdots & \vdots & \vdots & \dots & \vdots \end{bmatrix}$$

there are a total of n observations and k columns of traits. For the model used to generate continuous traits in this thesis:

$$P = G \cdot \beta + \epsilon$$

where P represents the phenotype, β denotes the effect sizes, and G represents the genotype matrix. The model used to generate binary traits in this thesis was:

$$\log\left(\frac{\pi}{1-\pi}\right) = G_i \cdot \beta + \epsilon$$
$$\pi_i = \frac{e^{G_i \cdot \beta + \epsilon}}{1 + e^{G_i \cdot \beta + \epsilon}}$$
$$P_i \sim \text{Bernoulli}(\pi_i)$$

here π_i is the probability that individual i has the trait of interest (based on genetic and environmental factors) and P_i denotes the binary indicator of whether or not they have the trait. The error terms include both non-genetic (environmental) effects and random noise, and, without loss of generality, are assumed to have a mean of zero. Three phenotype matrices were generated for this simulation: the $n_g \times k$ phenotype matrix for GWAS (P_g), and the $n_v \times k$ genotype matrix for the validation set (P_v). The number of traits in this simulation is set to $k = 4$.

These trait columns in P are correlated due to two factors: their error terms are

correlated (a non-genetic mechanism), and there are shared causal SNPs between traits (a genetic mechanism). In this thesis we will use the term *causal SNP* to refer to variants which are directly related to the trait of interest; that is, the true β corresponding to the SNP is non-zero. The correlation matrix of the errors is:

$$C_{\text{error}} = \begin{bmatrix} 1 & 0.125 & 0.125 & 0.125 \\ 0.125 & 1 & 0.125 & 0.125 \\ 0.125 & 0.125 & 1 & 0.125 \\ 0.125 & 0.125 & 0.125 & 1 \end{bmatrix}$$

and $\epsilon \sim MVN(0, \Sigma)$ where the variance-covariance matrix is:

$$\Sigma = \begin{bmatrix} 2 & 0.25 & 0.25 & 0.25 \\ 0.25 & 2 & 0.25 & 0.25 \\ 0.25 & 0.25 & 2 & 0.25 \\ 0.25 & 0.25 & 0.25 & 2 \end{bmatrix}$$

The true heritability vector and the true genetic correlation matrix are calculated prior to the data analysis, since these are true values (not estimates based on data), we did not use the methods discussed in previous papers for estimating heritability and the genetic correlation matrix, such as the method using LD scores described in Zaitlen and Kraft (2012). Instead, we used the model and true βs (effect sizes) to calculate them. The heritability formula for the j^{th} corresponding trait is now:

$$H_j^2 = \frac{\text{Var}(G)}{\text{Var}(P)} = \frac{\text{Var}(G)}{\text{Var}(G) + \text{Var}(\epsilon)} = \frac{\sum_{i=1}^{p} \beta_{i,j}^2 \cdot 2 \cdot p_i \cdot (1 - p_i)}{\sum_{i=1}^{p} \beta_{i,j}^2 \cdot 2 \cdot p_i \cdot (1 - p_i) + \text{Var}(\epsilon)} \tag{3.1.1}$$

We computed those values for both continuous and binary focal traits, since in the binary case we focus on the heritability on the liability scale, where liability-scale heritability refers to the heritability of an unobserved continuous liability that underlies the binary trait. The genetic correlation covariance between two traits is calculated using:

$$V_{l,m} = V_{m,l} = \sum_{i=1}^{p} \beta_{i,l} \cdot \beta_{i,m} \cdot 2 \cdot p_i \cdot (1 - p_i) \tag{3.1.2}$$

for l and $m = 1, ..., k$ and these covariances were then converted to correlations in the standard way.

In our studies, the four traits (for both continuous and binary case), have different truly associated SNPs. The first trait is truly associated with SNPs 1–10, the second is truly associated with SNPs 6–15, the third is truly associated with SNPs 8–17 and the last trait is truly associated with SNPs 9–18. For the j^{th} trait, the vector β_j contains the effect sizes of the SNPs that are truly correlated with this trait, with values set to zero for SNPs that are not correlated with it. The vector can be represented as:

$$\beta_j = \begin{bmatrix} 0.1 & 0.2 & 0.3 & 0.4 & 0.5 & 0.6 & 0.7 & 0.8 & 0.9 & 1 \end{bmatrix}^T$$

These parameter values result in true heritability for both the continuous and binary

cases of:

$$H^2 \approx \begin{bmatrix} 0.4 & 0.4 & 0.4 & 0.4 \end{bmatrix}$$

which are close to real-life scenarios. The true genetic correlation matrix is:

$$C_{\text{genetic}} = \begin{bmatrix} 1 & 0.37 & 0.15 & 0.08 \\ 0.37 & 1 & 0.63 & 0.48 \\ 0.15 & 0.63 & 1 & 0.82 \\ 0.08 & 0.48 & 0.82 & 1 \end{bmatrix}$$

3.1.1 Simulation framework

For the continuous traits, we fit a linear regression model for the GWAS with one SNP at a time as the covariate. For the binary traits, we will fit a logistic regression model. The simulation is repeated $R = 200$ times. The sample size and number of SNPs used in this thesis are smaller than those in real-life scenarios, as larger parameters would require significantly more time to run. Figure 3.1 outlines the basic framework for the simulation in this thesis, consisting of four steps:

Step 1. The inputs are the $n_g \times p$ genotype matrix for GWAS (G_g) and the $n_g \times k$ phenotype matrix for GWAS (P_g). By applying genome-wide association studies (GWAS), a $p \times k$ matrix of effect size estimates ($\hat{\beta}$) is produced as output.

Step 2. The inputs are the $n_t \times p$ genotype matrix for the training set (G_t), the $n_t \times k$ phenotype matrix for the training set (P_t), the $p \times k$ effect size estimates ($\hat{\beta}$) matrix from Step 1, a heritability vector (H^2), and a genetic correlation matrix (C_{genetic}). Additional inputs depend on the specific method applied; in this thesis, these include the various methods discussed in Chapter 2. The output is a $k \times 1$ vector of weight

estimates (w), indicating the weight for each column of polygenic risk scores (PRS).

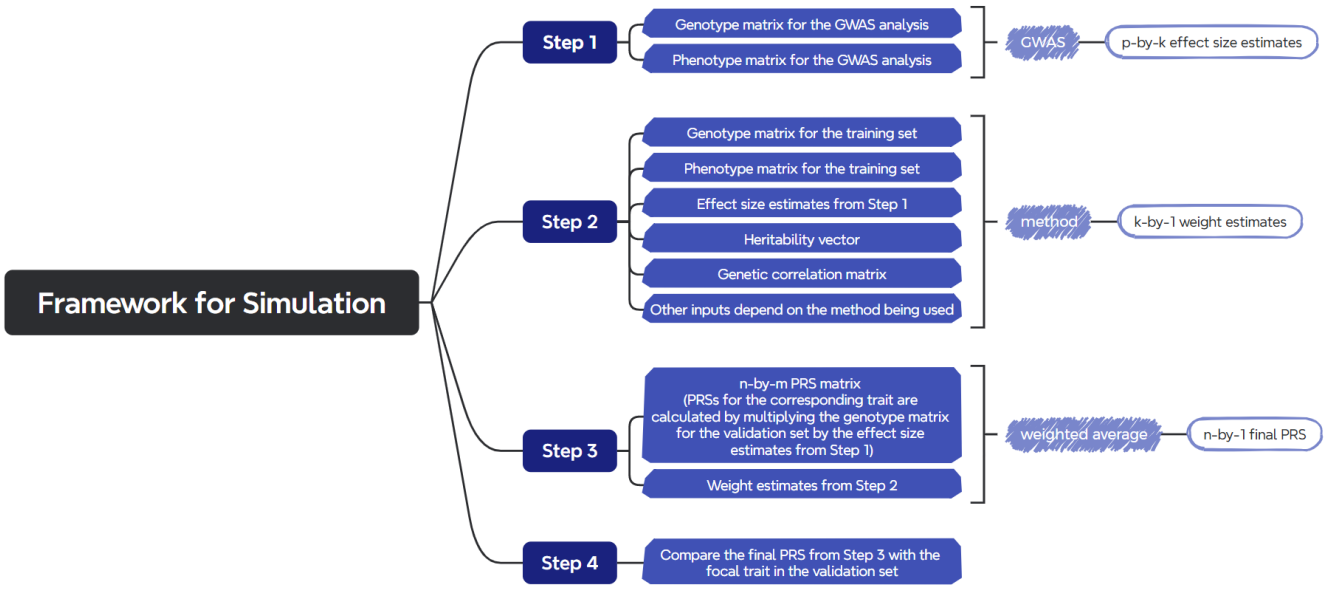

Figure 3.1: Framework for simulation in this thesis with steps

Step 3. In this step, we calculate the weighted average PRS for the validation set. The inputs are the $n_v \times p$ genotype matrix for the validation set (G_v), the $p \times k$ effect size estimates ($\hat{\beta}$) matrix from Step 1, and the $k \times 1$ weight estimates (w) vector from Step 2. The $n_v \times k$ PRS matrix is calculated by multiplying the genotype matrix (G_v) with the effect size estimates ($\hat{\beta}$) matrix, producing a score for each individual for each trait. The final $n_v \times 1$ polygenic risk score (PRS) is obtained by computing the weighted average of the trait-specific PRS values, using the weight estimates vector (w). This weighted average combines the individual PRS values across traits, where the weights represent the relative importance of each trait's contribution to the overall PRS.

Step 4. The final PRS from Step 3 is compared with the focal trait in the validation

set to assess the predictive accuracy.

In our simulation we compare the three methods described in Chapter 2 and also the uni-trait method which only used the PRS for the focal trait.

3.2 Continuous focal trait

In this thesis, the correlation between the target trait and the PRS generated from each method are examined to assess the accuracy of the methods (note that the correlation may not be very high, since PRS only uses genetic information, it can only predict the genetic component of risk for the trait).

Figure 3.2 shows the linear relationship between the PRS generated by each method and the corresponding trait for a single simulation run. We can see that the scatter plots exhibit a linear trend, indicating that the correlation coefficient is meaningful.

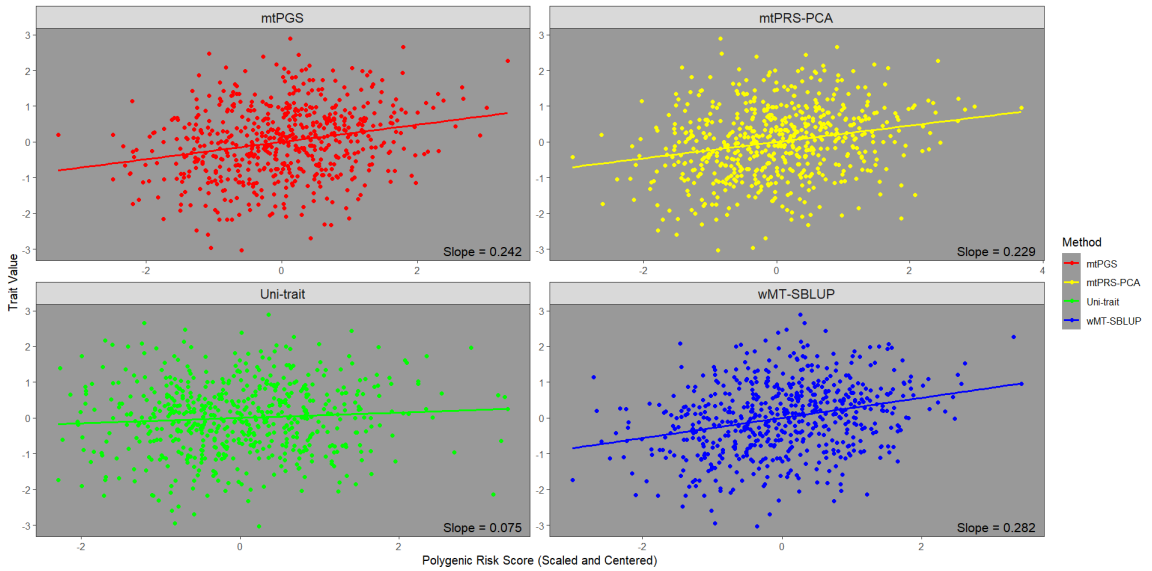

Figure 3.2: Plot for PRS and the first trait (as the average performance across all methods is the lowest for this trait)

Figure 3.3 presents boxplots for each method using the original parameters, these boxplots allow for a comparison of the correlation coefficients across methods for each trait. From Figure 3.3, we observe that the performance of the multi-trait method **wMT-SBLUP** method (show in blue) consistently outperforms the others, while **mtPGS** (show in red) performs worse. For the multi-trait methods **wMT-SBLUP** and **mtPRS-PCA**, the performance is comparatively lower for the first trait, while the other three traits show significantly higher performance. The **mtPGS** method performs best for the second trait, but still lags behind **wMT-SBLUP** and **mtPRS-PCA**. The uni-trait method shows consistent performance across all traits but delivers the worst results for all traits, except for the first one, when compared to all the multi-trait methods. One reason for this behavior may be that the first trait has the lowest average genetic correlation with the other traits (with the average correlation above 0.45 for the others, and the third trait having the highest at 0.53, while the first trait's is approximately 0.2), which could limit the extent to which information from the other traits improves prediction for the first trait. The performance of multi-trait methods, such as **mtPRS-PCA**, depends heavily on the genetic correlation matrix, which could explain their varied performance across traits. Furthermore, the consistently poor performance of **mtPGS** may be attributed to the fact that it is the only method that does not incorporate the PRS of the focal trait; its performance does not match the high performance observed in the paper, possibly because only ten SNPs are truly correlated with each trait, which may not provide enough predictive power for the model. Additionally, the low genetic correlation between the first trait and the others might constrain the advantages of multi-trait approaches, which could explain why uni-trait methods outperform some multi-trait methods for this specific

trait. Table 3.1 provides a summary of the average correlation between PRS and trait for each method across traits.

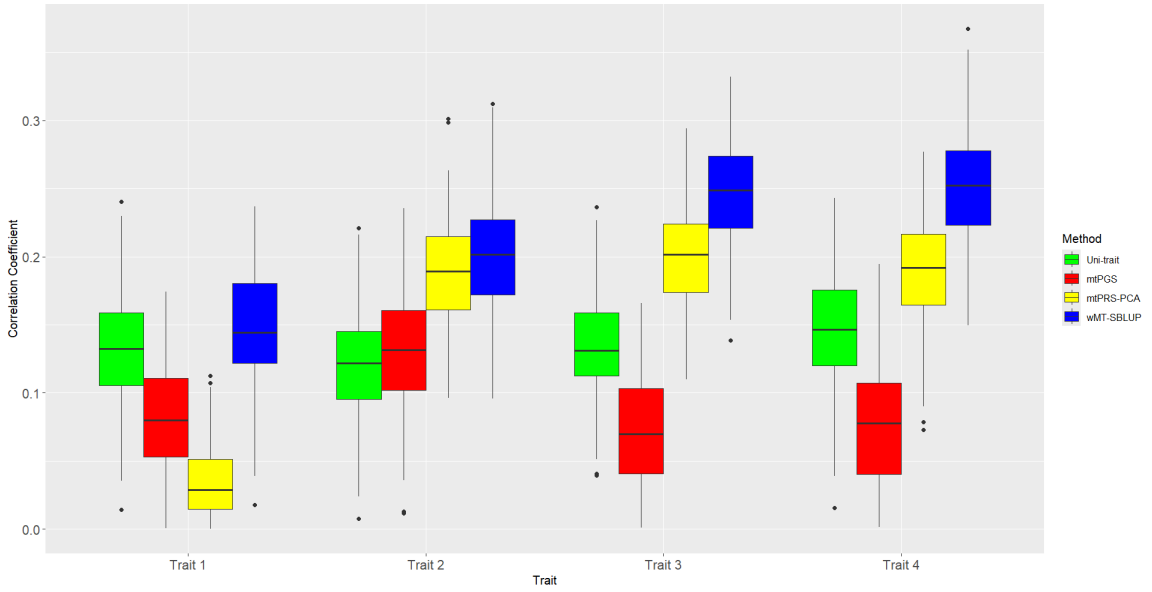

Figure 3.3: Comparison of correlation coefficients across methods using original parameters

Table 3.1: Mean correlation coefficients of original parameters

	Trait 1	**Trait 2**	**Trait 3**	**Trait 4**
Uni-trait	0.134	0.121	0.135	0.147
mtPGS	0.082	0.132	0.073	0.077
mtPRS-PCA	0.034	0.188	0.201	0.190
wMT-SBLUP	0.149	0.199	0.245	0.251

3.2.1 Effect of changing sample sizes

In this case, all parameters remained the same except for the sample sizes n_g and n_t.

Changing the GWAS sample size n_g

Figure 3.4 shows the performance of each method across traits as the GWAS sample size decreases from 1800 to 600. We observe that as the sample size n_g decreases, the correlation coefficient for the methods also decreases. The performance of each method across traits improves as the GWAS sample size increases from 1800 to 2700. The plot is qualitatively similar to Figure 3.4, with the major difference being the larger values on the Y-axis, the ordering of the performance remains the same, so we will not include this plot. We observe that as the sample size n_g increases, the correlation coefficient for the methods increases. The result aligns with expectations: the performance of all methods improves with larger GWAS sample sizes, as this increases statistical power and reduces sampling error. The relationship between the performance of the various methods, however, remains largely unchanged by the changing GWAS sample size.

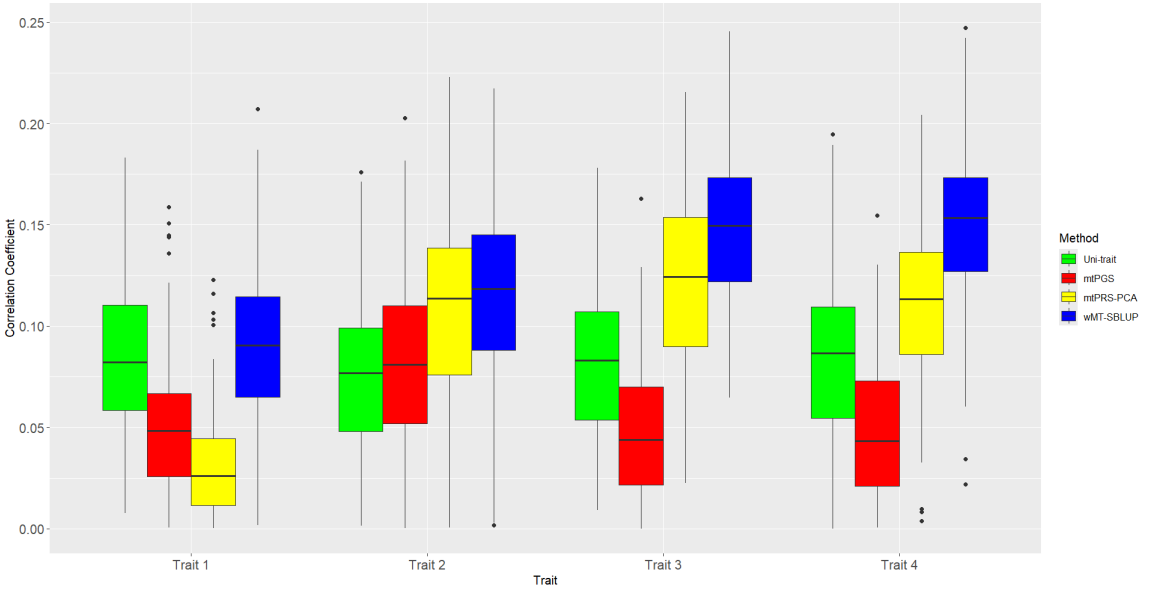

Figure 3.4: Comparison of correlation coefficients across methods when $n_g = 600$

Table 3.2: Mean correlation coefficients when n_g changes

	Method	**Trait 1**	**Trait 2**	**Trait 3**	**Trait 4**
$n_g = 600$	**Uni-trait**	0.085	0.074	0.081	0.086
	mtPGS	0.051	0.081	0.048	0.049
	mtPRS-PCA	0.031	0.109	0.123	0.113
	wMT-SBLUP	0.090	0.117	0.147	0.150
$n_g = 1800$	**Uni-trait**	0.134	0.121	0.135	0.147
	mtPGS	0.082	0.132	0.073	0.077
	mtPRS-PCA	0.034	0.188	0.201	0.190
	wMT-SBLUP	0.149	0.199	0.245	0.251
$n_g = 2700$	**Uni-trait**	0.166	0.145	0.169	0.179
	mtPGS	0.089	0.146	0.081	0.076
	mtPRS-PCA	0.038	0.223	0.237	0.221
	wMT-SBLUP	0.168	0.240	0.292	0.296

Changing the training sample size for the PRS n_t

Figure 3.5 shows the performance of each method across traits as the methods sample size decreases from 900 to 300; we observe that as the sample size n_t decreases, the performance gap between the uni-trait method and the multi-trait methods for each trait is smaller. As the sample size increases from 900 to 1350, the plot is very similar to Figure 3.5, with the major difference being the values on the Y-axis. The ordering of the methods in terms of performance stays the same.

Table 3.3: Mean correlation coefficients when n_t changes

	Method	Trait 1	Trait 2	Trait 3	Trait 4
$n_t = 300$	**Uni-trait**	0.138	0.127	0.136	0.138
	mtPGS	0.080	0.124	0.076	0.072
	mtPRS-PCA	0.059	0.186	0.195	0.172
	wMT-SBLUP	0.140	0.199	0.245	0.235
$n_t = 900$	**Uni-trait**	0.134	0.121	0.135	0.147
	mtPGS	0.082	0.132	0.073	0.077
	mtPRS-PCA	0.034	0.188	0.201	0.190
	wMT-SBLUP	0.149	0.199	0.245	0.251
$n_t = 1350$	**Uni-trait**	0.137	0.120	0.133	0.149
	mtPGS	0.072	0.129	0.066	0.054
	mtPRS-PCA	0.032	0.185	0.198	0.182
	wMT-SBLUP	0.141	0.197	0.243	0.247

3.2.2 Effect of the total number of SNPs p

In this case, all parameters remained the same as in the original simulation (including the sample size $n_g = 1800$ and $n_t = 900$, true effect size β, heritability and genetic correlation matrix) except for the number of SNPs p.

Figure 3.6 shows the performance of each method across traits as the number of SNPs decreases from 5000 to 3000. Compared to Figure 3.3, we observe that as the number of SNPs decreases, the correlation coefficient for the methods increases.

The plot showing the performance of each method across traits as the number of

SNPs increases from 5000 to 7000 is very similar to Figure 3.6. Compare to Figure 3.3, we observe that as the number of SNPs increases, the correlation coefficient for the methods decreases.

The results in this section may be due to the fact that, as the number of SNPs increases, we are adding SNPs that are not associated with any of the traits. This can decrease performance by introducing noise.

Table 3.4: Mean correlation coefficients when p changes

	Method	**Trait 1**	**Trait 2**	**Trait 3**	**Trait 4**
$p = 3000$	**Uni-trait**	0.169	0.151	0.173	0.184
	mtPGS	0.091	0.153	0.079	0.082
	mtPRS-PCA	0.036	0.233	0.248	0.224
	wMT-SBLUP	0.175	0.241	0.298	0.302
$p = 5000$	**Uni-trait**	0.134	0.121	0.135	0.147
	mtPGS	0.082	0.132	0.073	0.077
	mtPRS-PCA	0.034	0.188	0.201	0.190
	wMT-SBLUP	0.149	0.199	0.245	0.251
$p = 7000$	**Uni-trait**	0.123	0.105	0.114	0.128
	mtPGS	0.066	0.110	0.058	0.060
	mtPRS-PCA	0.033	0.163	0.170	0.156
	wMT-SBLUP	0.126	0.174	0.210	0.216

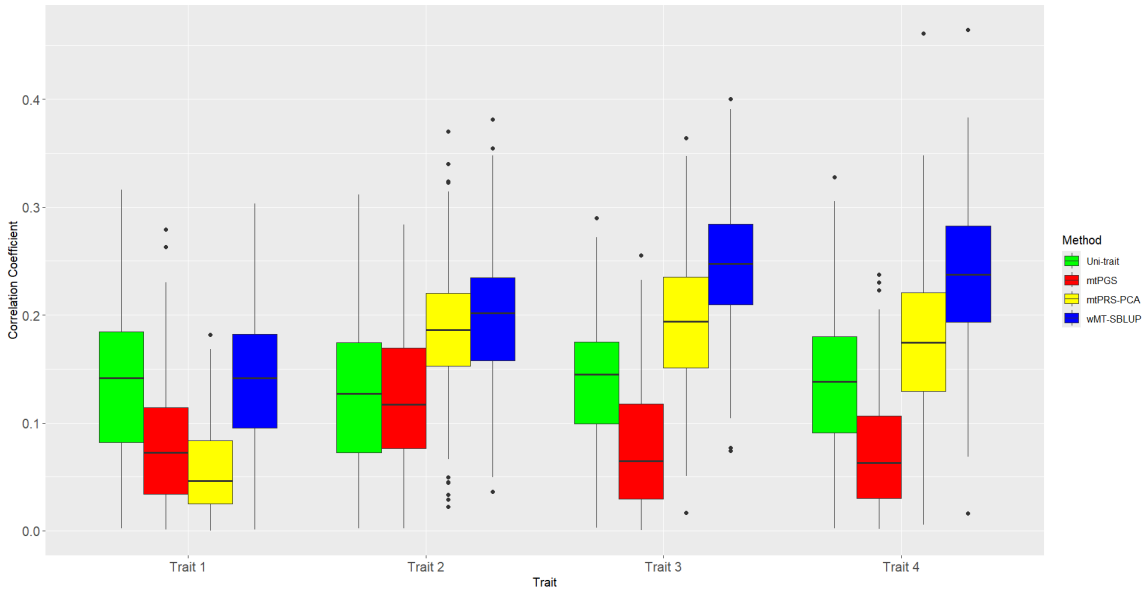

Figure 3.5: Comparison of correlation coefficients across methods when $n_t = 300$

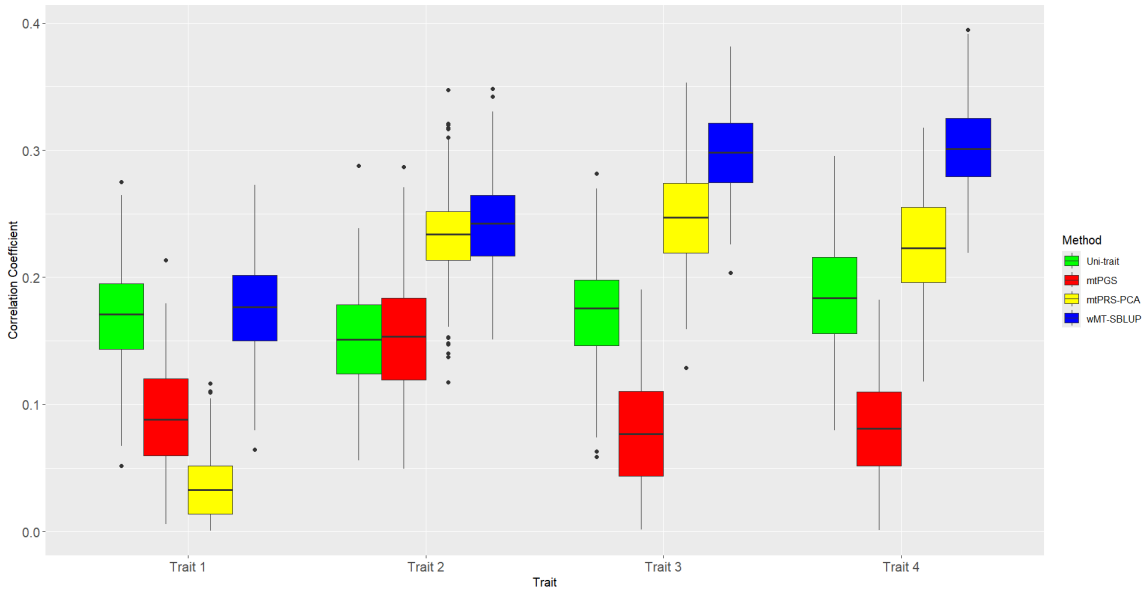

Figure 3.6: Comparison of correlation coefficients across methods when $p = 3000$

3.2.3 How changing the true genetic effect sizes β affects performance

In this case, all parameters remained the same. The change in true effect sizes will affect the heritability of each trait (while heritability across the traits remains approximately the same), but will not have much impact on the genetic correlation matrix.

Figure 3.7 shows the performance of each method across traits as the effect sizes decreases to:

$$\beta = \begin{bmatrix} 0.01 & 0.11 & 0.21 & 0.31 & 0.41 & 0.51 & 0.61 & 0.71 & 0.81 & 0.91 \end{bmatrix}^T$$

with heritability approximately 0.3 for each trait. The plot showing the performance of each method across traits as the effect sizes increases to

$$\beta = \begin{bmatrix} 0.3 & 0.4 & 0.5 & 0.6 & 0.7 & 0.8 & 0.9 & 1.0 & 1.1 & 1.2 \end{bmatrix}^T$$

with heritability approximately 0.5 for each trait is very similar to Figure 3.7. We observe that as the true effect sizes increase, the correlation coefficient for all methods increases, except for the multi-trait method **mtPRS-PCA**.

Additionally, compared to the original case, the distribution of the **mtPRS-PCA** method shows a significant shift, with the ranking of the traits reversed. In the original case, the second trait has the highest values, followed by the third and fourth traits. However, in this case, the second trait now has the lowest values, with the third and fourth traits showing progressively higher values.

Table 3.5: Mean correlation coefficients when β changes

	Method	**Trait 1**	**Trait 2**	**Trait 3**	**Trait 4**
$H^2 \approx 0.3$	**Uni-trait**	0.114	0.100	0.115	0.126
	mtPGS	0.062	0.106	0.059	0.065
	mtPRS-PCA	0.097	0.033	0.128	0.160
	wMT-SBLUP	0.124	0.166	0.212	0.219
$H^2 \approx 0.4$	**Uni-trait**	0.134	0.121	0.135	0.147
	mtPGS	0.082	0.132	0.073	0.077
	mtPRS-PCA	0.034	0.188	0.201	0.190
	wMT-SBLUP	0.149	0.199	0.245	0.251
$H^2 \approx 0.5$	**Uni-trait**	0.175	0.163	0.176	0.188
	mtPGS	0.128	0.181	0.102	0.097
	mtPRS-PCA	0.160	0.039	0.159	0.202
	wMT-SBLUP	0.197	0.262	0.308	0.310

3.2.4 The effect of variable heritability H^2 across traits

This case considers different heritabilities for each trait. Changing the variances for the error terms in the trait generation model can alter the heritability of the corresponding trait.

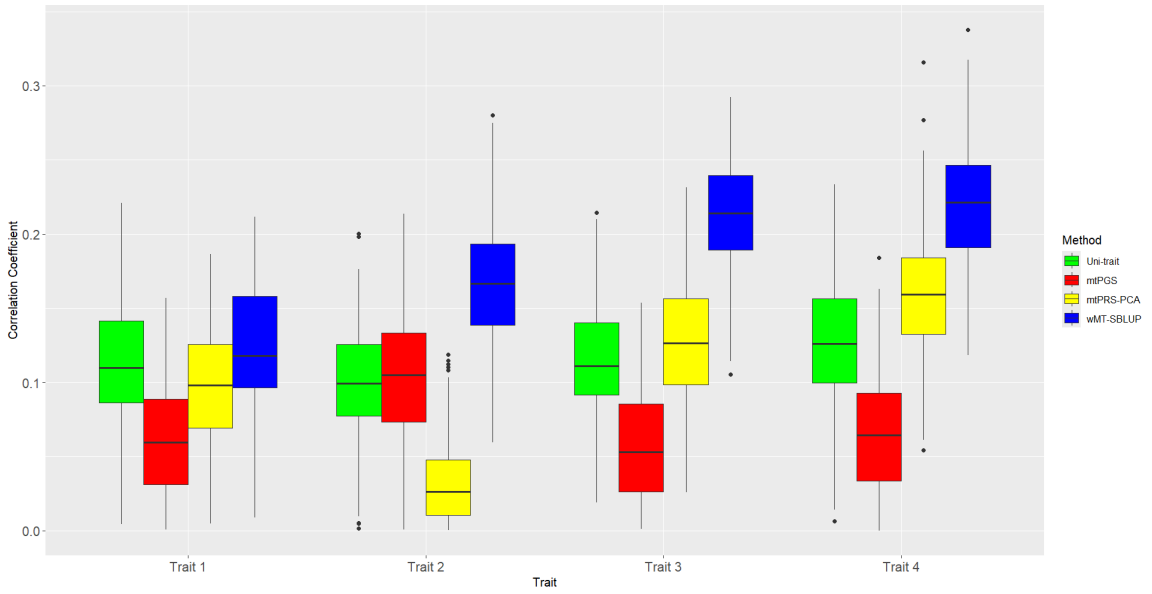

Figure 3.7: Comparison of correlation coefficients across methods when heritability approximately 0.3

Figure 3.8 shows the performance of each method across traits as the variance-covariance matrix becomes:

$$\Sigma = \begin{bmatrix} 2 & 0.25 & 0.25 & 0.25 \\ 0.25 & 3 & 0.25 & 0.25 \\ 0.25 & 0.25 & 5 & 0.25 \\ 0.25 & 0.25 & 0.5 & 7 \end{bmatrix} \qquad H^2 \approx \begin{bmatrix} 0.4 & 0.3 & 0.2 & 0.1 \end{bmatrix}$$

We observe that heritability significantly impacts the performance of the methods. In this case, the performance of the three multi-trait methods for the last trait is now closer to that of the first trait. Additionally, the uni-trait method no longer provides consistent performance across traits. Furthermore, among the multi-trait methods, **wMT-SBLUP** performs the worst for the second trait compared to all

other multi-trait methods for that trait.

Figure 3.9 shows the performance of each method across traits as the variance-covariance matrix becomes:

$$\Sigma = \begin{bmatrix} 7 & 0.25 & 0.25 & 0.25 \\ 0.25 & 5 & 0.25 & 0.25 \\ 0.25 & 0.25 & 3 & 0.25 \\ 0.25 & 0.25 & 0.5 & 2 \end{bmatrix} \qquad H^2 \approx \begin{bmatrix} 0.1 & 0.2 & 0.3 & 0.4 \end{bmatrix}$$

In this case, we observe the multi-trait method **mtPGS** demonstrates consistent performance across traits, while the other methods show an increasing trend in performance across traits.

The reason why these two plots show opposite trends is that we are flipping reversing the trend of the heritability. Higher heritability results in better performance.

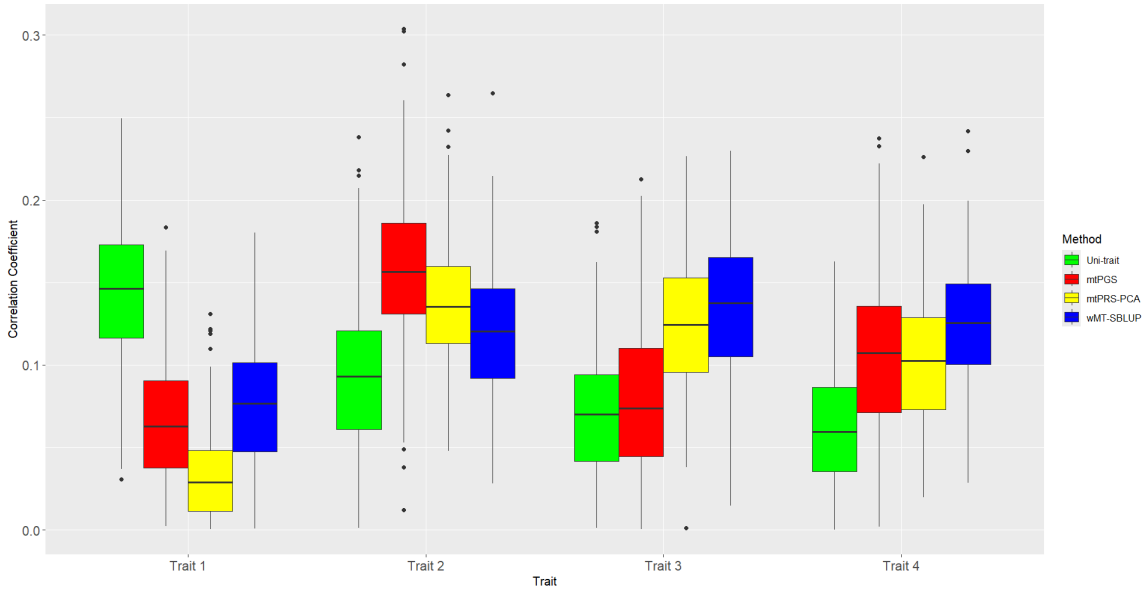

Figure 3.8: Comparison of correlation coefficients across methods when H^2 values are listed from high to low

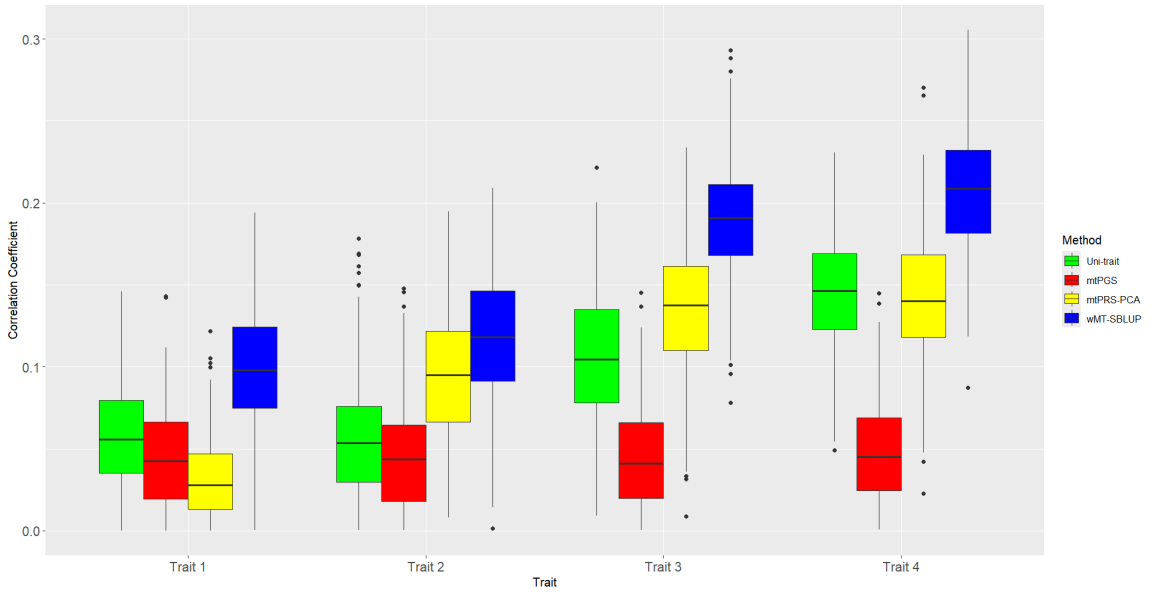

Figure 3.9: Comparison of correlation coefficients across methods when H^2 values are listed from low to high

3.2.5 The effect of altering the genetic correlations

In this case, all parameters remained the same except for the genetic correlation matrix. Adjusting the number of overlapping SNPs modifies the genetic correlation matrix, this adjustment does not affect other parameters, such as heritability, which remains unchanged.

Figure 3.10 shows the performance of each method across traits as genetic correlation matrix becomes:

$$C_{\text{genetic}} = \begin{bmatrix} 1 & 0.70 & 0.58 & 0.82 \\ 0.70 & 1 & 0.88 & 0.87 \\ 0.58 & 0.88 & 1 & 0.75 \\ 0.82 & 0.87 & 0.75 & 1 \end{bmatrix}$$

In this case, we observe that the multi-trait methods **mtPGS** exhibit consistent

performance across traits, similar to the uniform performance demonstrated by the uni-trait method, this may be because the genetic correlations across traits are almost the same. It is notable that the method **mtPGS** performs much better in this case compared to all the previous cases, this improvement is likely due to the increased genetic correlations, which enhance the influence of the other traits on the focal trait, thereby benefiting this multi-trait method (recall that **mtPGS** is the only method that does not use the focal PRS). Among the multi-trait methods, **wMT-SBLUP** consistently outperforms the others, while **mtPRS-PCA** exhibits comparatively lower performance.

Table 3.6: Mean correlation coefficients when H^2 different

	Method	**Trait 1**	**Trait 2**	**Trait 3**	**Trait 4**
H^2 high to low	**Uni-trait**	0.145	0.092	0.070	0.062
	mtPGS	0.067	0.156	0.079	0.103
	mtPRS-PCA	0.034	0.138	0.123	0.102
	wMT-SBLUP	0.075	0.120	0.136	0.124
$H^2 \approx 0.4$	**Uni-trait**	0.134	0.121	0.135	0.147
	mtPGS	0.082	0.132	0.073	0.077
	mtPRS-PCA	0.034	0.188	0.201	0.190
	wMT-SBLUP	0.149	0.199	0.245	0.251
H^2 low to high	**Uni-trait**	0.060	0.059	0.107	0.146
	mtPGS	0.046	0.047	0.046	0.049
	mtPRS-PCA	0.033	0.097	0.134	0.141
	wMT-SBLUP	0.097	0.118	0.191	0.207

Figure 3.11 shows the performance of each method across traits as genetic correlation matrix becomes:

$$C_{\text{genetic}} = \begin{bmatrix} 1 & 0.24 & 0.05 & 0.01 \\ 0.24 & 1 & 0.27 & 0.15 \\ 0.05 & 0.27 & 1 & 0.07 \\ 0.01 & 0.15 & 0.07 & 1 \end{bmatrix}$$

In this case, we observe that as the average genetic correlation for each trait decreases to approximately 0.1–0.2, the multi-trait methods lose their advantage except **wMT-SBLUP** , showing diminished performance compared to higher correlation scenarios.

Table 3.7: Mean correlation coefficients when average genetic correlation changes

	Method	**Trait 1**	**Trait 2**	**Trait 3**	**Trait 4**
For each trait about 0.7–0.8	**Uni-trait**	0.134	0.143	0.136	0.151
	mtPGS	0.213	0.225	0.224	0.230
	mtPRS-PCA	0.195	0.138	0.084	0.196
	wMT-SBLUP	0.243	0.267	0.243	0.275
About 0.4–0.5	**Uni-trait**	0.134	0.121	0.135	0.147
	mtPGS	0.082	0.132	0.073	0.077
	mtPRS-PCA	0.034	0.188	0.201	0.190
	wMT-SBLUP	0.149	0.199	0.245	0.251
For each trait about 0.1–0.2	**Uni-trait**	0.134	0.130	0.130	0.159
	mtPGS	0.054	0.138	0.099	0.111
	mtPRS-PCA	0.059	0.036	0.118	0.205
	wMT-SBLUP	0.166	0.176	0.166	0.261

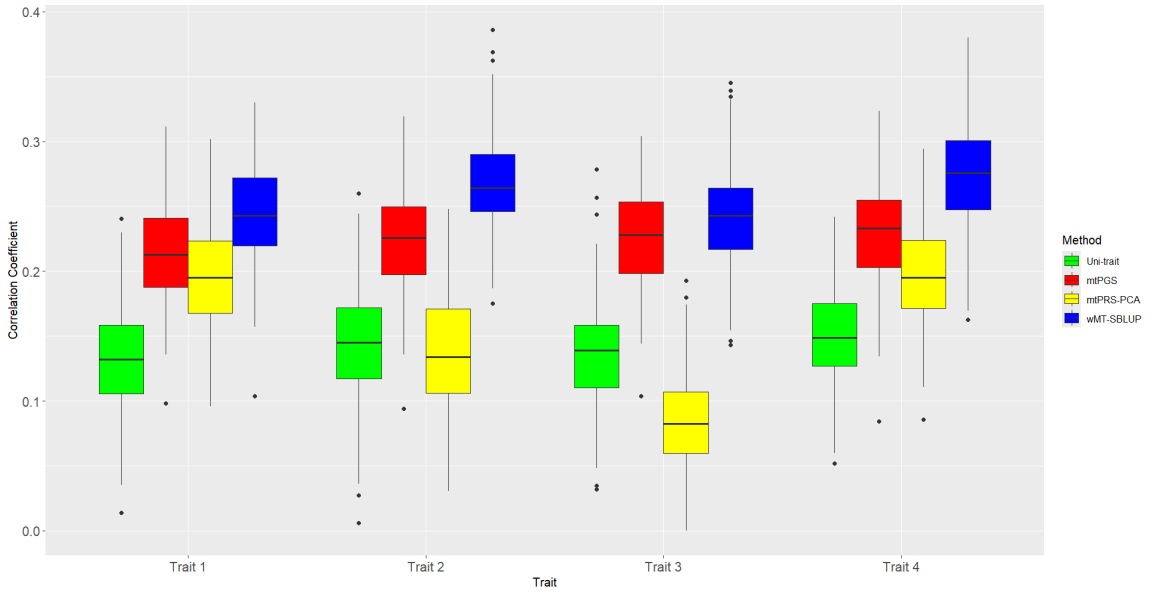

Figure 3.10: Comparison of correlation coefficients across methods when the off-diagonal elements of the genetic correlation matrix change (with an average genetic correlation for each trait about 0.7–0.8)

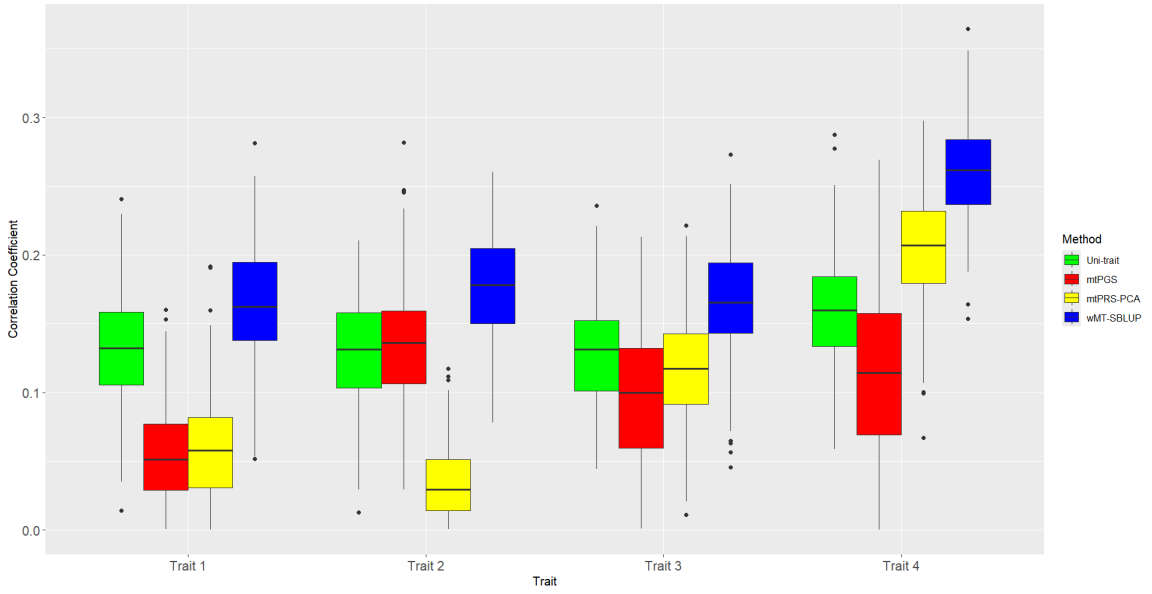

Figure 3.11: Comparison of correlation coefficients across methods when the off-diagonal elements of the genetic correlation matrix change (with an average genetic correlation for each trait about 0.1–0.2)

3.3 Binary focal trait

When looking at binary focal trait, we use the area under the receiver operating characteristic (ROC) curve to compare the polygenic risk score (PRS) with the observed trait values.

In this scenario we had two binary traits and two continuous traits. We will focus on evaluating and plotting the performance specifically for the binary focal traits. From Figure 3.12, we observe that the multi-trait method, **wMT-SBLUP**, consistently outperforms the others, while **mtPRS-PCA** exhibits comparatively lower performance. Additionally, compared to the continuous scenario, the multi-trait method **mtPGS** performs much better in the binary scenario. In contrast, the uni-trait method demonstrates consistent performance across all traits but yields the poorest results in every case. Table 3.8 provides a summary of the average performance for each method across traits.

Table 3.8: Mean AUC values of original parameters

	Trait 1	**Trait 3**
Uni-trait	0.583	0.582
mtPGS	0.602	0.648
mtPRS-PCA	0.599	0.642
wMT-SBLUP	0.614	0.655

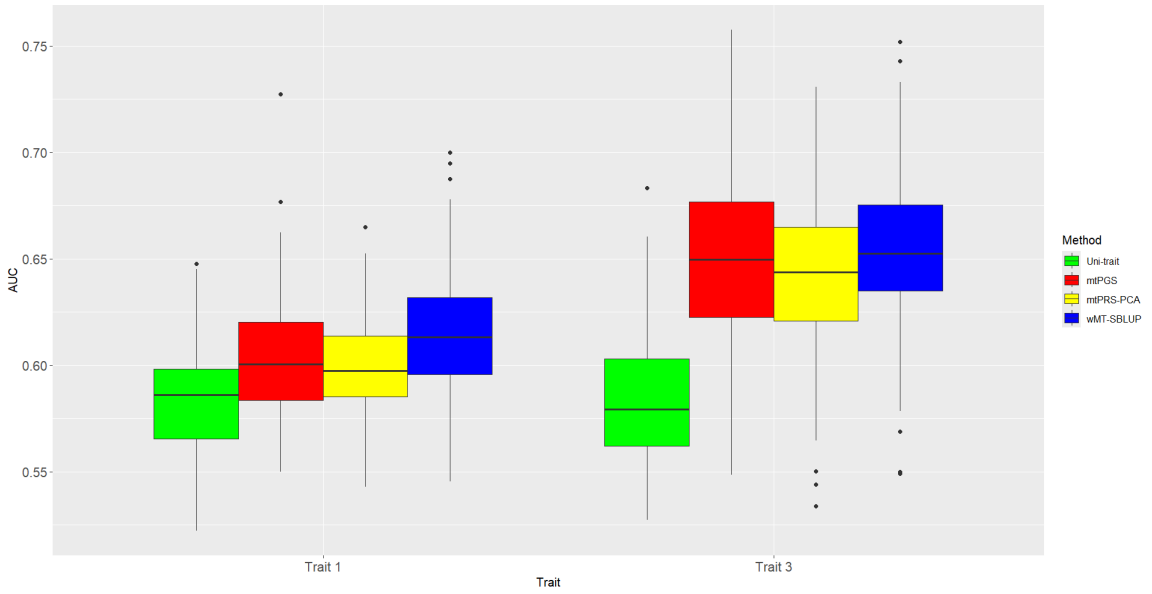

Figure 3.12: Comparison of AUC values across methods using the original parameters

3.3.1 The effect of variable heritability H^2 across traits

Figure 3.13 shows the performance of each method across traits as the variance-covariance matrix becomes:

$$\Sigma = \begin{bmatrix} 2 & 0.25 & 0.25 & 0.25 \\ 0.25 & 3 & 0.25 & 0.25 \\ 0.25 & 0.25 & 5 & 0.25 \\ 0.25 & 0.25 & 0.5 & 7 \end{bmatrix} \qquad H^2 \approx \begin{bmatrix} 0.4 & 0.3 & 0.2 & 0.1 \end{bmatrix}$$

In this case, among the multi-trait methods, **wMT-SBLUP** still performs the best across all traits compared to the other multi-trait methods, while **mtPGS** performs better than **mtPRS-PCA** for the first trait.

Figure 3.14 shows the performance of each method across traits as the variance-covariance matrix becomes:

$$\Sigma = \begin{bmatrix} 7 & 0.25 & 0.25 & 0.25 \\ 0.25 & 5 & 0.25 & 0.25 \\ 0.25 & 0.25 & 3 & 0.25 \\ 0.25 & 0.25 & 0.5 & 2 \end{bmatrix} \qquad H^2 \approx \begin{bmatrix} 0.1 & 0.2 & 0.3 & 0.4 \end{bmatrix}$$

In this case, unlike the continuous scenario, the performance of the multi-trait method **mtPGS** for the third trait is very close to that of **wMT-SBLUP**.

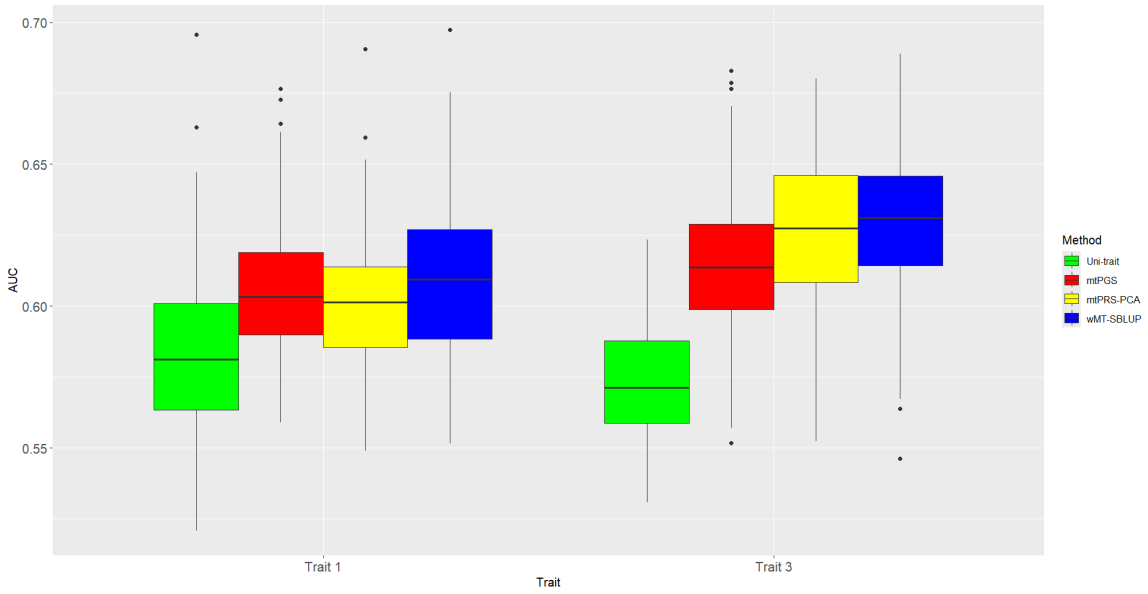

Figure 3.13: Comparison of AUC values across methods when H^2 values are listed from high to low

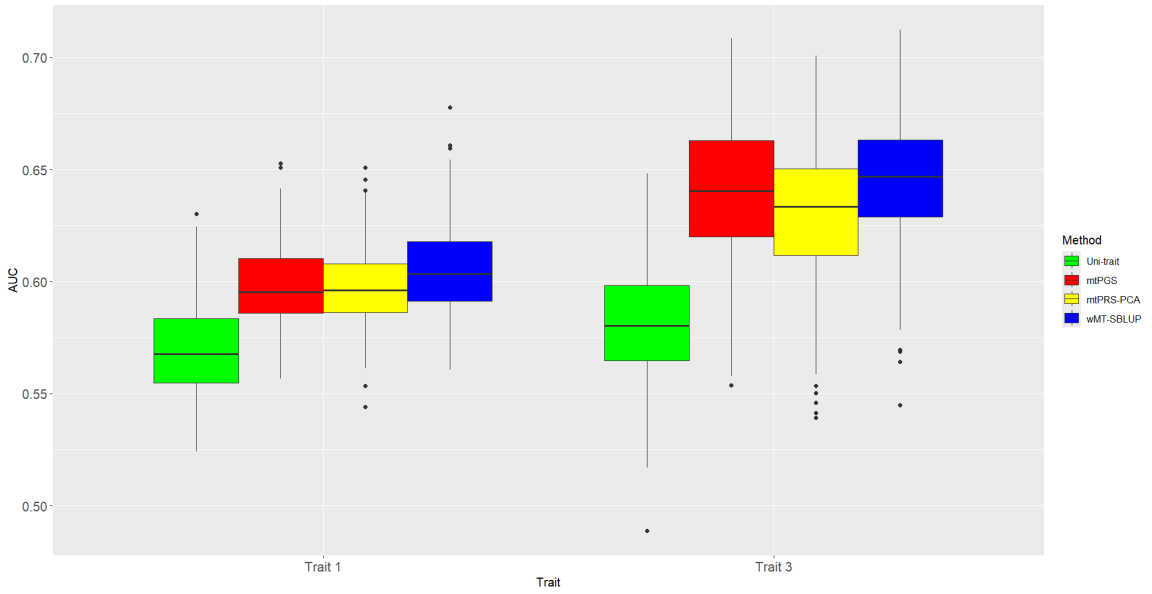

Figure 3.14: Comparison of AUC values across methods when H^2 values are listed from low to high

Table 3.9: Mean AUC values when H^2 different

	H^2 high to low		Original $H^2 \approx 0.4$		H^2 low to high	
	Trait 1	**Trait 3**	**Trait 1**	**Trait 3**	**Trait 1**	**Trait 3**
Uni-trait	0.583	0.573	0.583	0.582	0.570	0.581
mtPGS	0.605	0.615	0.602	0.648	0.598	0.639
mtPRS-PCA	0.601	0.627	0.599	0.642	0.597	0.631
wMT-SBLUP	0.609	0.630	0.614	0.655	0.605	0.644

3.3.2 Effect of other parameters

In the binary scenario, we altered all the parameters as we did in the continuous scenario. We do not provide results for some of the simulations, such as the effects

of changes in n_g (GWAS sample size), n_t (training set sample size), p (number of SNPs), β (true effect sizes), and the genetic correlation matrix, because their results were consistent with those observed in the continuous scenario (refer to Section 3.2). Specifically, these parameters did not exhibit substantially different effects on the model's performance when switching from a continuous to a binary outcome, reinforcing the idea that the relationships and trends are similar across different trait types.

Chapter 4

Two Novel Methods

4.1 Methods

After simulating the multi-trait methods from the literature, we began considering whether other methods might yield better performance. We developed two new methods to generate weighted average polygenic risk scores (PRS). To examine these methods we used the same simulation set-up as in Chapter 3. In each figure of this chapter, we include two novel methods along with the multi-trait method **wMT-SBLUP**, as this method performs the best in the majority of cases in Chapter 3. Both novel methods assume that we have access to the values for the trait of interest in the training dataset (same as method **mtPGS**) , which may limit their practical applicability.

4.1.1 A method based on multiple linear regression

The method **mt-lm** is a multiple linear regression for a continuous focal trait and a logistic regression for a binary focal trait, where the response variable is the focal trait

and the covariates are the corresponding PRS for each trait. We were surprised to find no reference paper specifically for this simple method, at least based on our research. The method calculates the weighted average PRS using the following formula:

$$\text{mt-lm} = \sum_{k=1}^{K} \gamma_k \cdot \text{PRS}_k \tag{4.1.1}$$

where weights $\gamma = (\gamma_1, ..., \gamma_k)$ are the estimates from multiple linear regression model.

4.1.2 Multi-trait cross-validation bagging method

This method (which will refer to as **mt-CVb**) combines cross-validation and bagging techniques in a hybrid approach to improve model performance. The method calculates the weighted average PRS using the following formula:

$$\text{mt-CVb} = \sum_{k=1}^{K} w_k \cdot \text{PRS}_k \tag{4.1.2}$$

where weights $w = (w_1, ..., w_k)$ are obtained using cross-validation bagging variable importance score. A higher score indicates that the specific covariate has a greater effect on the model, as it is weighted more heavily. The sum of the importance scores across all covariates in a single model is standardized to 1 (100%), this standardization ensures comparability and emphasizes the relative contribution of each covariate (PRS for the corresponding trait) to the model's predictions.

Bagging is an ensemble method that aims to improve model accuracy and reduce variance by training multiple models on different subsets of the data. These subsets are created using bootstrap sampling, where data points are randomly selected with replacement. Each model is trained on a bootstrap sample, and the final prediction is

made by combining the outputs, usually through averaging or voting. Cross-validation bagging further enhances this by using cross-validation during training.

As we did with **mtPGS** in Chapter 3, we used 10-fold cross-validation in our implementation. The cross-validation bagging model is then fitted with multiple linear regression, where the target trait is the response variable and all the PRS are the predictors. In the other words, bootstrap samples are generated from the original dataset, and a model is trained on each bootstrap sample to calculate variable importance scores or predict the response variable for each fold. After repeating the bagging and cross-validation steps, the variable importance scores are aggregated to provide a comprehensive measure of each covariate's contribution to the model.

4.2 Results for continuous focal trait

We used the same simulated data from Chapter 3 to apply these two novel methods. From Figure 4.1, we observe that the performance of the two new multi-trait methods, **mt-lm** and **mt-CVb**, are very similar. However, **mt-lm** exhibits higher average performance compared to **mt-CVb** for the third and fourth traits. The variability for the **wMT-SBLUP** method is lower, as indicated by its less dispersed box plot.

Table 4.1 summarizes the average performance of each method across traits (including all methods from Chapter 3 and the new methods), rounded to three decimal places. While the multi-trait method **wMT-SBLUP** consistently outperforms both of these new methods, the new methods remain competitive.

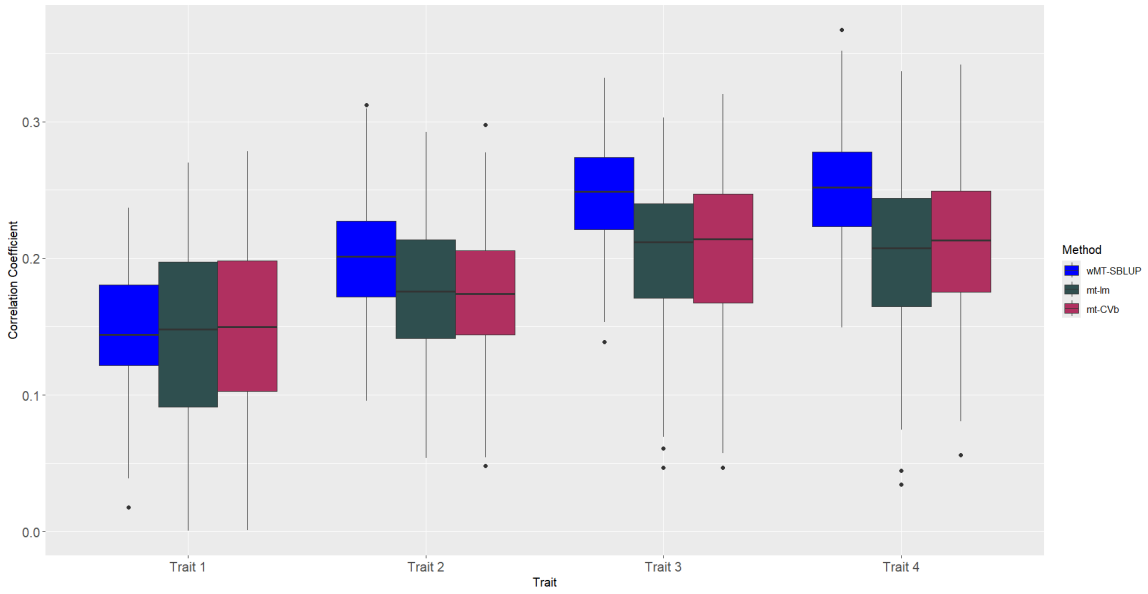

Figure 4.1: Comparison of correlation coefficients across methods using original parameters for new methods

Table 4.1: Mean correlation coefficients of original parameters for all methods

	Trait 1	**Trait 2**	**Trait 3**	**Trait 4**
Uni-trait	0.134	0.121	0.135	0.147
mtPGS	0.082	0.132	0.073	0.077
mtPRS-PCA	0.034	0.188	0.201	0.190
wMT-SBLUP	0.149	0.199	0.245	0.251
mt-lm	0.143	0.177	0.206	0.206
mt-CVb	0.148	0.175	0.205	0.210

4.2.1 Effect of changing sample sizes

In this section, we varied n_g and n_t in the same way as in Section 3.2.1 and Section 3.3.1.

Changing the GWAS sample size n_g

Table 4.2 shows the performance of the two new multi-trait methods across traits as the GWAS sample size decreases from 1800 to 600; we observe that as the sample size n_g decreases, Figure 4.2 shows that the correlation coefficient for the methods also decreases and method **mt-lm** always has better performance than method **mt-CVb** in this case. Table 4.2 shows the performance of the two new multi-trait methods across traits as the GWAS sample size increases from 1800 to 2700; we observe that as the sample size n_g increases, the correlation coefficient for the methods increases and Figure 4.3 shows that **mt-CVb** exhibits higher average performance compared to **mt-lm** for the second and third traits.

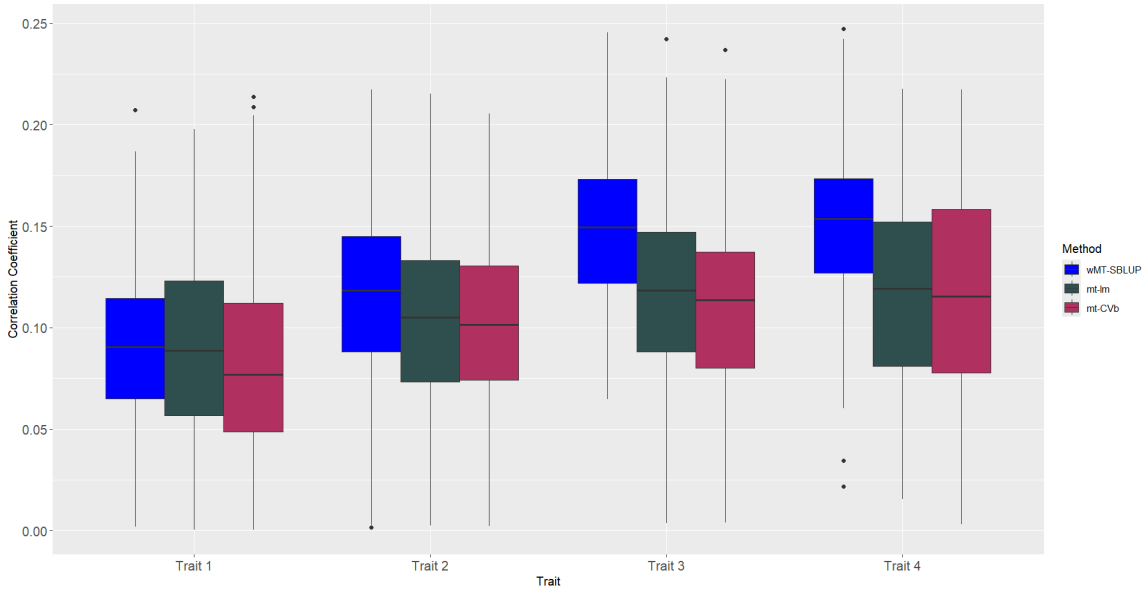

Figure 4.2: Comparison of correlation coefficients across methods when $n_g = 600$ for new methods

Table 4.2: Mean correlation coefficients when n_g changes for all methods

	Method	**Trait 1**	**Trait 2**	**Trait 3**	**Trait 4**
$n_g = 600$	**Uni-trait**	0.085	0.074	0.081	0.086
	mtPGS	0.051	0.081	0.048	0.049
	mtPRS-PCA	0.031	0.109	0.123	0.113
	wMT-SBLUP	0.090	0.117	0.147	0.150
	mt-lm	0.087	0.103	0.119	0.117
	mt-CVb	0.080	0.102	0.113	0.116
$n_g = 1800$	**Uni-trait**	0.134	0.121	0.135	0.147
	mtPGS	0.082	0.132	0.073	0.077
	mtPRS-PCA	0.034	0.188	0.201	0.190
	wMT-SBLUP	0.149	0.199	0.245	0.251
	mt-lm	0.143	0.177	0.206	0.206
	mt-CVb	0.148	0.175	0.205	0.210
$n_g = 2700$	**Uni-trait**	0.166	0.145	0.169	0.179
	mtPGS	0.089	0.146	0.081	0.076
	mtPRS-PCA	0.038	0.223	0.237	0.221
	wMT-SBLUP	0.168	0.240	0.292	0.296
	mt-lm	0.183	0.208	0.244	0.263
	mt-CVb	0.180	0.215	0.256	0.253

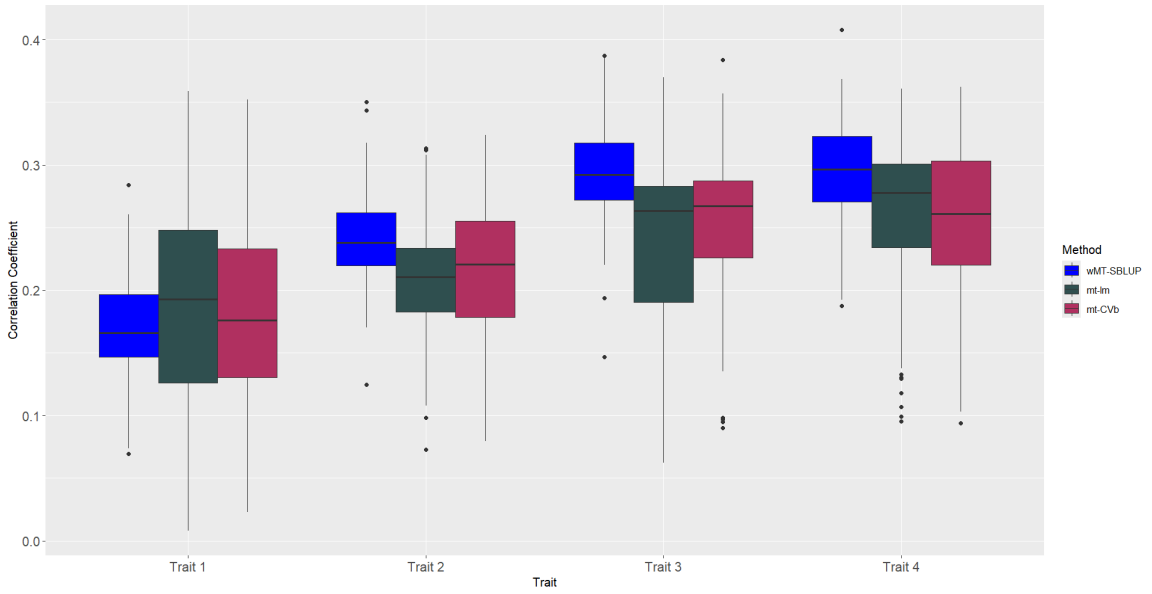

Figure 4.3: Comparison of correlation coefficients across methods when $n_g = 2700$ for new methods

Changing the training sample size for the PRS n_t

Table 4.3 shows the performance of the two new multi-trait methods across traits as the methods sample size decreases from 900 to 300; we observe that as the sample size n_t decreases, Figure 4.4 shows the performance gap between the uni-trait method and the multi-trait methods for each trait is smaller and method **mt-CVb** exhibits higher average performance compare to **mt-lm** for the first and second traits. Table 4.3 shows the performance of the two new multi-trait methods across traits as the methods sample size increases from 900 to 1350; we observe that as the sample size n_t increases, method **mt-CVb** exhibits higher average performance compare to **mt-lm** for all the traits except the third one, Figure 4.5 shows that the performance of methods **mt-lm** and **mt-CVb** are very similar for the first two traits.

Table 4.3: Mean correlation coefficients when n_t changes for all methods

	Method	**Trait 1**	**Trait 2**	**Trait 3**	**Trait 4**
$n_t = 300$	**Uni-trait**	0.138	0.127	0.136	0.138
	mtPGS	0.080	0.124	0.076	0.072
	mtPRS-PCA	0.059	0.186	0.195	0.172
	wMT-SBLUP	0.140	0.199	0.245	0.235
	mt-lm	0.131	0.183	0.208	0.191
	mt-CVb	0.135	0.186	0.204	0.189
$n_t = 900$	**Uni-trait**	0.134	0.121	0.135	0.147
	mtPGS	0.082	0.132	0.073	0.077
	mtPRS-PCA	0.034	0.188	0.201	0.190
	wMT-SBLUP	0.149	0.199	0.245	0.251
	mt-lm	0.143	0.177	0.206	0.206
	mt-CVb	0.148	0.175	0.205	0.210
$n_t = 1350$	**Uni-trait**	0.137	0.120	0.133	0.149
	mtPGS	0.072	0.129	0.066	0.054
	mtPRS-PCA	0.032	0.185	0.198	0.182
	wMT-SBLUP	0.141	0.197	0.243	0.247
	mt-lm	0.149	0.179	0.212	0.218
	mt-CVb	0.152	0.180	0.210	0.220

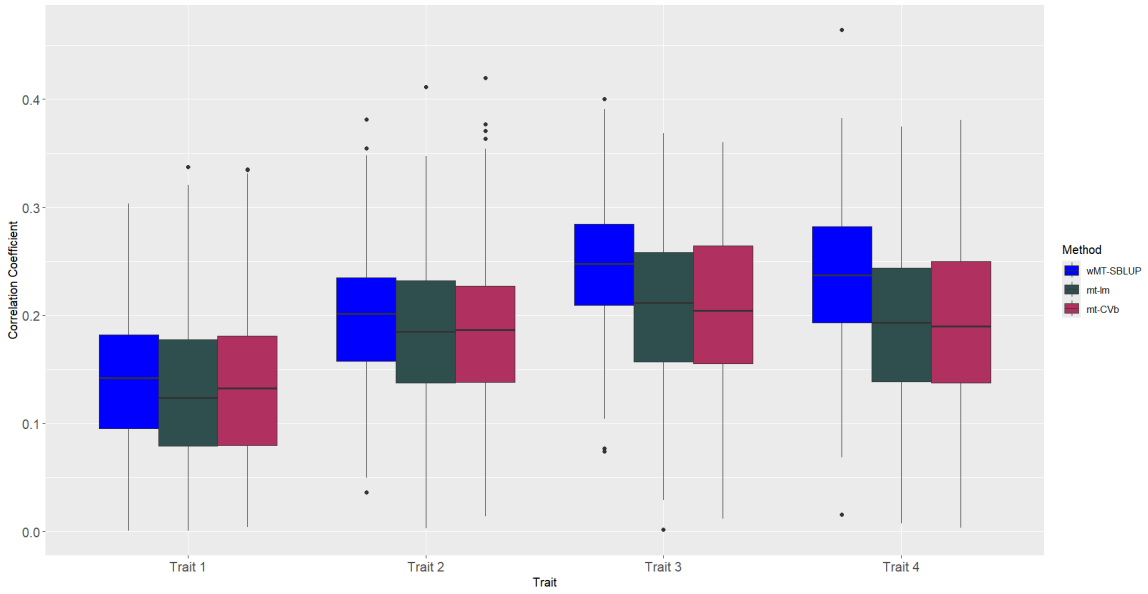

Figure 4.4: Comparison of correlation coefficients across methods when $n_t = 300$ for new methods

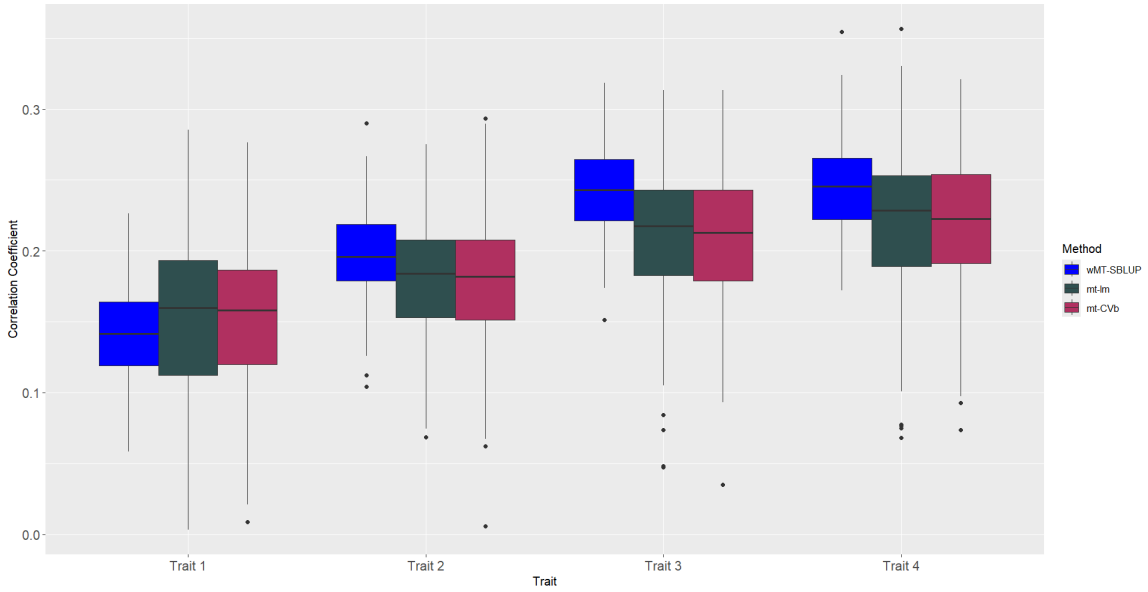

Figure 4.5: Comparison of correlation coefficients across methods when $n_t = 1350$ for new methods

4.2.2 Effect of the total number of SNPs p

Table 4.4 shows the performance of the two new multi-trait methods across traits as the number of SNPs decreases from 5000 to 3000. Compare to Figure 4.1, we observe that from Figure 4.6, as the number of SNPs decreases, the correlation coefficient for the methods increases and method **mt-CVb** exhibits higher performance compared to **mt-lm** for all the traits.

Table 4.4 shows the performance of the two new multi-trait methods across traits as the number of SNPs increases from 5000 to 7000. Compared to Figure 4.1, we observe that in Figure 4.7, as the number of SNPs increases, the correlation coefficient for the methods decreases and method **mt-lm** exhibits higher performance compared to **mt-CVb** for all the traits.

The results indicate that **mt-CVb** performs better with a smaller number of SNPs, while **mt-lm** performs better for larger numbers of SNPs. This difference could be because, with fewer SNPs, models like **mt-lm** are more susceptible to overfitting due to limited flexibility in handling variability. In contrast, the bagging approach in **mt-CVb** mitigates overfitting by averaging predictions across multiple subsamples, providing more robust results in such case. The multi-trait method **wMT-SBLUP** consistently outperforms the others, except for the first trait when $p = 3000$, where the multi-trait method **mt-CVb** shows the best performance, according to Table 4.4.

Table 4.4: Mean correlation coefficients when p changes for all methods

	Method	**Trait 1**	**Trait 2**	**Trait 3**	**Trait 4**
$p = 3000$	**Uni-trait**	0.169	0.151	0.173	0.184
	mtPGS	0.091	0.153	0.079	0.082
	mtPRS-PCA	0.036	0.233	0.248	0.224
	wMT-SBLUP	0.175	0.241	0.298	0.302
	mt-lm	0.181	0.215	0.262	0.265
	mt-CVb	0.186	0.216	0.261	0.266
$p = 5000$	**Uni-trait**	0.134	0.121	0.135	0.147
	mtPGS	0.082	0.132	0.073	0.077
	mtPRS-PCA	0.034	0.188	0.201	0.190
	wMT-SBLUP	0.149	0.199	0.245	0.251
	mt-lm	0.143	0.177	0.206	0.206
	mt-CVb	0.148	0.175	0.205	0.210
$p = 7000$	**Uni-trait**	0.123	0.105	0.114	0.128
	mtPGS	0.066	0.110	0.058	0.060
	mtPRS-PCA	0.033	0.163	0.170	0.156
	wMT-SBLUP	0.126	0.174	0.210	0.216
	mt-lm	0.120	0.150	0.178	0.173
	mt-CVb	0.112	0.153	0.174	0.174

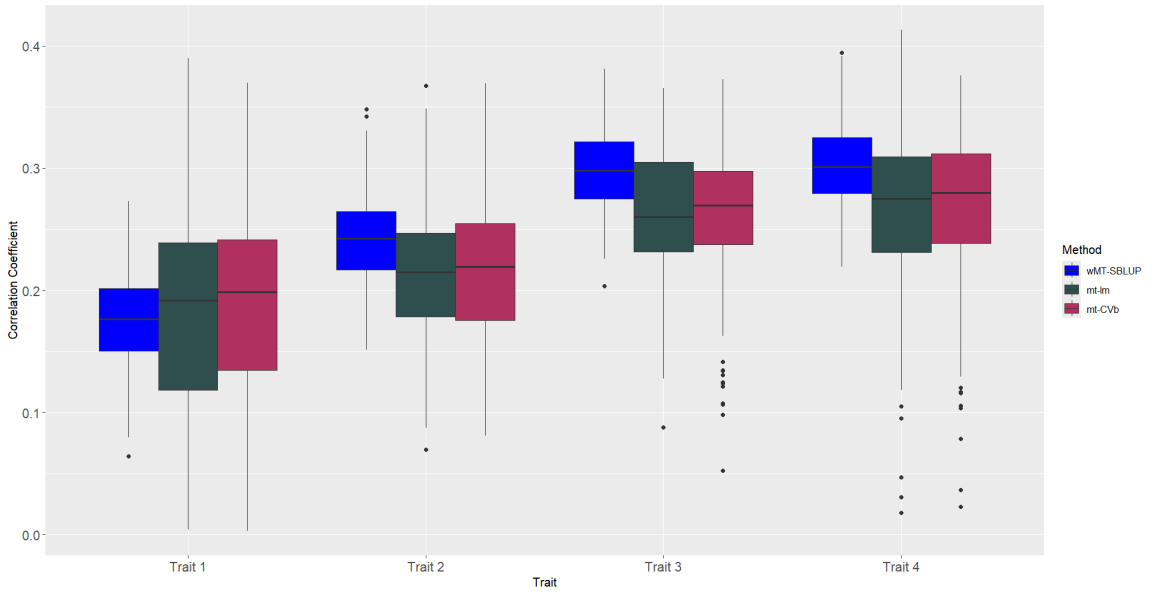

Figure 4.6: Comparison of correlation coefficients across methods when $p = 3000$ for new methods

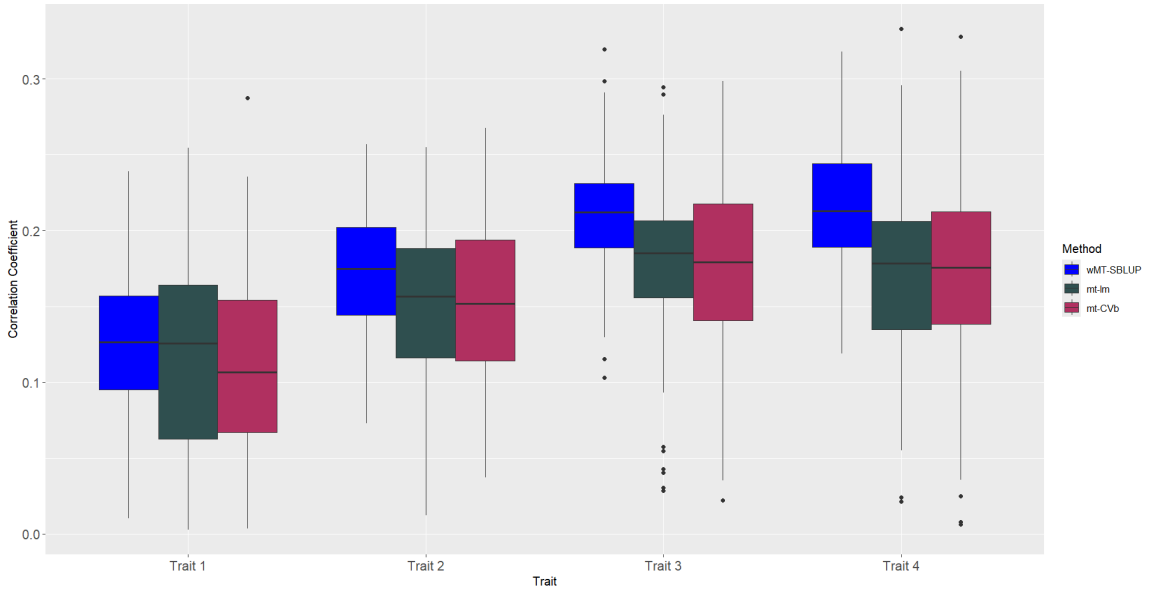

Figure 4.7: Comparison of correlation coefficients across methods when $p = 7000$ for new methods

4.2.3 How changing the true genetic effect sizes β affects performance

Table 4.5 presents the mean correlation coefficients obtained by varying the effect sizes (β). From the table, we observe that as the true effect sizes decrease, the correlation coefficient for all methods also decreases, and as the true effect sizes increase, the correlation coefficient for all methods increases. The method **mt-lm** performs better for lower heritability, while the method **mt-CVb** performs better for higher heritability.

Figure 4.8 includes boxplots of the correlation coefficients across methods when $H^2 \approx 0.3$, demonstrating that the performance of these two methods is very similar. The plot for the boxplots of the correlation coefficients across methods when $H^2 \approx 0.5$ is qualitatively similar to Figure 4.8, with the major difference being the Y-axis scale (larger values).

According to Table 4.5, the multi-trait method **wMT-SBLUP** consistently outperforms the others, except for the first trait when heritability for each trait approximately equals to 0.5, where the multi-trait method **mt-lm** shows the best performance.

4.2.4 The effect of variable heritability H^2 across traits

Figure 4.9 contains the boxplots for the correlation coefficients across methods, with H^2 values listed from highest to lowest. From Figure 4.9, the performance of the two new multi-trait methods shows a decreasing trend across traits. Additionally, the **mt-CVb** method exhibits higher performance compared to **mt-lm** for the first and fourth traits. Table 4.6 presents the mean correlation coefficients for different values

of H^2, according to Table 4.6, we can see that for the first trait, the uni-trait method performs better than all multi-trait methods.

Table 4.5: Mean correlation coefficients when β changes for all methods

	Method	**Trait 1**	**Trait 2**	**Trait 3**	**Trait 4**
$H^2 \approx 0.3$	**Uni-trait**	0.114	0.100	0.115	0.126
	mtPGS	0.062	0.106	0.059	0.065
	mtPRS-PCA	0.097	0.033	0.128	0.160
	wMT-SBLUP	0.124	0.166	0.212	0.219
	mt-lm	0.115	0.141	0.168	0.172
	mt-CVb	0.113	0.139	0.169	0.171
$H^2 \approx 0.4$	**Uni-trait**	0.134	0.121	0.135	0.147
	mtPGS	0.082	0.132	0.073	0.077
	mtPRS-PCA	0.034	0.188	0.201	0.190
	wMT-SBLUP	0.149	0.199	0.245	0.251
	mt-lm	0.143	0.177	0.206	0.206
	mt-CVb	0.148	0.175	0.205	0.210
$H^2 \approx 0.5$	**Uni-trait**	0.175	0.163	0.176	0.188
	mtPGS	0.128	0.181	0.102	0.097
	mtPRS-PCA	0.160	0.039	0.159	0.202
	wMT-SBLUP	0.197	0.262	0.308	0.310
	mt-lm	0.207	0.242	0.276	0.279
	mt-CVb	0.202	0.243	0.279	0.281

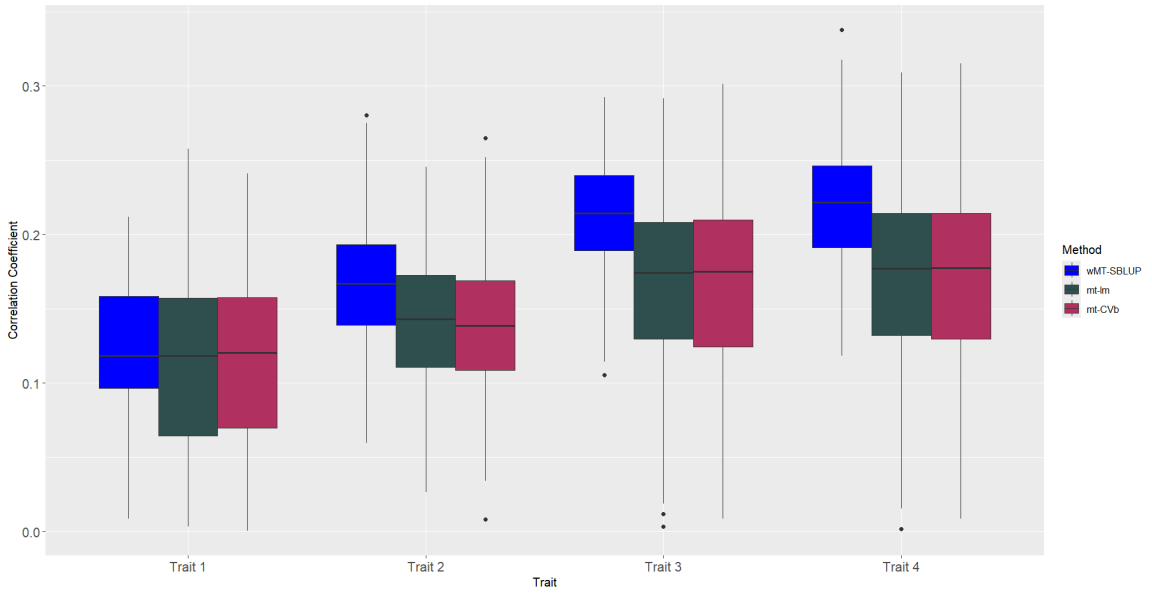

Figure 4.8: Comparison of correlation coefficients across methods when heritability approximately 0.3 for new methods

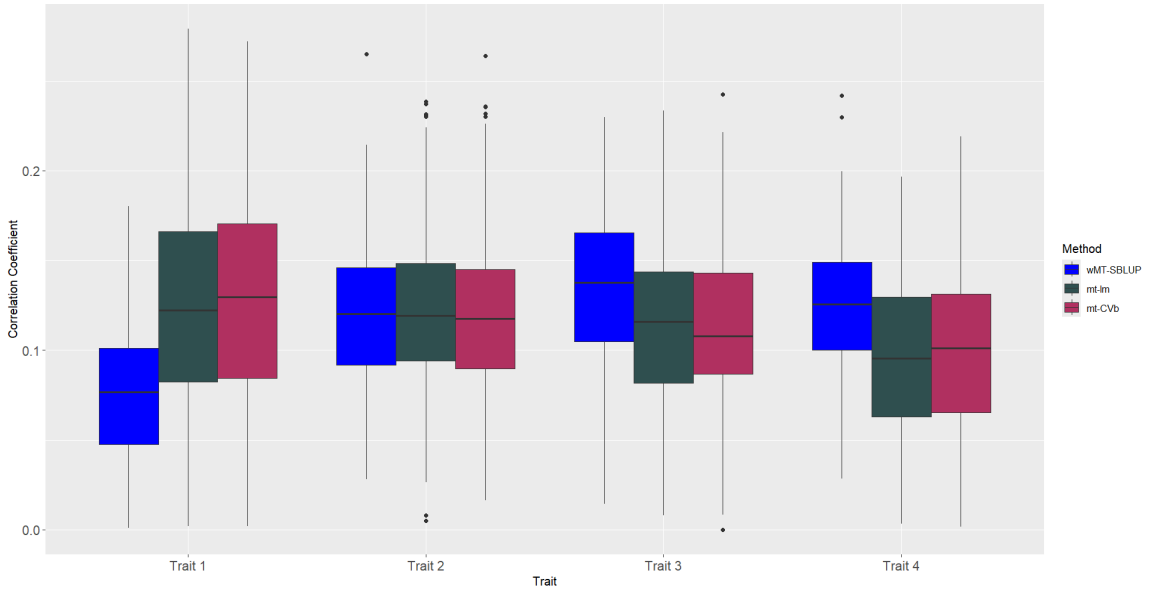

Figure 4.9: Comparison of correlation coefficients across methods when H^2 values are listed from high to low for new methods

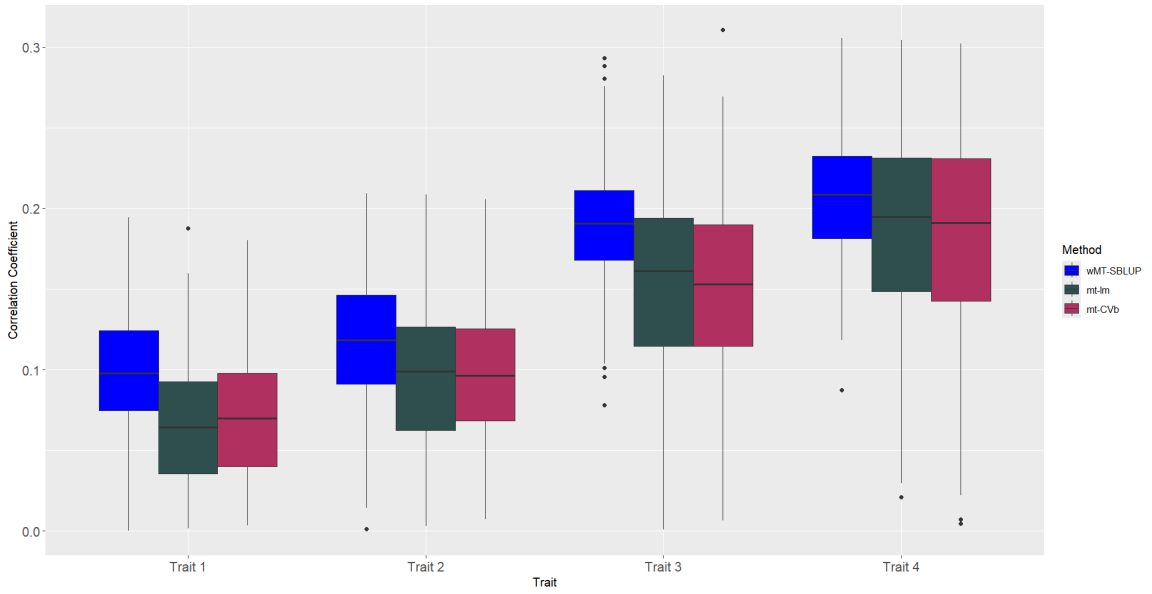

Figure 4.10: Comparison of correlation coefficients across methods when H^2 values are listed from low to high for new methods

Figure 4.10 contains the boxplots for the correlation coefficients across methods, with H^2 values listed from lowest to highest. From Figure 4.10, we observe the performance of the two new multi-trait methods shows a increasing trend across traits. Additionally, according to Table 4.6, the **mt-CVb** method exhibits lower average performance compared to **mt-lm** for all the traits except the first one.

The results indicate that heritability also affects the performance of the novel methods **mt-lm** and **mt-CVb**; higher heritability leads to better performance.

4.2.5 The effect of altering the genetic correlations

Figure 4.11 shows the boxplots for the average genetic correlation of each trait, which ranges from approximately 0.7 to 0.8. The plot for the range of approximately 0.1 to 0.2 is very similar to Figure 4.11. Table 4.7 includes the mean correlation coefficients

when the average genetic correlation of each trait changes. According to Figure 4.11 and Table 4.7, among the two novel methods, **mt-CVb** performs better in high correlation cases, while **mt-lm** performs better in low correlation cases.

Table 4.6: Mean correlation coefficients when H^2 different for all methods

	Method	**Trait 1**	**Trait 2**	**Trait 3**	**Trait 4**
H^2 high to low	**Uni-trait**	0.145	0.092	0.070	0.062
	mtPGS	0.067	0.156	0.079	0.103
	mtPRS-PCA	0.034	0.138	0.123	0.102
	wMT-SBLUP	0.075	0.120	0.136	0.124
	mt-lm	0.123	0.122	0.113	0.095
	mt-CVb	0.127	0.119	0.114	0.098
$H^2 \approx 0.4$	**Uni-trait**	0.134	0.121	0.135	0.147
	mtPGS	0.082	0.132	0.073	0.077
	mtPRS-PCA	0.034	0.188	0.201	0.190
	wMT-SBLUP	0.149	0.199	0.245	0.251
	mt-lm	0.143	0.177	0.206	0.206
	mt-CVb	0.148	0.175	0.205	0.210
H^2 low to high	**Uni-trait**	0.060	0.059	0.107	0.146
	mtPGS	0.046	0.047	0.046	0.049
	mtPRS-PCA	0.033	0.097	0.134	0.141
	wMT-SBLUP	0.097	0.118	0.191	0.207
	mt-lm	0.067	0.096	0.154	0.186
	mt-CVb	0.072	0.097	0.150	0.183

Table 4.7: Mean correlation coefficients when average genetic correlation changes for all methods

	Method	**Trait 1**	**Trait 2**	**Trait 3**	**Trait 4**
For each trait about 0.7–0.8	**Uni-trait**	0.134	0.143	0.136	0.151
	mtPGS	0.213	0.225	0.224	0.230
	mtPRS-PCA	0.195	0.138	0.084	0.196
	wMT-SBLUP	0.243	0.267	0.243	0.275
	mt-lm	0.224	0.249	0.228	0.256
	mt-CVb	0.226	0.249	0.230	0.260
About 0.4–0.5	**Uni-trait**	0.134	0.121	0.135	0.147
	mtPGS	0.082	0.132	0.073	0.077
	mtPRS-PCA	0.034	0.188	0.201	0.190
	wMT-SBLUP	0.149	0.199	0.245	0.251
	mt-lm	0.143	0.177	0.206	0.206
	mt-CVb	0.148	0.175	0.205	0.210
For each trait about 0.1–0.2	**Uni-trait**	0.134	0.130	0.130	0.159
	mtPGS	0.054	0.138	0.099	0.111
	mtPRS-PCA	0.059	0.036	0.118	0.205
	wMT-SBLUP	0.166	0.176	0.166	0.261
	mt-lm	0.134	0.140	0.120	0.169
	mt-CVb	0.121	0.140	0.118	0.166

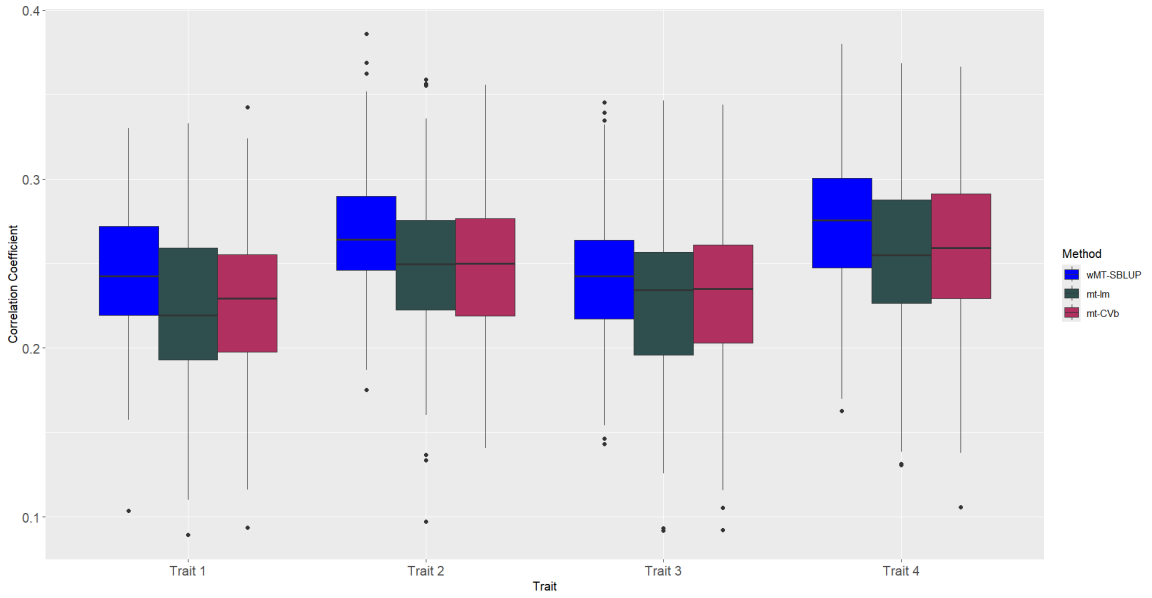

Figure 4.11: Comparison of correlation coefficients across methods when the off-diagonal elements of the genetic correlation matrix change (with an average genetic correlation for each trait about 0.7–0.8) for new methods

4.3 Results for binary focal trait

Figure 4.12 presents boxplots for the binary focal traits, allowing for a comparison of AUC values across methods for each trait. From Figure 4.12, we observe that the performance of the two new multi-trait methods, **mt-lm** and **mt-CVb**, differs more significantly than in the continuous scenario. In the binary scenario, **mt-lm** demonstrates a higher average performance compared to **mt-CVb** across all traits. Table 4.8 summarizes the average performance of each method across traits (including those from Chapter 3 and the new methods), rounded to three decimal places. The multi-trait method **wMT-SBLUP** no longer consistently outperforms both new

methods, **mt-lm** achieves the best performance for all traits. Overall, the new methods remain competitive, even though **mt-CVb** does not always deliver the best results.

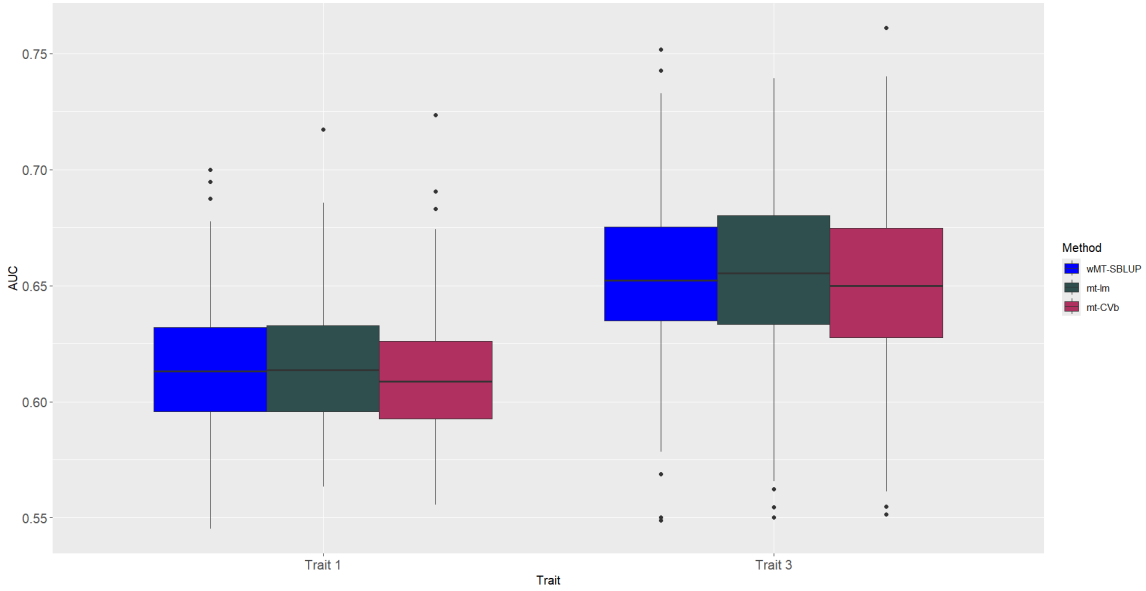

Figure 4.12: Comparison of AUC values across methods using original parameters for new methods

Table 4.8: Mean AUC values of original parameters for all methods

	Trait 1	**Trait 3**
Uni-trait	0.583	0.582
mtPGS	0.602	0.648
mtPRS-PCA	0.599	0.642
wMT-SBLUP	0.614	0.655
mt-lm	0.616	0.656
mt-CVb	0.610	0.649

4.3.1 How changing the true genetic effect sizes β affects performance

Nothing changes significantly compared to the previous results for the case where heritability is approximately equal to 0.3 for each trait. Therefore, the plot for this case will not be included.

Table 4.9 presents the mean AUC values for all methods when varying β, while Figure 4.13 displays the boxplots for the three methods (**mt-lm**, **mt-CVb** and **wMT-SBLUP**). According to Figure 4.13 and Table 4.9, when the heritability for each trait is approximately 0.5, the multi-trait method **mt-lm** consistently demonstrates the best performance among all methods in the binary scenario.

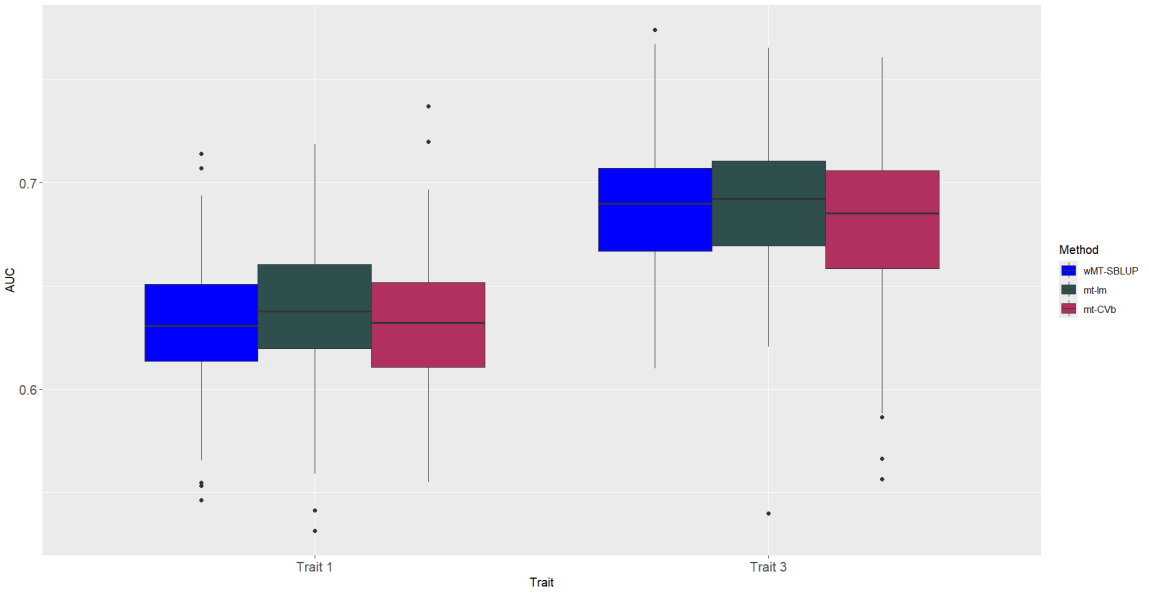

Figure 4.13: Comparison of AUC values across methods when heritability approximately 0.5 for new methods

Table 4.9: Mean AUC values when β changes for all methods

	$H^2 \approx 0.3$		$H^2 \approx 0.4$		$H^2 \approx 0.5$	
	Trait 1	**Trait 3**	**Trait 1**	**Trait 3**	**Trait 1**	**Trait 3**
Uni-trait	0.577	0.580	0.583	0.582	0.595	0.596
mtPGS	0.598	0.635	0.602	0.648	0.618	0.683
mtPRS-PCA	0.604	0.616	0.599	0.642	0.622	0.627
wMT-SBLUP	0.609	0.643	0.614	0.655	0.633	0.688
mt-lm	0.605	0.639	0.616	0.656	0.640	0.690
mt-CVb	0.603	0.637	0.610	0.649	0.630	0.681

4.3.2 Effect of other parameters

Similar to Section 3.3.2, for the binary scenario, we altered all the parameters as in the continuous scenario. For parameters not mentioned above, the results were qualitatively very similar to those observed for continuous focal traits.

Chapter 5

Correlated SNPs

In Chapters 3 and 4, we focused exclusively on simulations involving independent SNPs. However, in real-life, SNPs are often correlated. This chapter examines the performance of the methods discussed in Chapter 2 and Chapter 4 under the consideration of correlated SNPs, comparing their performance across various scenarios.

All the parameters remain the same as in Chapter 3, except for the genotype matrix G; since heritability and the genetic correlation matrix have the greatest impact on performance, as discussed in Chapters 3 and 4. We focus exclusively on these two parameters in this chapter. Recall there are a total of n observations and p columns of single nucleotide polymorphisms (SNPs) as described in Chapter 3. Specifically, there are $n_g = 1800$ observations for G_g, $n_t = 900$ observations for G_t, and $n_v = 900$ observations for G_v, with a total of $p = 5000$ SNPs. The minor allele frequency (MAF) is different across SNPs; and they assumed to follow a uniform distribution $\text{MAF} \sim \text{Uniform}(0.05, 0.5)$.

To generate the correlated SNP data, the total number of p SNPs is partitioned into $m = p/d$ blocks, where d represents the number of SNPs in each block. For n

individuals, each carrying two copies of each autosomal (non-sex) chromosome, two independent multivariate normal vectors Z of dimension d were generated within each block. Both Z were sampled from a multivariate normal distribution with a mean vector of zero and a specified covariance matrix, reflecting the underlying linkage disequilibrium structure among the SNPs in the block. The covariance which is defined with diagonal elements are set to 1, while the off-diagonal elements decay according to an auto-regressive structure of order 1, specifically $0.8^{|i-j|}$, where the i^{th} and j^{th} SNPs are in the same block. This choice reflects the biological principle that SNPs located closer together on the chromosome tend to exhibit stronger correlations, which is reasonable given the structure of genetic linkage.

The cutoff values help translate continuous allele probabilities into binary values, reflecting the presence or absence of alleles based on the normal distribution threshold. These cutoff values are determined using the MAF of each SNP and the quantiles of the standard normal distribution. Specifically, the cutoff for each SNP is given by $\Phi^{-1}(\text{MAF})$, where Φ is the cumulative distribution function (CDF) of the standard normal distribution. The cutoff values are then applied to each SNP within a block and repeated across rows to match the dimensions of the generated matrix Z. For each individual, two d-dimensional vectors of ones and zeros are created, each representing one copy of the genome block; each element in the vector corresponds to the presence (1) or absence (0) of the allele. The genotype of an individual at this block is calculated by summing these two vectors.

The SNPs that truly impact the traits selected in this chapter are the middle ones, specifically the tenth SNP in each block (when the block size is 20), for the first 18 blocks. Choosing edge SNPs (the first SNP or the last SNP) may not fully capture

the correlation structure within the block, by selecting the middle SNPs, we ensure that the SNPs are more representative of the genetic variation within the block, capturing the central correlation patterns that are more relevant for modeling trait associations. Appendix A includes thea function for generating the correlated SNPs matrix as described above.

5.1 Continuous focal trait

From Figure 5.1, we observe that the performance of the multi-trait method **wMT-SBLUP** method consistently outperforms the others except for the second trait, while **mtPGS** exhibits lower performance. For all the multi-trait methods except **mt-lm**, the first trait shows lower performance, while the other three traits display significantly higher results. The **mtPGS** method performs best (compared to its own performance across different traits) for the second trait, but still lags behind all the other multi-trait methods. The uni-trait method shows consistent performance across all traits but delivers the worst results for all traits, except for the first one, when compared to the multi-trait methods. However, it does outperform the multi-trait method **mtPGS** for the third trait. Recall that the first trait has the lowest average genetic correlation with the other traits (with the average correlation above 0.45 for the others, and the third trait having the highest at 0.53, while the first trait's is approximately 0.2). The performance of the multi-trait method **mtPRS-PCA** depends on the genetic correlation matrix, which explains why it performed better for the third trait than for the other traits. Overall, the novel methods remain competitive, particularly **mt-lm**, which delivers the highest performance for both the first and second traits. Table 5.1 provides a summary of the average performance for

each method across traits, rounded to three decimal places.

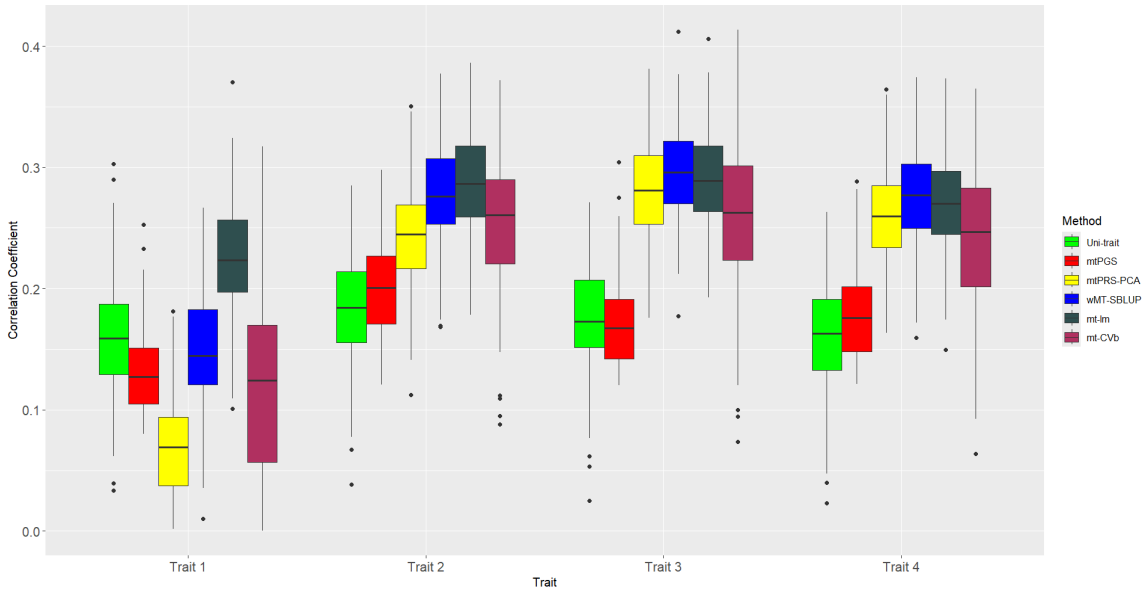

Figure 5.1: Comparison of correlation coefficients across methods using original parameters when SNPs are correlated

Table 5.1: Mean correlation coefficients of original parameters when SNPs are correlated

	Trait 1	**Trait 2**	**Trait 3**	**Trait 4**
Uni-trait	0.158	0.183	0.176	0.161
mtPGS	0.131	0.199	0.172	0.177
mtPRS-PCA	0.068	0.243	0.281	0.261
wMT-SBLUP	0.150	0.277	0.295	0.278
mt-lm	0.226	0.288	0.290	0.272
mt-CVb	0.119	0.254	0.259	0.240

5.1.1 How changing the true genetic effect sizes β affects performance

Table 5.2 includes the average correlation coefficients when altering the β values and considering correlated SNPs. We observe that as the true effect sizes decrease, the correlation coefficients of all methods decrease from Table 5.2. According to Figure 5.2, across all multi-trait methods, method **mt-lm** delivers the highest performance for both the first and second traits, method **wMT-SBLUP** delivers the highest performance for both the third and fourth traits, and method **mt-CVb** is not as competitive as before when heritability decreases.

We observe that as the true effect sizes increase, the correlation coefficients of all methods increase from Table 5.2. The plot for $H^2 \approx 0.5$ is very similar to Figure 5.2. According to the plot and Table 5.2, we observe that across all multi-trait methods, method **mt-lm** and method **mt-CVb** deliver better performance than the other methods for the first trait, method **mt-lm** delivers the highest performance for both the first and second traits, method **wMT-SBLUP** delivers the highest performance for both the third and fourth traits.

According to Table 5.2, for the two novel methods, method **mt-lm** exhibits higher average performance compared to **mt-CVb** for all traits in all cases.

Table 5.2: Mean correlation coefficients when β changes and SNPs are correlated

	Method	**Trait 1**	**Trait 2**	**Trait 3**	**Trait 4**
$H^2 \approx 0.3$	**Uni-trait**	0.127	0.155	0.149	0.132
	mtPGS	0.120	0.182	0.164	0.170
	mtPRS-PCA	0.064	0.214	0.245	0.224
	wMT-SBLUP	0.119	0.240	0.257	0.239
	mt-lm	0.180	0.250	0.251	0.230
	mt-CVb	0.098	0.211	0.214	0.199
$H^2 \approx 0.4$	**Uni-trait**	0.158	0.183	0.176	0.161
	mtPGS	0.131	0.199	0.172	0.177
	mtPRS-PCA	0.068	0.243	0.281	0.261
	wMT-SBLUP	0.150	0.277	0.295	0.278
	mt-lm	0.226	0.288	0.290	0.272
	mt-CVb	0.119	0.254	0.259	0.240
$H^2 \approx 0.5$	**Uni-trait**	0.223	0.238	0.231	0.219
	mtPGS	0.166	0.237	0.186	0.185
	mtPRS-PCA	0.077	0.294	0.346	0.331
	wMT-SBLUP	0.214	0.343	0.365	0.351
	mt-lm	0.317	0.360	0.361	0.347
	mt-CVb	0.219	0.335	0.344	0.327

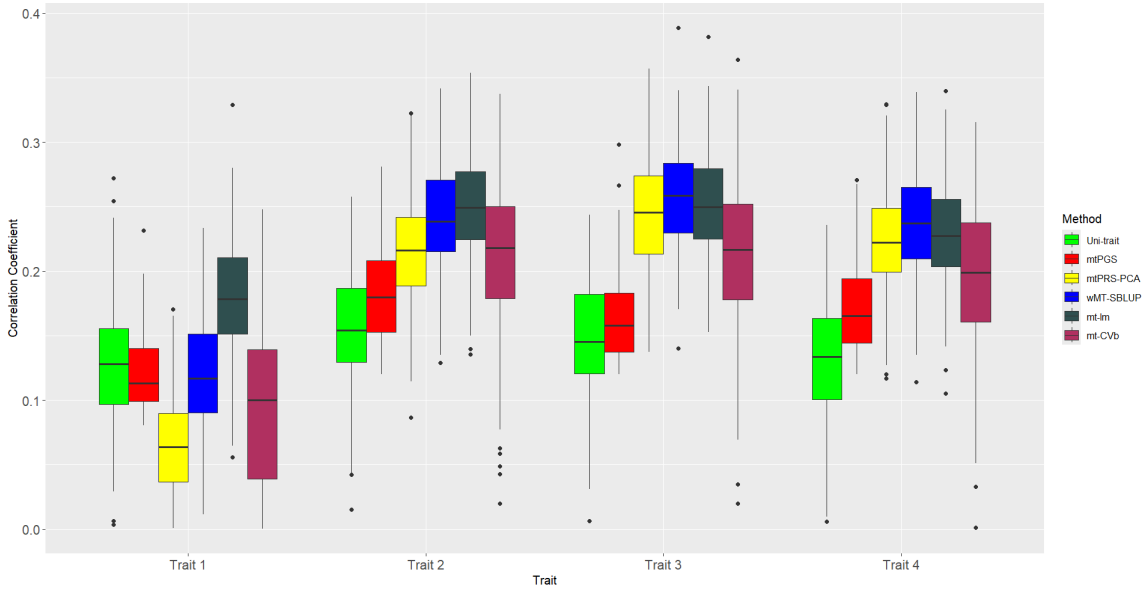

Figure 5.2: Comparison of correlation coefficients across methods when heritability approximately 0.3 for all methods

5.1.2 The effect of altering the genetic correlations

Figure 5.3 presents the boxplots for all the methods when the average genetic correlation of each trait is approximately between 0.7 and 0.8 with correlated SNPs. From Figure 5.3, we observe that for the two new multi-trait methods, method **mt-lm** always exhibits higher performance compared to **mt-CVb** for all traits. Among all the multi-trait methods from Table 5.3, **wMT-SBLUP** does not consistently outperform the others anymore, while method **mtPGS** performs the best for all traits. That might be because when the SNPs are correlated, increasing the genetic correlation between traits allows the other traits to better contribute to predicting the focal trait.

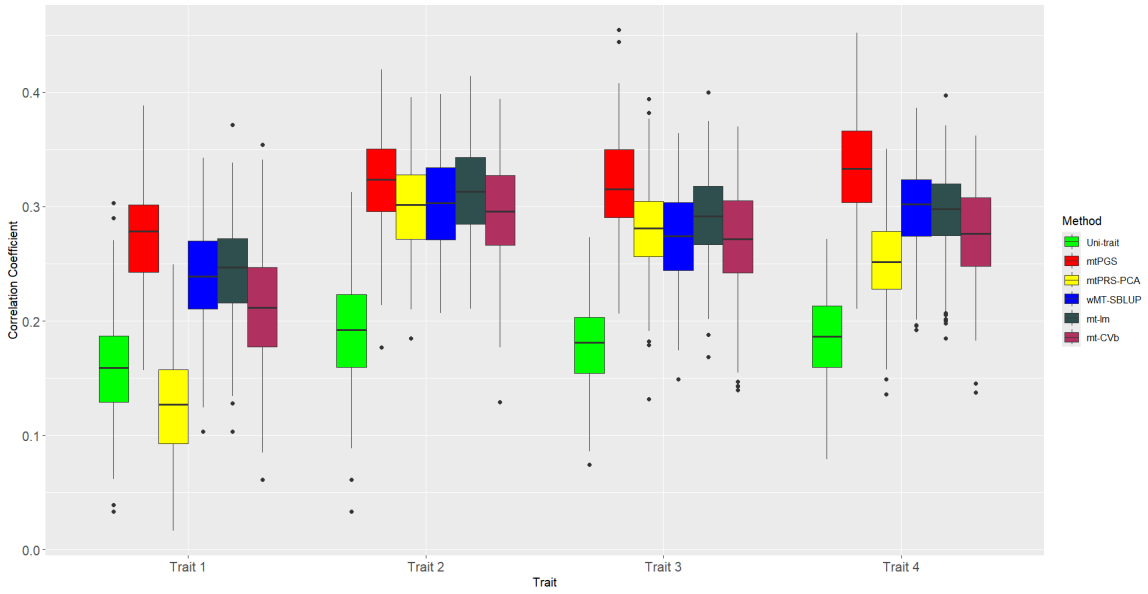

Figure 5.3: Comparison of correlation coefficients across methods when the off-diagonal elements of the genetic correlation matrix change (with an average genetic correlation for each trait about 0.7–0.8) for all methods

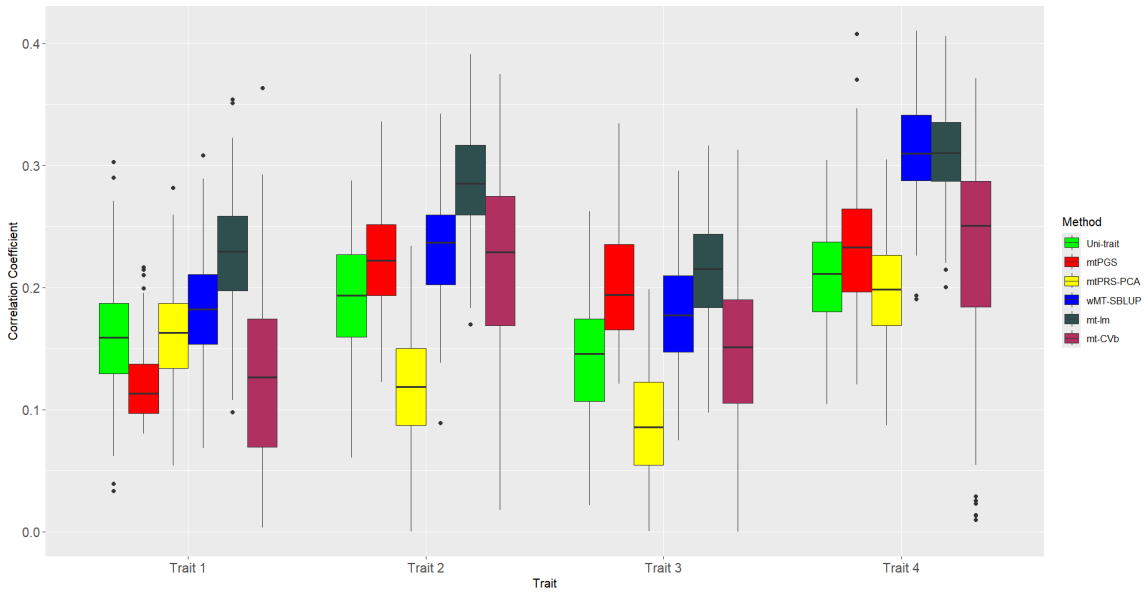

Figure 5.4: Comparison of correlation coefficients across methods when the off-diagonal elements of the genetic correlation matrix change (with an average genetic correlation for each trait about 0.1–0.2) for all methods

Table 5.3: Mean correlation coefficients when average genetic correlation changes and SNPs are correlated

	Method	**Trait 1**	**Trait 2**	**Trait 3**	**Trait 4**
For each trait about 0.7–0.8	**Uni-trait**	0.158	0.191	0.179	0.184
	mtPGS	0.273	0.322	0.319	0.332
	mtPRS-PCA	0.124	0.300	0.280	0.251
	wMT-SBLUP	0.240	0.302	0.272	0.297
	mt-lm	0.245	0.314	0.291	0.295
	mt-CVb	0.210	0.295	0.271	0.275
About 0.4–0.5	**Uni-trait**	0.158	0.183	0.176	0.161
	mtPGS	0.131	0.199	0.172	0.177
	mtPRS-PCA	0.068	0.243	0.281	0.261
	wMT-SBLUP	0.150	0.277	0.295	0.278
	mt-lm	0.226	0.288	0.290	0.272
	mt-CVb	0.119	0.254	0.259	0.240
For each trait about 0.1–0.2	**Uni-trait**	0.158	0.192	0.142	0.208
	mtPGS	0.120	0.223	0.202	0.232
	mtPRS-PCA	0.160	0.119	0.088	0.198
	wMT-SBLUP	0.182	0.231	0.178	0.312
	mt-lm	0.228	0.287	0.214	0.311
	mt-CVb	0.125	0.215	0.147	0.234

Figure 5.4 presents the boxplots for all the methods when the average genetic correlation of each trait is approximately between 0.1 and 0.2 with correlated SNPs. From Figure 5.4, we observe that as the average genetic correlation for each trait

decreases to approximately 0.1–0.2, the multi-trait methods lose their advantage, showing diminished performance compared to higher correlation scenarios. Among all the multi-trait methods, method **mt-lm** performs the best for all traits except for the fourth one, method **mtPRS-PCA** even worse than the uni-trait method for all the traits except the first one.

5.2 Binary focal trait

From Figure 5.5, we observe that the multi-trait method **wMT-SBLUP** performs the best for the first trait, while **mt-lm** performs the best for the third trait. In contrast, **mtPRS-PCA** exhibits the lowest performance for the first trait, and **mtPGS** shows the lowest performance for the fourth trait. The uni-trait method demonstrates consistent performance across all traits but yields the poorest results in every case. Overall, the new methods remain competitive, particularly **mt-lm**. Table 5.4 provides a summary of the average performance for each method across traits, rounded to three decimal places.

Table 5.4: Mean AUC values of original parameters when SNPs are correlated

	Trait 1	**Trait 3**
Uni-trait	0.587	0.589
mtPGS	0.611	0.643
mtPRS-PCA	0.610	0.651
wMT-SBLUP	0.625	0.666
mt-lm	0.620	0.671
mt-CVb	0.618	0.656

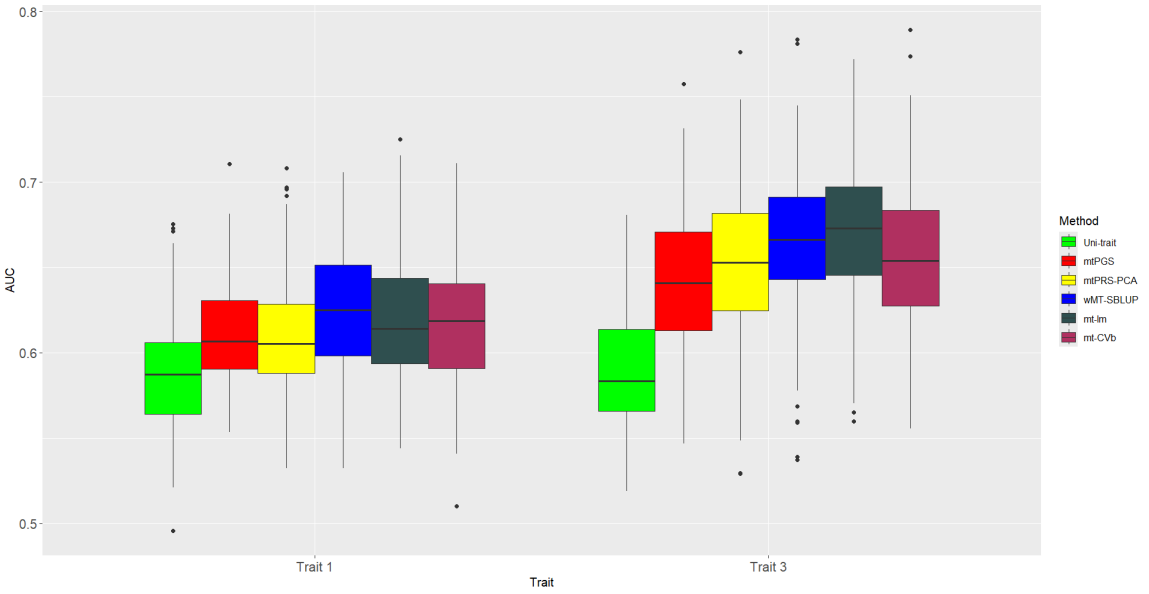

Figure 5.5: Comparison of AUC values across methods using original parameters when SNPs are correlated

5.2.1 How changing the true genetic effect sizes β affects performance

Nothing changes significantly compared to the continuous scenario. The patterns observed in the continuous scenario (for the correlated SNPs) hold true, with similar trends in performance across methods.

5.2.2 The effect of altering the genetic correlations

Figure 5.6 presents the boxplots when the average genetic correlation for each trait is approximately 0.7–0.8, while Figure 5.7 presents the boxplots when the average genetic correlation for each trait is approximately 0.1–0.2. Table 5.3 includes the average AUC values when altering the genetic correlations.

From Figure 5.6, we observe that for the two new multi-trait methods, method **mt-lm** exhibits higher performance compared to **mt-CVb** for all the traits. Among all the multi-trait methods from Table 5.3, **wMT-SBLUP** does not consistently outperform the others anymore, method **mt-lm** performs the best for the third trait and both new methods demonstrate higher performance than **wMT-SBLUP** for the third trait.

From Figure 5.7, we observe that as the average genetic correlation for each trait decreases to approximately 0.1–0.2, the multi-trait methods lose their advantage, showing diminished performance compared to higher correlation scenarios although they still have better performance than uni-trait method. According to Table 5.3, among all the multi-trait methods, method **wMT-SBLUP** performs the best for the first trait, method **mt-CVb** performs the best for the third trait and both new methods demonstrate higher performance than all the other methods for the third trait.

Table 5.5: Mean AUC values when average genetic correlation changes and SNPs are correlated

	For each trait about 0.7–0.8		Original about 0.4–0.5		For each trait about 0.1–0.2	
	Trait 1	**Trait 3**	**Trait 1**	**Trait 3**	**Trait 1**	**Trait 3**
Uni-trait	0.587	0.590	0.587	0.589	0.587	0.586
mtPGS	0.636	0.640	0.611	0.643	0.610	0.612
mtPRS-PCA	0.627	0.657	0.610	0.651	0.618	0.617
wMT-SBLUP	0.654	0.658	0.625	0.666	0.628	0.622
mt-lm	0.649	0.671	0.620	0.671	0.616	0.623
mt-CVb	0.650	0.666	0.618	0.656	0.617	0.626

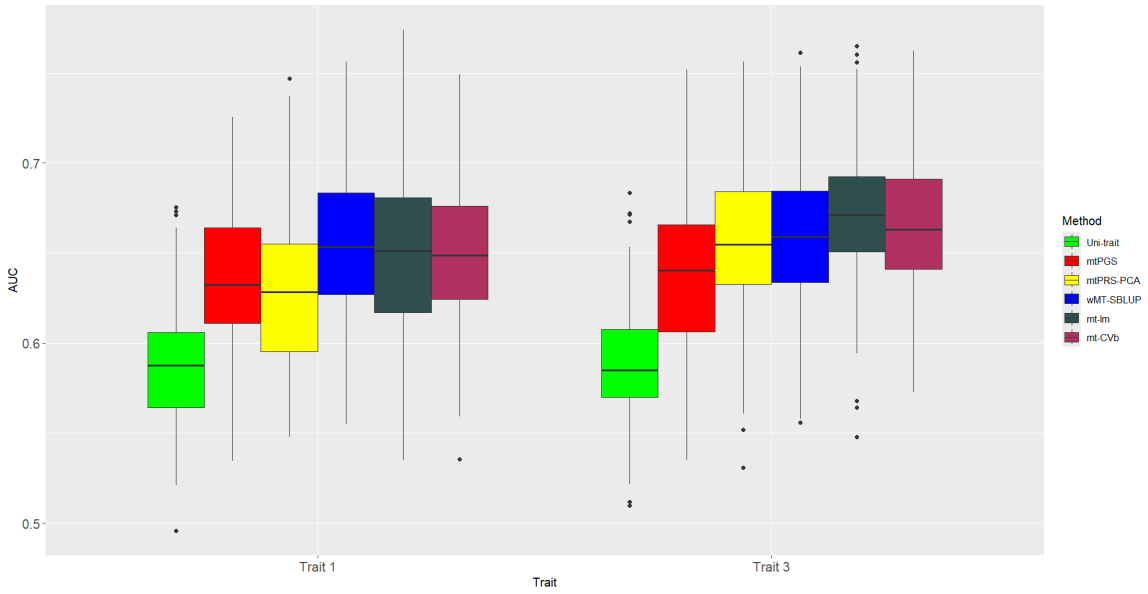

Figure 5.6: Comparison of AUC values across methods when the off-diagonal elements of the genetic correlation matrix change (with an average genetic correlation for each trait about 0.7–0.8) for all methods

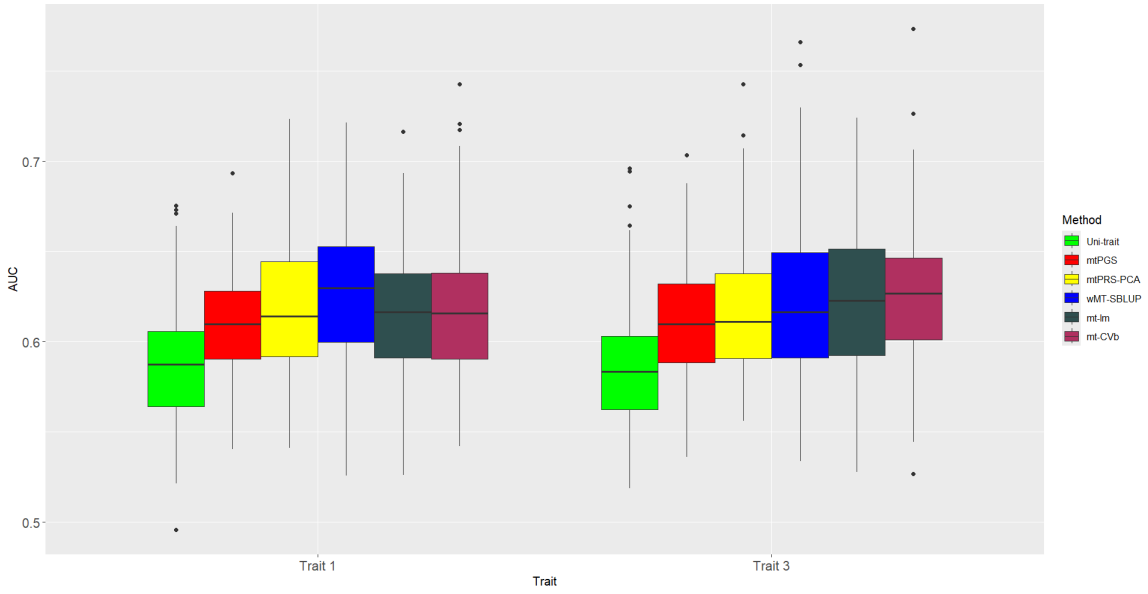

Figure 5.7: Comparison of AUC values across methods when the off-diagonal elements of the genetic correlation matrix change (with an average genetic correlation for each trait about 0.1–0.2) for all methods

Chapter 6

Conclusion

6.1 Discussion

Based on Chapter 3 and Chapter 4, we can conclude that all parameters influence the final performance in both continuous and binary scenarios, particularly heritability and the genetic correlation matrix. In both scenarios, the genome-wide association study (GWAS) sample size n_g is positively related to performance, while the number of SNPs p is negatively related to performance. Although true effect sizes (β) and minor allele frequency (MAF) also contribute to the genetic correlation matrix, their impact is relatively minor compared to heritability. Under the same heritability assumptions, lower heritability and lower genetic correlation across traits reduce the performance of multi-trait methods. Lower heritability is more reflective of real-life scenarios, and even when heritability is low, multi-trait methods generally outperform uni-trait methods, although their advantage decreases. The multi-trait method **wMT-SBLUP** consistently outperforms all other methods across Chapter 3 and Chapter 4, the **mtPGS** method shows improved performance when the genetic correlation is

high, while the uni-trait method generally exhibits lower performance compared to the multi-trait methods.

The performance of the two novel multi-trait methods developed in Chapter 4, **mt-lm** and **mt-CVb**, is competitive with the other three multi-trait methods, although they do not always deliver the best results. In certain cases, one or both of the two novel methods can even outperform **wMT-SBLUP**. The advantage of the method **mt-lm** is that it is faster when dealing with large datasets, as shown in Table 6.1.

Table 6.1: The average time over 100 datasets for three of the multi-trait methods

	Average time over 100 **datasets (in seconds)**
wMT-SBLUP	0.026
mt-lm	0.022
mt-CVb	1.425

Chapter 5 focuses solely on the cases where heritability and the genetic correlation matrix change when single-nucleotide polymorphisms (SNPs) are correlated. Method **mtPGS** shows overall higher performance compared to its results in Chapter 3 and Chapter 4, where SNPs are independent. Additionally, the performance of method **mt-lm** remains more consistent across all traits compared to the other multi-trait methods.

6.2 Limitations

The number of replicates in the simulation is relatively small ($R = 200$), which may introduce some variability in the results. Moreover, both the sample sizes (n_g and n_t)

and the number of SNPs p used in this study are much smaller than those typically found in real-life datasets (where sample sizes can reach up to one million and the number of SNPs may be in the several millions). SNPs selection was not performed in this simulation, which could be an area for future work. Additionally, the heritability values used in the study were calculated as true heritabilities, whereas in real-life scenarios, these would generally be estimated. Furthermore, linkage disequilibrium (LD) scores were not used in the calculation of heritability, which may affect the accuracy and relevance of the results in comparison to real-life data.

In the binary scenario, we consider a total of four traits: two binary and two continuous, with the primary focus of this thesis in this scenario being on the binary focal trait, while the continuous traits are not considered. In Chapters 3 and 4, the section discussing changes to the number of SNPs p only includes SNPs that are not associated with any of the traits when increasing p. Chapter 5 focuses on exploring heritability and the genetic correlation matrix within the simulation.

6.3 Future work

To make the simulation more realistic, we plan to increase the number of replicates, as well as the sample size and the number of SNPs. Additionally, we may explore SNP selection using clumping and thresholding (C+T), as this thesis did not include any selection process. Furthermore, we will consider using publicly available linkage disequilibrium (LD) scores to calculate the estimated heritability and genetic correlation matrix, which will better reflect the true genetic architecture of traits.

We would also like to examine the performance of the continuous focal trait in the binary scenario to determine if there are any notable differences compared to the

continuous scenario. We plan to explore the effects of adding causal SNPs as p increases. For the scenario considered in Chapter 5, we aim to further investigate the effects of parameters on correlated SNPs.

We plan to explore methods for generating polygenic risk scores (PRS) across global populations (cross-ethnic) using real-life datasets, such as the UK Biobank. As discussed in Chapter 1, genetic architectures can vary among populations due to differences in allele frequencies, linkage disequilibrium patterns, and environmental interactions. These variations can affect the accuracy of PRS when applied to diverse groups, underscoring the need for accurate and equitable risk assessments across different populations. To develop a final PRS that can accurately quantify genetic risk across all ethnic groups, it is necessary to combine PRS from multiple populations. This can also be related to the approach taken in this thesis. Specifically, we can calculate separate GWAS for each ethnic group, obtain their respective weights, and then combine the cross-ethnic PRS into a single score. This method improving the generalizability of the polygenic risk scores across diverse ethnic groups.

Appendix A

Related R Code

This appendix includes the function used to generate the correlated single nucleotide polymorphisms (SNPs) in Chapter 5, in case there is any confusion.

Listing A.1: Function for generating the correlated SNPs matrix

```
generate_snp_matrix <- function(n, p, maf, block_size=20)
{
  num_blocks <- ceiling(p/block_size)
  #initialize genotype matrix
  snp_matrix <- matrix(nrow=n, ncol=p)
  cutoff <- qnorm(maf) #cutoff
  #covariance matrix for each block
  sigma <- diag(block_size)
  sigma <- 0.8^(abs(row(sigma)-col(sigma)))
  mu <- rep(0, block_size)
  for (i in 1:num_blocks)
  {
```

```
13      Z <- mvrnorm(n=2*n, mu=mu, Sigma=sigma)
14      index <- ((i-1)*block_size+1):(i*block_size)
15      cutoff_i <- cutoff[index]
16      cutoff_i <- matrix(cutoff_i, nrow=2*n, ncol=block_size,
           byrow=TRUE)
17      Z1 <- Z<cutoff_i
18      snp_matrix[, index] <- Z1[1:n, ]+Z1[((n+1):(2*n)), ]
19      while (any(colMaxs(snp_matrix[, index])-colMins(snp_matrix
           [, index]) == 0)) #avoid same value
20      {
21        Z <- mvrnorm(n=2*n, mu=mu, Sigma=sigma)
22        index <- ((i-1)*block_size-1):(i*block_size)
23        cutoff_i <- cutoff[index]
24        cutoff_i <- matrix(cutoff_i, nrow=2*n, ncol=block_size,
             byrow=TRUE)
25        Z1 <- Z<cutoff_i
26        snp_matrix[, index] <- Z1[1:n, ]+Z1[((n+1):(2*n)), ]
27      }
28    }
29    return(snp_matrix)
30  }
```

Bibliography

C. Albiñana, Z. Zhu, A. J. Schork, et al. Multi-PGS enhances polygenic prediction by combining 937 polygenic scores. *Nature Communications*, 14(1):4702, 2023.

B. Bulik-Sullivan, H. K. Finucane, V. Anttila, et al. An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, 47(11):1236–1241, 2015.

M. E. Goddard, N. R. Wray, K. Verbyla, et al. Estimating effects and making predictions from genome-wide marker data. *Statistical Science*, 24(4):517–529, 2009.

L. Kachuri, N. Chatterjee, J. Hirbo, et al. Principles and methods for transferring polygenic risk scores across global populations. *Nature Reviews Genetics*, 25(1): 8–25, 2024.

H. Klinkhammer, C. Staerk, C. Maj, et al. A statistical boosting framework for polygenic risk scores based on large-scale genotype data. *Frontiers in Genetics*, 13: 1076440, 2023.

E. Krapohl, H. Patel, S. Newhouse, et al. Multi-polygenic score approach to trait prediction. *Molecular Psychiatry*, 23(5):1368–1374, 2018.

M. Kuhn. Applied Predictive Modeling. *Springer*, 2013.

Y. Ma and X. Zhou. Genetic prediction of complex traits with polygenic scores: a statistical review. *Trends in Genetics*, 37(11):995–1011, 2021.

R. M. Maier, Z. Zhu, S. H. Lee, et al. Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nature Communications*, 9 (1):989, 2018.

G. Paré, S. Mao, and W. Q. Deng. A machine-learning heuristic to improve gene score prediction of polygenic traits. *Scientific Reports*, 7(1):12665, 2017.

C. Xu, S. K. Ganesh, and X. Zhou. mtPGS: Leverage multiple correlated traits for accurate polygenic score construction. *The American Journal of Human Genetics*, 110(10):1673–1689, 2023.

N. Zaitlen and P. Kraft. Heritability in the genome-wide association era. *Human Genetics*, 131:1655–1664, 2012.

S. Zhai, B. Guo, B. Wu, et al. Integrating multiple traits for improving polygenic risk prediction in disease and pharmacogenomics GWAS. *Briefings in Bioinformatics*, 24(4):1–11, 2023.

S. Zhai, D. V. Mehrotra, and J. Shen. Applying polygenic risk score methods to pharmacogenomics GWAS: challenges and opportunities. *Briefings in Bioinformatics*, 25(1):1–18, 2024.