Statistical methods for integrative analysis of genomic data

Jingsi Ming

Follow this and additional works at: https://repository.hkbu.edu.hk/etd_oa

Recommended Citation
https://repository.hkbu.edu.hk/etd_oa/545

This Thesis is brought to you for free and open access by the Electronic Theses and Dissertations at HKBU Institutional Repository. It has been accepted for inclusion in Open Access Theses and Dissertations by an authorized administrator of HKBU Institutional Repository. For more information, please contact repository@hkbu.edu.hk.
DATE: August 24, 2018

STUDENT'S NAME: MING Jingsi

THESIS TITLE:  Statistical Methods for Integrative Analysis of Genomic Data

This is to certify that the above student's thesis has been examined by the following panel members and has received full approval for acceptance in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chairman:  Prof Ng Joseph K Y
Professor, Department of Computer Science, HKBU
(Designated by Dean of Faculty of Science)

Internal Members:  Dr Peng Heng
Associate Professor, Department of Mathematics, HKBU
(Designated by Head of Department of Mathematics)

Dr Ling Leevan
Associate Professor, Department of Mathematics, HKBU

External Members:  Dr Zhao Xingqiu
Associate Professor
Department of Applied Mathematics
The Hong Kong Polytechnic University

Dr Yao Yuan
Associate Professor
Department of Mathematics
The Hong Kong University of Science and Technology

Proxy:  Dr Fan Jun
Assistant Professor, Department of Mathematics, HKBU

In-attendance:  Prof Zhu Lixing
Chair Professor in Statistics, Department of Mathematics, HKBU

Issued by Graduate School, HKBU
Statistical Methods for Integrative Analysis of Genomic Data

MING Jingsi

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Principal Supervisor:
Prof. Zhu Lixing (Hong Kong Baptist University)

August 2018
DECLARATION

I hereby declare that this thesis represents my own work which has been done after registration for the degree of PhD at Hong Kong Baptist University, and has not been previously included in a thesis or dissertation submitted to this or any other institution for a degree, diploma or other qualifications.

I have read the University’s current research ethics guidelines, and accept responsibility for the conduct of the procedures in accordance with the University’s Committee on the Use of Human & Animal Subjects in Teaching and Research (HASC). I have attempted to identify all the risks related to this research that may arise in conducting this research, obtained the relevant ethical and/or safety approval (where applicable), and acknowledged my obligations and the rights of the participants.

Signature: Ming Jingzi
Date: August 2018
Abstract

Thousands of risk variants underlying complex phenotypes (quantitative traits and diseases) have been identified in genome-wide association studies (GWAS). However, there are still several challenges towards deepening our understanding of the genetic architectures of complex phenotypes. First, the majority of GWAS hits are in non-coding region and their biological interpretation is still unclear. Second, most complex traits are suggested to be highly polygenic, i.e., they are affected by a vast number of risk variants with individually small or moderate effects, whereas a large proportion of risk variants with small effects remain unknown. Third, accumulating evidence from GWAS suggests the pervasiveness of pleiotropy, a phenomenon that some genetic variants can be associated with multiple traits, but there is a lack of unified framework which is scalable to reveal relationship among a large number of traits and prioritize genetic variants simultaneously with functional annotations integrated. In this thesis, we propose two statistical methods to address these challenges using integrative analysis of summary statistics from GWASs and functional annotations.

In the first part, we propose a latent sparse mixed model (LSMM) to integrate functional annotations with GWAS data. Not only does it increase the statistical power of identifying risk variants, but also offers more biological insights by detecting relevant functional annotations. To allow LSMM scalable to millions of variants and hundreds of functional annotations, we developed an efficient variational expectation-maximization (EM) algorithm for model parameter estimation and statistical inference. We first conducted comprehensive simulation studies to evaluate the performance of LSMM. Then we applied it to analyze 30 GWASs of complex phenotypes integrated with nine genic category annotations and 127 cell-type specific functional annotations from the Roadmap project. The results demonstrate that our method possesses more statistical power than conventional methods, and can help researchers achieve deeper understanding of genetic architecture of these complex phenotypes.

In the second part, we propose a latent probit model (LPM) which combines summary statistics from multiple GWASs and functional annotations, to characterize relationship and increase statistical power to identify risk variants. LPM can also perform hypothesis testing for pleiotropy and annotations enrichment. To enable the scalability of LPM as the number of GWASs increases, we developed an efficient parameter-expanded EM (PX-EM) algorithm which can execute parallelly. We first validated the performance of LPM through comprehensive simulations, then applied it to analyze 44 GWASs with nine genic category annotations. The results demonstrate the benefits of LPM and can offer new insights of disease etiology.
Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisors Prof. ZHU Lixing and Dr. YANG Can for all of their tremendous academic support and inspiring guidance during my PhD study in Hong Kong Baptist University. I learned a lot from them in the field of statistics and biostatistics. Their broad knowledge, passion of research extraordinarily inspire me to keep studying and improving myself.

I would like to thank the members of our group, for their invaluable advices, collaboration, encouragement and our friendships. Moreover, I’m very grateful to all my friends in the mathematics department for all the great times we have shared in the last three years.

Last but not the least, I would like to express my deepest gratitude to my beloved family, my parents and my husband Hang, for supporting me for everything and encouraging me all the time.
Table of Contents

Declaration i

Abstract ii

Acknowledgements iii

Table of Contents iv

List of Figures v

List of Tables vi

Chapter 1 Introduction 1

1.1 Complex Traits are Highly Polygenic . . . . . . . . . . . . . . . . . . 1
1.2 Integrative Analysis of GWAS with Functional Annotations . . . . . . 2
1.3 Integrative Analysis of Multiple GWASs . . . . . . . . . . . . . . . . 3
1.4 Major Contributions . . . . . . . . . . . . . . . . . . . . . . . . . . . 5
1.5 Outline of the Thesis . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

Chapter 2 LSMM: A statistical approach to integrating functional annotations

with genome-wide association studies 7

2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7
2.2 Model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8
2.3 Algorithm . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
   2.3.1 The Variational EM Algorithm . . . . . . . . . . . . . . . . . . 10
   2.3.2 Details of the Four-stage Algorithm . . . . . . . . . . . . . . . 17
2.4 Inference . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29
# List of Figures

2.1 Performance of LSMM, LFM and TGM for identification of risk SNPs with $\alpha = 0.2$. ........................................ 31  
2.2 Performance of LSMM, LFM and TGM for identification of risk SNPs with $\alpha = 0.4$. ........................................ 32  
2.3 Performance of LSMM, LFM and TGM for identification of risk SNPs with $\alpha = 0.6$. ........................................ 32  
2.4 FDR of LSMM and LSMM (treat the effects of all covariates as fixed effects) for identification of risk SNPs. ........................................ 33  
2.5 Performance of LSMM for detection of relevant annotations with $\alpha = 0.2$. ........................................ 34  
2.6 Performance of LSMM for detection of relevant annotations with $\alpha = 0.4$. ........................................ 34  
2.7 Performance of LSMM for detection of relevant annotations with $\alpha = 0.6$. ........................................ 35  
2.8 Performance of LSMM for detection of relevant annotations when $M$ is large. ........................................ 35  
2.9 Performance in estimation of parameter $\alpha$. ........................................ 36  
2.10 Performance in estimation of parameter $b_0$. ........................................ 37  
2.11 Performance in estimation of parameter $b_1$. ........................................ 37  
2.12 Performance in estimation of parameter $b_2$. ........................................ 38  
2.13 Performance in estimation of parameter $b_3$. ........................................ 38  
2.14 Performance in estimation of parameter $b_4$. ........................................ 39  
2.15 Performance in estimation of parameter $b_5$. ........................................ 39  
2.16 Performance in estimation of parameter $b_6$. ........................................ 40  
2.17 Performance in estimation of parameter $b_7$. ........................................ 40  
2.18 Performance in estimation of parameter $b_8$. ........................................ 41  
2.19 Performance in estimation of parameter $b_9$. ........................................ 41
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.20</td>
<td>Performance in estimation of parameter $b_{10}$</td>
<td>42</td>
</tr>
<tr>
<td>2.21</td>
<td>Performance in estimation of parameter $\omega$</td>
<td>42</td>
</tr>
<tr>
<td>2.22</td>
<td>Parameter estimation ($\hat{\pi}_1$ v.s. true $\pi_1$) and FDR for identification of risk SNPs using LSMM (default setting and random setting).</td>
<td>44</td>
</tr>
<tr>
<td>2.23</td>
<td>Computational time of LSMM.</td>
<td>44</td>
</tr>
<tr>
<td>2.24</td>
<td>Performance of LSMM for identification of risk SNPs based on probit model.</td>
<td>46</td>
</tr>
<tr>
<td>2.25</td>
<td>Performance of LSMM for detection of relevant annotations based on probit model.</td>
<td>46</td>
</tr>
<tr>
<td>2.26</td>
<td>FDR of LSMM and LSMM without fixed effects for detection of relevant annotations.</td>
<td>47</td>
</tr>
<tr>
<td>2.27</td>
<td>The estimation of parameter $\alpha$ using individual-level data.</td>
<td>48</td>
</tr>
<tr>
<td>2.28</td>
<td>FDR of LSMM, LFM and TGM when $P$-values are not from beta distribution.</td>
<td>50</td>
</tr>
<tr>
<td>2.29</td>
<td>FDR of LSMM for identification of risk SNPs with different distance thresholds.</td>
<td>51</td>
</tr>
<tr>
<td>2.30</td>
<td>Parameter estimation ($\hat{\pi}_1$ v.s. true $\pi_1$) using LSMM and Higher Criticism and FDR for identification of risk SNPs using LSMM.</td>
<td>53</td>
</tr>
<tr>
<td>2.31</td>
<td>Performance of LSMM, LFM and TGM for identification of risk SNPs based on probit model.</td>
<td>54</td>
</tr>
<tr>
<td>2.32</td>
<td>Performance of LSMM for detection of relevant annotations based on probit model.</td>
<td>55</td>
</tr>
<tr>
<td>2.33</td>
<td>Performance of LSMM, LFM and TGM for identification of risk SNPs when random effects are correlated.</td>
<td>55</td>
</tr>
<tr>
<td>2.34</td>
<td>Performance of LSMM for detection of relevant annotations when random effects are correlated.</td>
<td>55</td>
</tr>
<tr>
<td>2.35</td>
<td>Performance of LSMM, LFM and TGM for identification of risk SNPs when the variance of random effects are different.</td>
<td>56</td>
</tr>
<tr>
<td>2.36</td>
<td>Performance of LSMM, LFM and TGM for identification of risk SNPs when the variance of random effects are different.</td>
<td>56</td>
</tr>
<tr>
<td>Section</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>2.37</td>
<td>Performance of LSMM, LSMM without fixed effects and GPA for identification of risk SNPs.</td>
<td></td>
</tr>
<tr>
<td>2.38</td>
<td>Performance of LSMM and cmfdr for identification of risk SNPs.</td>
<td></td>
</tr>
<tr>
<td>2.39</td>
<td>Performance of LSMM, LFM and GenoWAP for identification of risk SNPs.</td>
<td></td>
</tr>
<tr>
<td>2.40</td>
<td>Performance of LSMM and GenoWAP for identification of risk SNPs.</td>
<td></td>
</tr>
<tr>
<td>2.41</td>
<td>The number of risk variants identified by TGM, LFM and LSMM for 30 GWAS.</td>
<td></td>
</tr>
<tr>
<td>2.42</td>
<td>Relevant cell-type specific functional annotations for 30 GWASs.</td>
<td></td>
</tr>
<tr>
<td>2.43</td>
<td>Manhattan plots of Schizophrenia1-4 using TGM and LSMM.</td>
<td></td>
</tr>
<tr>
<td>2.44</td>
<td>Computational time using LSMM and cmfdr for 30 GWASs.</td>
<td></td>
</tr>
<tr>
<td>2.45</td>
<td>Relevant functional annotations for 30 GWASs without integrating genic category annotations.</td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>The true correlation graph among eight traits.</td>
<td></td>
</tr>
<tr>
<td>3.2</td>
<td>The average estimated correlation graph using LPM.</td>
<td></td>
</tr>
<tr>
<td>3.3</td>
<td>Relationship test graph of LPM, GPA and graph-GPA.</td>
<td></td>
</tr>
<tr>
<td>3.4</td>
<td>FDR and AUC of LPM and GPA for identification of risk SNPs for P1 and P2.</td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td>FDR and AUC of LPM and GPA for identification of risk SNPs for P3, P4 and P5.</td>
<td></td>
</tr>
<tr>
<td>3.6</td>
<td>FDR and AUC of LPM and GPA for identification of risk SNPs for P6, P7 and P8.</td>
<td></td>
</tr>
<tr>
<td>3.7</td>
<td>FDR and AUC of LPM and GPA for identification of risk SNPs for both P1 and P4.</td>
<td></td>
</tr>
<tr>
<td>3.8</td>
<td>FDR and AUC of LPM and GPA for identification of risk SNPs for three specific traits.</td>
<td></td>
</tr>
<tr>
<td>3.9</td>
<td>Parameter estimation performance of LPM.</td>
<td></td>
</tr>
<tr>
<td>3.10</td>
<td>Comparison of the parameter estimation of $\alpha$ using LPM and $\tilde{\alpha}$ using bLPM.</td>
<td></td>
</tr>
<tr>
<td>3.11</td>
<td>Comparison of the parameter estimation of $\beta$ using LPM and $\tilde{\beta}$ using bLPM for P1.</td>
<td></td>
</tr>
</tbody>
</table>
3.12 Comparison of the parameter estimation of $\beta$ using LPM and $\tilde{\beta}$ using bLPM for P2. ........................................ 103
3.13 Comparison of the parameter estimation of $\beta$ using LPM and $\tilde{\beta}$ using bLPM for P3. ........................................ 103
3.14 Comparison of the parameter estimation of $\beta$ using LPM and $\tilde{\beta}$ using bLPM for P4. ........................................ 104
3.15 Comparison of the parameter estimation of $\beta$ using LPM and $\tilde{\beta}$ using bLPM for P5. ........................................ 104
3.16 Comparison of the parameter estimation of $\beta$ using LPM and $\tilde{\beta}$ using bLPM for P6. ........................................ 104
3.17 Comparison of the parameter estimation of $\beta$ using LPM and $\tilde{\beta}$ using bLPM for P7. ........................................ 105
3.18 Comparison of the parameter estimation of $\beta$ using LPM and $\tilde{\beta}$ using bLPM for P8. ........................................ 105
3.19 Comparison of the parameter estimation of $R$ using LPM and $\tilde{R}$ using bLPM. ........................................ 106
3.20 Relationship test graph of LPM, GPA and graph-GPA when there is no functional annotation. ......................... 107
3.21 FDR and AUC of LPM, GPA and graph-GPA for identification of risk SNPs for P1 and P2 when there is no functional annotations. .......... 108
3.22 FDR and AUC of LPM, GPA and graph-GPA for identification of risk SNPs for P3, P4 and P5 when there is no functional annotations. ... 109
3.23 FDR and AUC of LPM, GPA and graph-GPA for identification of risk SNPs for P6, P7 and P8 when there is no functional annotations. ... 110
3.24 Type I error rate and power of LPM for the relationship test between two traits. ........................................ 111
3.25 FDR and AUC of LPM for identification of risk SNPs for the target trait when correlated trait is integrated. ..................... 112
3.26 Type I error rate and power of LPM for the annotation enrichment test. ........................................ 113
3.27 Computational time of LPM. ........................................ 113
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.28</td>
<td>Type I error rate of LPM for the relationship test between the two traits when ( P )-values are simulated from individual-level data</td>
<td>115</td>
</tr>
<tr>
<td>3.29</td>
<td>FDR of LPM for identification of risk SNPs for P1, P2, P1 and P2 when ( P )-values are simulated from individual-level data</td>
<td>116</td>
</tr>
<tr>
<td>3.30</td>
<td>Type I error rate of LPM for the relationship test between the two traits when ( P )-values are not from beta distribution</td>
<td>117</td>
</tr>
<tr>
<td>3.31</td>
<td>FDR of LPM for identification of risk SNPs for P1, P2, P1 and P2 when ( P )-values are not from beta distribution</td>
<td>118</td>
</tr>
<tr>
<td>3.32</td>
<td>The estimated ( \hat{R} ) for 44 GWASs with 9 genic category annotation integrated</td>
<td>120</td>
</tr>
<tr>
<td>3.33</td>
<td>The number of variants identified to be associated with each of the 44 traits using LPM</td>
<td>121</td>
</tr>
<tr>
<td>3.34</td>
<td>The estimated coefficients of downstream, exonic, intergenic and intronic for 44 GWASs</td>
<td>122</td>
</tr>
<tr>
<td>3.35</td>
<td>The estimated coefficients of ncRNA_exonic, ncRNA_intronic, upstream and UTR'3 for 44 GWASs</td>
<td>123</td>
</tr>
<tr>
<td>3.36</td>
<td>The estimated coefficients of UTR’5 for 44 GWASs</td>
<td>124</td>
</tr>
<tr>
<td>3.37</td>
<td>The estimated ( \hat{R} ) for 44 GWAS with no annotation integrated</td>
<td>125</td>
</tr>
<tr>
<td>3.38</td>
<td>Relationship inferred by graph-GPA when different numbers of GWASs are considered</td>
<td>126</td>
</tr>
<tr>
<td>3.39</td>
<td>Manhattan plots of LDL and TC</td>
<td>127</td>
</tr>
</tbody>
</table>
# List of Tables

2.1 Alternative distributions for $z$-scores. ........................................ 49  
2.2 The source of the 30 GWASs. .................................................. 61  
2.3 Summary of results for Schizophrenia. ........................................ 65  
2.4 The number of risk SNPs identified by GenoWAP, TGM, LFM and LSMM for Schizophrenia. .................................................. 66  
3.1 Alternative distributions for $z$-scores. ........................................ 115  
3.2 The source of the 44 GWASs. .................................................. 119  
3.3 The top two correlated traits of the 44 GWASs. ............................ 128  
3.4 The number of traits with significant coefficients for each genic category annotation at different significance level. ................................. 129  
3.5 The estimated correlations among four schizophrenia GWASs. .......... 129  
3.6 The estimated coefficients of annotations for four schizophrenia GWASs. 130  
3.7 Sample sizes for four schizophrenia GWASs. ............................... 130
Chapter 1

Introduction

Genome-wide association studies (GWASs) were designed to detect associations between genetic variants and complex traits from populations. Since the success of the first GWAS on age-related macular degeneration [Klein et al., 2005], more than 40,000 single-nucleotide polymorphisms (SNPs) have been reported to be associated with one or more complex traits in about 3,300 GWAS at the genome-wide significance level (see GWAS Catalog http://www.ebi.ac.uk/gwas/) [Welter et al., 2014]. Despite these fruitful discoveries, the emerging evidence from GWAS presents great challenges towards deeper understanding of the genetic architectures of complex phenotypes.

1.1 Complex Traits are Highly Polygenic

Complex phenotypes are often highly polygenic, i.e., they are affected by a vast number of risk variants with individually small effects. For example, it is widely accepted that 70%-80% of the variation in human height can be attributed to genetics [Visscher et al., 2008]. However, Wood et al. [2014] collected more than 250,000 samples and identified 697 variants at genome-wide significance level, and all these variants together can only explain 20% of heritability. A recent estimate [Boyle et al., 2017] suggests that about 100,000 variants may be associated with human height. Given current sample sizes, a large proportion of risk variants underlying complex phenotypes remain unknown yet. It motivates us to consider integrative
analysis which combines indirect evidence of relevant information beyond GWAS and direct evidence of GWAS, in order to increase the statistical power of identifying risk SNPs. The indirect evidence we discuss here includes functional annotations and other GWASs which are obtained from different data sources or based on other correlated traits.

1.2 Integrative Analysis of GWAS with Functional Annotations

More than 85% genome-wide significant hits are located in the non-coding region \cite{Welter2014}. However, most of their functional roles are still elusive. Fortunately, an increasing number of reports suggest that the functional importance of SNPs may not be equal \cite{Schork2013}, which provides a direction to address the above challenge. On the one hand, SNPs in or near genic regions can explain more heritability of complex phenotypes \cite{Smith2011, Yang2011}. For example, the partition of genic category annotations for SNPs has revealed that SNPs in 5’UTR, exon and 3’UTR are significantly enriched, SNPs in introns are moderately enriched and intergenic SNPs are negatively enriched across diverse complex traits \cite{Schork2013}. It is also coincidence with the finding that pleiotropic SNPs are more often exonic and less often intergenic compared with non-pleiotropic SNPs \cite{Sivakumaran2011}. On the other hand, cell-type specific functional annotations can provide information that is complementary to genic category annotations, for dissecting genetic contribution to complex diseases in a cell-type specific manner. To name a few, genetic variants related to functions of immune cells are significantly enriched for immune diseases, such as rheumatoid arthritis, coeliac disease and type 1 diabetes; variants with liver functions are enriched for metabolic traits, such as LDL, HDL and total cholesterol; variants with pancreatic islet functions are enriched for fasting glucose \cite{Kundaje2015}. Additionally, SNPs in genes that are preferentially expressed in the central nervous system are significantly enriched in psychiatric disorders (e.g., schizophrenia and bipolar disorder) \cite{Chung2014}.

A large amount of functional annotation data has become publicly available and
the volume is still expanding. The Encyclopedia of DNA Elements (ENCODE) project [The ENCODE Project Consortium, 2012] has conducted more than 1,650 experiments on 147 cell lines to delineate functional elements across the human genome, such as DNase I hypersensitive sites and transcription factor binding. The NIH Roadmap Epigenomics Mapping Consortium [Kundaje et al., 2015] is generating high-quality genome-wide human epigenomic maps of histone modifications, chromatin accessibility, DNA methylation and mRNA expression across more than one hundred of human cell types and tissues.

Statistical methods to incorporate functional annotations with GWAS can be roughly divided into two categories: methods based on individual-level genotype data, e.g., bfGWAS [Yang et al., 2017] and FST [He et al., 2017], and methods based on summary statistics. In practice, the access to the individual-level data of a large sample size is often very difficult. Therefore, methods in the second category play the major role and also have computational advantages. However, existing methods based on summary statistics, e.g., stratified FDR methods [Schork et al., 2013], cmfdr [Zablocki et al., 2014], GPA [Chung et al., 2014], GenoWAP [Lu et al., 2016] and EPS [Liu et al., 2016], were designed to handle only a few number of functional annotations and can not be scalable to a large-scale integrative analysis. For example, GPA can integrate GWAS datasets with functional annotations to prioritizing risk variants and perform hypothesis testing for the enrichment of functional annotations. However, GPA assumes conditional independence among annotations, which is often not the case in real data of cell-type specific functional annotations. cmfdr is a fully Bayesian approach to incorporate genic category annotations in GWAS using MCMC sampling algorithm. As a result, cmfdr is not able to handle a large number of annotations and the MCMC sampling algorithm is very time-consuming. GenoWAP is a GWAS signal prioritization method that integrates genomic functional annotation and GWAS test statistics. However, GenoWAP is only designed to integrate one annotation at a time.
1.3 Integrative Analysis of Multiple GWASs

By exploring the fruitful findings of GWASs, a phenomenon that some genetic variants can be associated with multiple traits, which is called pleiotropy, was discovered. And accumulating studies suggest the pervasiveness of pleiotropy. The analysis of pleiotropy can be divided into two main directions. One is to identify the underlying causal SNPs for two or more traits. For example, a nonsynonymous variant in the zinc transporter SLC39A8 influences both schizophrenia and Parkinson disease (Pickrell et al., 2016). Ellinghaus et al. identified 244 independent multi-disease loci in an analysis of five chronic inflammatory diseases. Overall, 16.9% genes and 4.6% SNPs are reported to have pleiotropic effects [Sivakumaran et al., 2011]. The other direction is to measure the genetic correlation among traits. Substantial genetic correlation has been revealed among psychiatric disorders, such as the high correlation between schizophrenia and bipolar, moderate correlation between schizophrenia and major depressive disorder, bipolar disorder and major depressive disorder, and attention-deficit/hyperactivity disorder and major depressive disorder [Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013a]. For autoimmune diseases, primary sclerosing cholangitis and ulcerative colitis, as well as ulcerative colitis and Crohn’s disease, are suggested to have a relatively high genome-wide genetic correlation [Ji et al., 2016].

The study of pleiotropy can facilitate our understanding of the genetic architecture of complex traits. Moreover, the existence of widespread pleiotropy provides a direction to boost statistical power of identifying risk variants for one specific trait by jointly analyzing multiple GWASs. To name a few, joint analysis of schizophrenia and bipolar disorder could significantly improve association mapping power for each of the diseases [Chung et al., 2014]. The power to detect associated variants for systolic blood pressure was increased by considering GWASs of other comorbid phenotypes, such as low-density lipoprotein, body mass index and type 1 diabetes mellitus [Andreassen et al., 2014].

Some existing statistical methods based on summary statistics, such as cross-trait LD Score regression [Bulik-Sullivan et al., 2015] and GNOVA [Lu et al., 2017a], provide genetic correlation estimation for pair of traits, but are not able to prioritize...
GWAS results. Other methods such as GPA [Chung et al., 2014] and graph-GPA [Chung et al., 2017], can both infer the relationship among traits and identify causal variants. However, important statistical and computational challenges emerge as the number of traits increases. GPA assumes a four-groups model for the case of two GWASs. With increasing number of traits $K$, the number of groups will increase exponentially ($2^K$). Graph-GPA is not able to integrate functional annotations and its implement is based on an MCMC algorithm which is time-consuming. Additionally, the pleiotropic results of graph-GPA are hard to interpret in real data analysis because it suggests a graphical model based on a Markov random field which represents a conditional independent structure for genetic relationship among traits.

### 1.4 Major Contributions

For integrative analysis of GWAS with functional annotations, we proposed a latent sparse mixed model (LSMM), which can (i) integrate genic category annotations and cell-type specific functional annotations with GWAS to increase the statistical power of identifying risk SNPs, and (ii) detect relevant cell-type specific functional annotations among a large amount of available annotation data to have a more biologically insightful interpretation of GWAS results. Compared with existing methods, LSMM can be scalable to a large number of functional annotations and achieve a higher power to detect risk SNPs. This work is published in Bioinformatics [Ming et al., 2018]. We developed an efficient R package which is available at https://github.com/mingjingsi/LSMM. All the simulations are reproducible using the codes at https://github.com/mingjingsi/sim-LSMM.

For integrative analysis of multiple GWASs and functional annotations, we proposed a latent probit model (LPM), which is a unified framework to (i) characterize relationship among complex traits, including identifying pleiotropic associations and estimating correlation among traits, (ii) increase the association mapping power for one or more traits, and (iii) investigate the effect of functional annotations. LPM can not only fulfill the three goals simultaneously but also achieve a better performance than conventional methods. The algorithm based on PX-EM algorithm for pair of
GWASs is efficient and can be implemented parallelly. This work is in submission. R package for LPM is available at https://github.com/mingjingsi/LPM. Codes for simulations are available at https://github.com/mingjingsi/sim-LPM.

1.5 Outline of the Thesis

In Chapter 2, we introduce the latent sparse mixed model (LSMM) for integrative analysis of GWAS with functional annotations. We used a mixed model to integrate both genic category annotations and cell-type specific functional annotations with GWAS data. After the description of model, algorithm and method of inference, we evaluated the performance of LSMM in the identification of risk SNPs and detection of relevant cell-type specific functional annotations through comprehensive simulations and then applied LSMM for real data analysis to obtain new insights of genetic architecture of complex phenotypes.

In Chapter 3, we introduce the latent probit model (LPM) for integrative analysis of multiple GWASs with functional annotations. First, we described the details of model, algorithm, inference method and related theorem. Then we evaluated the performance of LPM in the identification of risk SNPs, relationship test among traits and annotation enrichment test through comprehensive simulations. Finally, we analyzed the results of real datasets obtained using LPM to provide better understanding of disease etiology.

In Chapter 4, we make conclusions of the results we obtained. We also provide some extensions about our methods and propose some directions for future works.
Chapter 2

LSMM: A statistical approach to integrating functional annotations with genome-wide association studies

2.1 Introduction

In this chapter, we propose a Latent Sparse Mixed Model (LSMM) to integrate genic category annotations and cell-type specific functional annotations with GWAS data. The “latent” statuses are used to connect the observed summary statistics from GWAS with functional annotations. “Mixed” models are designed to simultaneously consider both genic category and cell-type specific annotations, where genic category annotations are put into the design matrix of fixed effects, and cell-type specific annotations are encoded in the design matrix of random effects. We further impose a “sparse” structure on the random effects to adaptively select relevant cell-type specific annotations. LSMM can not only increase the statistical power of identifying risk SNPs but also detect relevant cell-type specific functional annotations among a large amount of available annotation data.

We first described the model in Section 2.2 introduced the efficient variational EM algorithm which enables LSMM scalable to millions of variants and hundreds of
functional annotations in Section 2.3, and methods of statistical inference in Section 2.4. Then we conducted comprehensive simulations to investigate the properties of LSMM in Section 2.5. We also applied LSMM to real data for integrative analysis of 30 GWASs with 9 genic category annotations and 127 cell-type specific functional annotations from the Roadmap project in Section 2.6.

2.2 Model

Suppose we have the summary statistics (P-values) of M SNPs from GWAS. Consider the two-groups model [Efron, 2008], i.e., SNPs either belong to null or non-null group. Let $\gamma_j$ be the latent variable indicating the membership of the $j$-th SNP, i.e., $\gamma_j = 0$ or $\gamma_j = 1$ indicates the $j$-th SNP from null or non-null group, respectively. The proportion of null and non-null group are denoted as $\pi_0$ and $\pi_1$, respectively. Then we model the observed $P$-values as [Chung et al., 2014],

$$
 p_j \sim \begin{cases} 
 U[0,1], & \gamma_j = 0, \\
 \text{Beta}(\alpha,1), & \gamma_j = 1,
\end{cases}
$$

(2.1)

where $U[0,1]$ denotes the uniform distribution on $[0,1]$ and $\text{Beta}(\alpha,1)$ is the beta distribution with parameter $(\alpha,1)$. We constrain $0 < \alpha < 1$ to model the fact that $P$-values from the non-null group tend to be closer to 0 rather than 1.

Suppose that we have collected not only the $P$-values of $M$ SNPs from GWAS, but also functional annotations of these SNPs. To incorporate information from functional annotations for prioritization of risk variants and detection of cell-type specific functions for a complex phenotype, we consider the following latent sparse mixed model:

$$
 \log \frac{\Pr(\gamma_j = 1|Z_j, A_j)}{\Pr(\gamma_j = 0|Z_j, A_j)} = Z_j b + A_j \beta,
$$

(2.2)

where $Z \in \mathbb{R}^{M \times (L+1)}$ is the design matrix for fixed effects, comprised of an intercept and $L$ covariates, $b \in \mathbb{R}^{L+1}$ is the vector of fixed effects, $A \in \mathbb{R}^{M \times K}$ is the design matrix for random effects, $\beta \in \mathbb{R}^{K}$ is the vector of random effects, and $K$ is the number of random effects. Both the $j$-th row of $Z$ (i.e., $Z_j$) and $A$ (i.e., $A_j$) corresponds to
the $j$-th SNP. Note that $\gamma_j$ is a latent variable in model (2.2) but its corresponding $p_j$ is observed. This makes our model different from the standard generalized linear mixed model.

Now we partition functional annotations into two categories: genic category annotations and cell-type specific annotations. According to [Schork et al., 2013], genomic regions, such as exon, intron, 5'UTR and 3'UTR, are considered as genic category annotations. For cell-type specific annotations, we used epigenetic markers (H3k4me1, H3k4me3, H3k36me3, H3k27me3, H3k9me3, H3k27ac, H3k9ac, and DNase I Hypersensitivity) of multiple tissues from the Roadmap project. As we are more interested in the detection of cell-type specific results, we put genic category annotation data into $Z$ and cell-type specific annotation data into $A$, where each column of $Z$ corresponds to a genic functional category and each column of $A$ corresponds to a cell-type specific functional category. In the simplest case, the entries in $Z$ and $A$ are binary. For example, $Z_{jl} = 1$ means that the $j$-th SNP has a function in the $l$-th genic category and $Z_{jl} = 0$ otherwise. Our model also allows the entries in $Z$ and $A$ to be continuous variables, e.g., a score $Z_{jl}$ between 0 and 1 can be used to indicate the degree that the $j$-the SNP has a function in the $l$-th category. The closer to 1, the more likely it has a functional role. The entries in $A$ are defined in the same way as those of $Z$.

To adaptively select cell-type specific annotations, we assign a spike-slab prior on $\beta_k$:

$$
\beta_k \sim \begin{cases} 
N(\beta_k|0,\sigma^2), & \eta_k = 1, \\
\delta_0(\beta_k), & \eta_k = 0,
\end{cases}
$$

(2.3)

where $N(\beta_k|0,\sigma^2)$ denotes the Gaussian distribution with mean 0 and variance $\sigma^2$, $\delta_0$ denotes the Dirac delta function at zero, $\eta_k = 1$ or $\eta_k = 0$ means the $k$-th annotation is relevant or irrelevant to the given phenotype, respectively. Here $\eta_k$ is a Bernoulli variable with probability $\omega$ being 1:

$$
\eta_k \sim \omega^{\eta_k}(1-\omega)^{1-\eta_k},
$$

where $\omega$ can be interpreted as the proportion of relevant annotations corresponding
to this phenotype.

Let $\theta = \{\alpha, b, \sigma^2, \omega\}$ be the collection of model parameters. The logarithm of the marginal likelihood can be written as

$$\log \Pr (p|Z, A; \theta) = \log \sum_{\gamma} \sum_{\eta} \int \Pr (p, \gamma, \beta, \eta|Z, A; \theta) d\beta,$$  \tag{2.4}

where

$$\Pr (p, \gamma, \beta, \eta|Z, A; \theta) = \Pr (p|\gamma; \alpha) \Pr (\gamma|Z, A, \beta; b) \Pr (\beta|\eta; \sigma^2) \Pr (\eta|\omega).$$

Our goal is to maximize the marginal likelihood to obtain the estimation $\hat{\theta}$ of $\theta$ and compute the posterior

$$\Pr (\gamma, \beta, \eta|p, Z, A; \hat{\theta}) = \frac{\Pr (p, \gamma, \beta, \eta|Z, A; \hat{\theta})}{\Pr (p|Z, A; \hat{\theta})}. \tag{2.5}$$

Then we can infer the risk SNPs and relevant cell-type specific functional annotations for this phenotype and calculate the false discovery rate.

### 2.3 Algorithm

#### 2.3.1 The Variational EM Algorithm

Exact evaluation of posterior (2.5) is intractable. One difficulty is due to the sigmoid function resulting from the logistic model. The other comes from the spike-slab prior. To address this issue, we propose a variational EM algorithm for parameter estimation and posterior approximation.

Before starting the derivation of our algorithm, we first reparametrize the spike-slab prior (2.3) by introducing a new Gaussian variable $\tilde{\beta}_k \sim N(0, \sigma^2)$, then the product $\eta_k \tilde{\beta}_k$ has the same distribution with $\beta_k$ in model (2.3). So model (2.2) can be written as

$$\log \frac{\Pr (\gamma_j = 1|Z_j, A_j)}{\Pr (\gamma_j = 0|Z_j, A_j)} = Z_j b + \sum_{k=1}^{K} A_j k \tilde{\beta}_k = Z_j b + \sum_{k=1}^{K} A_j k \eta_k \tilde{\beta}_k.$$
Hence the complete-data likelihood \( Pr(p, \gamma, \beta, \eta | Z, A; \theta) \) can be re-written as

\[
Pr (p, \gamma, \tilde{\beta}, \eta | Z, A; \theta) = Pr (p | \gamma; \alpha) Pr (\gamma | Z, A, \tilde{\beta}, \eta; b) Pr (\tilde{\beta}, \eta | \sigma^2, \omega),
\]

where

\[
Pr (p | \gamma; \alpha) = \prod_{j=1}^{M} Pr (p_j | \gamma_j; \alpha) = \prod_{j=1}^{M} (\alpha p_j^{\alpha-1})^{\gamma_j},
\]

\[
Pr (\gamma | Z, A, \tilde{\beta}, \eta; b) = \prod_{j=1}^{M} Pr (\gamma_j | Z_j, A_j, \tilde{\beta}, \eta; b)
\]

\[
= \prod_{j=1}^{M} e^{\gamma_j (Z_j b + \sum_k A_{jk} \eta_k \tilde{\beta}_k) S (-Z_j b - \sum_{k=1}^{K} A_{jk} \eta_k \tilde{\beta}_k)}, \tag{2.6}
\]

\[
Pr (\tilde{\beta}, \eta | \sigma^2, \omega) = Pr (\tilde{\beta} | \sigma^2) Pr (\eta | \omega) = \prod_{k=1}^{K} N (\tilde{\beta}_k | 0, \sigma^2) \omega^{\eta_k} (1 - \omega)^{1-\eta_k},
\]

where \( S (\cdot) \) is the sigmoid function and \( S (x) = (1 + e^{-x})^{-1} \). With this reparameterization, we get rid of the Dirac delta function.

Due to the intractability caused by the sigmoid function inside integration (2.4), we consider the JJ bound \cite{Jaakkola and Jordan 2000}:

\[
S (x) \geq S (\xi) \exp \left\{ (x - \xi) / 2 - \lambda (\xi) (x^2 - \xi^2) \right\}, \tag{2.7}
\]

where \( \lambda (\xi) = \frac{1}{2\xi} \left[ S (\xi) - \frac{1}{2} \right] \) and the right-hand-side of the inequality (2.7) is the JJ bound. Clearly, the JJ bound is in the exponential of a quadratic form. Applying this bound to (2.6), we can get a tractable lower bound of \( Pr (\gamma | Z, A, \tilde{\beta}, \eta; b) \), denoted
as \( h(\gamma|Z, A, \tilde{\beta}, \eta; b, \xi) \):

\[
\Pr(\gamma_j|Z_j, A_j, \tilde{\beta}, \eta; b) = e^{\gamma_j(z_jb + \sum_k A_{jk}\eta_k\tilde{\beta}_k)} S(-z_jb - \sum_k A_{jk}\eta_k\tilde{\beta}_k) \\
\geq e^{\gamma_j(z_jb + \sum_k A_{jk}\eta_k\tilde{\beta}_k)} S(\xi_j) \exp(-\lambda(\xi_j) \left( (z_jb + \sum_k A_{jk}\eta_k\tilde{\beta}_k)^2 - \xi_j^2 \right)) \\
+ e^{\gamma_j(z_jb + \sum_k A_{jk}\eta_k\tilde{\beta}_k)} S(\xi_j) \exp(-z_jb + \sum_k A_{jk}\eta_k\tilde{\beta}_k + \xi_j) \\
=h(\gamma_j|Z_j, A_j, \tilde{\beta}, \eta; b, \xi_j),
\]

where \( \xi \in \mathbb{R}^M \) is variational parameter and

\[
\lambda(\xi_j) = \frac{1}{2\xi_j} \left( S(\xi_j) - \frac{1}{2} \right).
\]

Let \( \Theta = \{\alpha, b, \xi, \sigma^2, \omega\} \). The lower bound of the complete-data likelihood is defined as

\[
f(p, \gamma, \tilde{\beta}, \eta|Z, A; \Theta) = \Pr(p|\gamma; \alpha) h(\gamma|Z, A, \tilde{\beta}, \eta; b, \xi) \Pr(\tilde{\beta}, \eta|\sigma^2, \omega).
\]

Next we derive the variational EM algorithm. Let \( q(\gamma, \tilde{\beta}, \eta) \) be an approximation of the posterior \( \Pr(\gamma, \tilde{\beta}, \eta|p, Z, A; \theta) \). We can obtain a lower bound of the logarithm
of the marginal likelihood

\[ \log \Pr (p|Z, A; \theta) \]
\[ = \log \sum_{\gamma} \sum_{\eta} \int \Pr (p, \gamma, \tilde{\beta}, \eta|Z, A; \theta) \, d\tilde{\beta} \]
\[ \geq \log \sum_{\gamma} \sum_{\eta} \int f (p, \gamma, \tilde{\beta}, \eta|Z, A; \Theta) \, d\tilde{\beta} \]
\[ \geq \sum_{\gamma} \sum_{\eta} \int q (\gamma, \tilde{\beta}, \eta) \log \frac{f (p, \gamma, \tilde{\beta}, \eta|Z, A; \Theta)}{q (\gamma, \tilde{\beta}, \eta)} \, d\tilde{\beta} \]
\[ = \mathbb{E}_q \left[ \log f (p, \gamma, \tilde{\beta}, \eta|Z, A; \Theta) - \log q (\gamma, \tilde{\beta}, \eta) \right] \]
\[ \triangleq L (q), \]

where \( L(q) \) is the lower bound. The first inequality is based on the JJ bound. The second inequality follows Jensen’s inequality. To make it feasible to evaluate the lower bound, we assume that \( q (\gamma, \tilde{\beta}, \eta) \) can be factorized as

\[ q (\gamma, \tilde{\beta}, \eta) = \left( \prod_{k=1}^{K} q (\tilde{\beta}_k, \eta_k) \right) \left( \prod_{j=1}^{M} q (\gamma_j) \right), \]

where \( q (\tilde{\beta}_k, \eta_k) = q (\tilde{\beta}_k|\eta_k) q (\eta_k), q (\gamma_j = 1) = \pi_j, q (\eta_k = 1) = \omega_k. \)

We can obtain an approximation according to the mean-field method:

\[ \log q (\tilde{\beta}_i, \eta_i) = \mathbb{E}_{k \neq i} \mathbb{E}_\gamma \left[ \log f (p, \gamma, \tilde{\beta}, \eta|Z, A; \Theta) \right] \]
\[ = \left( -\frac{1}{2\sigma^2} - \sum_{j=1}^{M} \lambda (\xi_j) A^2_{ji} \eta_i^2 \right) \tilde{\beta}_i^2 + \eta_i \log \omega + (1 - \eta_i) \log (1 - \omega) + \text{const} \]
\[ + \sum_{j=1}^{M} \left( \pi_j - \frac{1}{2} - 2\lambda (\xi_j) Z_j b \right) A_{ji} - 2\lambda (\xi_j) A_{ji} \sum_{k \neq i} A_{jk} E_k \left[ \eta_k \tilde{\beta}_k \right] \eta_i \tilde{\beta}_i, \]

where the expectation is taken under the distribution \( q (\tilde{\beta}_-, \eta_-) = \prod_{k \neq i} q (\tilde{\beta}_k, \eta_k) \) and \( q (\gamma) \).
When $\eta_i = 1$, we have

$$
\log q \left( \tilde{\beta}_i | \eta_i = 1 \right) \\
= \left( -\frac{1}{2\sigma^2} - \sum_{j=1}^{M} \lambda (\xi_j) A_{ji}^2 \right) \tilde{\beta}_i^2 \\
+ \sum_{j=1}^{M} \left( \left( \pi_j - \frac{1}{2} - 2\lambda (\xi_j) Z_j b \right) A_{ji} - 2\lambda (\xi_j) A_{ji} \sum_{k \neq i} A_{jk} E_k \left[ \eta_k \tilde{\beta}_k \right] \right) \tilde{\beta}_i + \text{const},
$$

where $E_k$ denotes the expectation under $q \left( \tilde{\beta}_k, \eta_k \right)$, and the constant doesn’t depend on $\tilde{\beta}_i$. Because $\log q \left( \tilde{\beta}_i | \eta_i = 1 \right)$ is a quadratic form,

$$
q \left( \tilde{\beta}_i | \eta_i = 1 \right) = N \left( \mu_i, s_i^2 \right),
$$

where

$$
\mu_i = s_i^2 \sum_{j=1}^{M} \left( \pi_j - \frac{1}{2} - 2\lambda (\xi_j) \left( Z_j b + \sum_{k \neq i} A_{jk} E_k \left[ \eta_k \tilde{\beta}_k \right] \right) \right) A_{ji},
$$

$$
s_i^2 = \frac{\sigma^2}{1 + 2\sigma^2 \sum_{j=1}^{M} \lambda (\xi_j) A_{ji}^2}.
$$

When $\eta_i = 0$, we have

$$
\log q \left( \tilde{\beta}_i | \eta_i = 0 \right) = -\frac{1}{2\sigma^2} \tilde{\beta}_i^2 + \text{const}.
$$

So

$$
q \left( \tilde{\beta}_i | \eta_i = 0 \right) = N \left( 0, \sigma^2 \right).
$$

Therefore we have

$$
q \left( \tilde{\beta}_i, \eta_i \right) = \left[ \omega_i N \left( \mu_i, s_i^2 \right) \right]^{\eta_i} \left[ \left( 1 - \omega_i \right) N \left( 0, \sigma^2 \right) \right]^{1-\eta_i}.
$$
Now we evaluate the variational lower bound

\[ L(q) = E_q \left[ \log \Pr(p|\gamma, \alpha) + \log h \left( \gamma|Z, A, \bar{\beta}, \eta, \xi, b, \xi \right) + \log \Pr \left( \bar{\beta}, \eta|\sigma^2, \omega \right) \right] \]
\[ - E_q \left[ \log q \left( \gamma, \bar{\beta}, \eta \right) \right]. \]

We set the partial derivative of the lower bound \( L(q) \) w.r.t to \( \omega_k, \pi_j \) and \( \xi_j \) be 0 to get the variational parameters \( \omega_k, \pi_j \) and \( \xi_j \):

\[ \omega_k = \frac{1}{1 + \exp(-u_k)}, \text{ where } u_k = \log \frac{\omega}{1 - \omega} + \frac{1}{2} \log \frac{s_k^2}{\sigma^2} + \frac{\mu_k^2}{2s_k^2}, \]
\[ \pi_j = \frac{1}{1 + \exp(-v_j)}, \text{ where } v_j = \log \alpha + (\alpha - 1) \log p_j + Z_j b + \sum_{k=1}^{K} A_{jk} \omega_k \mu_k, \]
\[ \xi_j^2 = \left( Z_j b + \sum_{k} A_{jk} \omega_k \mu_k \right)^2 + \sum_{k} A_{jk}^2 \left( \omega_k \left( s_k^2 + \mu_k^2 \right) - \omega_k \mu_k^2 \right). \]

Thus the lower bound \( L(q) \) can be exactly evaluated.

\[
L(q) = \sum_{j=1}^{M} \left( \pi_j \left( \log \alpha + (\alpha - 1) \log p_j \right) + \pi_j \left( Z_j b + \sum_{k} A_{jk} \omega_k \mu_k \right) + \log S(\xi_j) \right) \\
+ \sum_{j=1}^{M} \left( -\lambda(\xi_j) \left( \left( \beta_0 + \sum_{k} A_{jk} \omega_k \mu_k \right)^2 - \xi_j^2 \right) - \left( Z_j b + \sum_{k} A_{jk} \omega_k \mu_k + \xi_j \right) / 2 \right) \\
+ \sum_{j=1}^{M} \left( \lambda(\xi_j) \sum_{k} A_{jk}^2 \omega_k \mu_k^2 - \lambda(\xi_j) \sum_{k} A_{jk}^2 \omega_k \left( s_k^2 + \mu_k^2 \right) \right) \\
- \frac{1}{2\sigma^2} \sum_{k=1}^{K} \left( \omega_k \left( s_k^2 + \mu_k^2 \right) - \omega_k \sigma^2 \right) + \sum_{k=1}^{K} \omega_k \log \omega + \sum_{k=1}^{K} \left( 1 - \omega_k \right) \log \left( 1 - \omega \right) \\
+ \sum_{k=1}^{K} \left( \frac{1}{2} \omega_k \left( \log s_k^2 - \log \sigma^2 \right) - \omega_k \log \omega_k - \left( 1 - \omega_k \right) \log \left( 1 - \omega_k \right) \right) \\
- \sum_{j=1}^{M} \left( \pi_j \log \pi_j + \left( 1 - \pi_j \right) \log \left( 1 - \pi_j \right) \right). \]

We can obtain the following updating equations by setting the derivative of \( L(q) \)
with respect to $\alpha$, $\sigma^2$, $\omega$ be zero,

$$\alpha = - \frac{\sum_{j=1}^{M} \pi_j}{\sum_{j=1}^{M} \pi_j \log p_j},$$

$$\sigma^2 = \frac{\sum_{k=1}^{K} \omega_k \left( s_k^2 + \mu_k^2 \right)}{\sum_{k=1}^{K} \omega_k},$$

$$\omega = \frac{1}{K} \sum_{k=1}^{K} \omega_k,$$

and use Newton’s method to update $b$:

$$b = b_{old} - H^{-1}g,$$

where

$$g = \sum_{j=1}^{M} Z_j^T \left( \pi_j - 2\lambda (\xi_j) \left( Z_j b + \sum_k A_{jk} \omega_k \mu_k \right) - \frac{1}{2} \right),$$

$$H = -2Z_j^T \lambda (\xi_j) Z_j.$$

It is worth noting that LSMM covers two special cases: (1) Two-groups model only (denoted as TGM) when all the coefficients in $b$ (except the intercept term) and $\beta$ are zero; (2) Two-groups model plus fixed effects model only (denoted as LFM for the abbreviation of latent fixed effect model) when all coefficients in $\beta$ are zero.

This motivates us developing a four-stage algorithm based on warm starts. More specifically, in the first stage, we run an EM algorithm to obtain the two parameters ($\alpha$ and the proportion of non-null group $\pi_1$) in the TGM. Then we use the estimated parameters as the starting point to run the second stage variational EM algorithm to fit the LFM and obtain the parameter $\alpha$, $b$ and the posterior probability of $\gamma$. In the third stage, we treat the obtained posterior as the value of $\gamma$ and fit the logistic sparse mixed model to obtain the required initial value for the parameters in the next stage. Finally, in the fourth stage, we run the above variational EM algorithm with the obtained parameters at the second and third stage until convergence. Since all the iterations are built upon the framework of EM algorithm, the lower bound is guaranteed to increase at each iteration. The details of the algorithm design are
provided in Section 2.3.2

2.3.2 Details of the Four-stage Algorithm

Stage 1: Two-groups Model (TGM)

Suppose we have the $P$-values of $M$ SNPs for a given phenotype. Let $\gamma_j$ be the latent variable indicating whether the $j$-th SNP is associated with this phenotype. Here $\gamma_j = 0$ means unassociated and $\gamma_j = 1$ means associated. Then we have the following two-groups model:

$$p_j \sim \begin{cases} U [0, 1], & \gamma_j = 0, \\ Beta (\alpha, 1), & \gamma_j = 1, \end{cases}$$

where $p \in \mathbb{R}^M$ are the $P$-values, $0 < \alpha < 1$ and $Pr (\gamma_j = 1) = \pi_1$.

We can use EM algorithm to compute the posterior and parameter estimation.

Let $\theta = \{\alpha, \pi_1\}$ be the collection of model parameters. The logarithm of the marginal likelihood is

$$\log Pr (p|\theta) = \log \sum_{\gamma} Pr (p, \gamma|\theta) = \log \sum_{\gamma} Pr (p|\gamma; \alpha) Pr (\gamma|\pi_1),$$

where

$$Pr (p|\gamma; \alpha) = \prod_{j=1}^{M} Pr (p_j|\gamma_j; \alpha) = \prod_{j=1}^{M} (\alpha p_j^{\alpha-1})^{\gamma_j},$$

$$Pr (\gamma|\pi_1) = \prod_{j=1}^{M} \pi_1^{\gamma_j} (1 - \pi_1)^{1-\gamma_j}.$$ 

In the E step, we compute the posterior:

$$\tilde{\gamma}_j = q (\gamma_j = 1) = \frac{\pi_1 \alpha p_j^{\alpha-1}}{\pi_1 \alpha p_j^{\alpha-1} + 1 - \pi_1},$$
and get the Q function:

\[ Q = \mathbb{E}_q \left[ \log \Pr(p|\gamma; \alpha) + \log \Pr(\gamma|\pi_1) \right] \]
\[ = \sum_{j=1}^{M} \tilde{\gamma}_j \left( \log \alpha + (\alpha - 1) \log p_j + \log \pi_1 \right) + \sum_{j=1}^{M} (1 - \tilde{\gamma}_j) \log (1 - \pi_1). \]

The incomplete log likelihood can be evaluated as:

\[ L = \sum_{j=1}^{M} \tilde{\gamma}_j \left( \log \alpha + (\alpha - 1) \log p_j + \log \pi_1 - \log \tilde{\gamma}_j \right) \]
\[ + \sum_{j=1}^{M} (1 - \tilde{\gamma}_j) \left( \log (1 - \pi_1) - \log (1 - \tilde{\gamma}_j) \right). \]

In the M step, we update \( \alpha \) and \( \pi_1 \) by maximizing the Q function. We have

\[ \alpha = - \frac{\sum_{j=1}^{M} \tilde{\gamma}_j}{\sum_{j=1}^{M} \tilde{\gamma}_j \log p_j}, \]
\[ \pi_1 = \frac{1}{M} \sum_{j=1}^{M} \tilde{\gamma}_j. \]

**Implementation**

Input: \( p \), Initialize: \( \alpha = 0.1, \pi_1 = 0.1 \), Output: \( \alpha, \pi_1, \{\tilde{\gamma}_j\}_{j=1,...,M}. \)

- Initialize \( \alpha = 0.1, \pi_1 = 0.1. \)
- E-step: For \( j = 1, ..., M \), calculate \( \tilde{\gamma}_j \) as follows

\[ \tilde{\gamma}_j = \frac{\pi_1 \alpha p_j^{\alpha-1}}{\pi_1 \alpha p_j^{\alpha-1} + 1 - \pi_1}. \]

Calculate \( L \):

\[ L = \sum_{j=1}^{M} \tilde{\gamma}_j \left( \log \alpha + (\alpha - 1) \log p_j + \log \pi_1 - \log \tilde{\gamma}_j \right) \]
\[ + \sum_{j=1}^{M} (1 - \tilde{\gamma}_j) \left( \log (1 - \pi_1) - \log (1 - \tilde{\gamma}_j) \right). \]
• M-step:

\[
\alpha = - \frac{\sum_{j=1}^{M} \tilde{\gamma}_j}{\sum_{j=1}^{M} \tilde{\gamma}_j \log p_j},
\]

\[
\pi_1 = \frac{1}{M} \sum_{j=1}^{M} \tilde{\gamma}_j.
\]

• Check convergence.

**Stage 2: Latent fixed-effect model (LFM)**

Suppose we have the P-values of M SNPs for a given phenotype. Similarly, we assume

\[
p_j \sim \begin{cases} 
U[0, 1], & \gamma_j = 0, \\
\text{Beta}(\alpha, 1), & \gamma_j = 1,
\end{cases}
\]

where \( p \in \mathbb{R}^M \) are the P-values, \( \gamma_j = 1 \) indicates the \( j \)-th is associated with this phenotype and \( \gamma_j = 0 \) otherwise, and \( 0 < \alpha < 1 \).

To integrate more information, we consider the logistic fixed-effect model:

\[
\log \frac{\Pr(\gamma_j = 1|Z_j)}{\Pr(\gamma_j = 0|Z_j)} = Z_j b,
\]

where \( Z \in \mathbb{R}^{M \times (L+1)} \) and \( b = [b_0, b_1, b_2, ..., b_L]^T \) is an unknown vector of fixed effects, \( L \) is the number of covariates.

We can use EM algorithm to compute the posterior and parameter estimation.

Let \( \theta = \{\alpha, b\} \) be the collection of model parameters. The complete data likelihood can be written as

\[
\Pr(p, \gamma|Z; \theta) = \Pr(p|\gamma; \alpha) \Pr(\gamma|Z; b),
\]
where

\[
\Pr (p|\gamma; \alpha) = \prod_{j=1}^{M} \Pr (p_j|\gamma_j; \alpha) = \prod_{j=1}^{M} (\alpha p_j^{\alpha-1})^{\gamma_j},
\]

\[
\Pr (\gamma|Z; b) = \prod_{j=1}^{M} e^{\gamma_j Z_j b} S (-Z_j b).
\]

In the E step, we compute the posterior:

\[
\tilde{\gamma}_j = q (\gamma_j = 1) = \frac{e^{Z_j b \alpha p_j^{\alpha-1}}}{e^{Z_j b \alpha p_j^{\alpha-1}} + 1},
\]

and get the Q function:

\[
Q = \sum_{j=1}^{M} \tilde{\gamma}_j (\log \alpha + (\alpha - 1) \log p_j + Z_j b) + \sum_{j=1}^{M} \log S (-Z_j b).
\]

The incomplete log likelihood can be evaluated as:

\[
L = \sum_{j=1}^{M} \tilde{\gamma}_j (\log \alpha + (\alpha - 1) \log p_j + Z_j b - \log \tilde{\gamma}_j)
\]

\[
- \sum_{j=1}^{M} (1 - \tilde{\gamma}_j) \log (1 - \tilde{\gamma}_j) + \sum_{j=1}^{M} \log S (-Z_j b).
\]

In the M step, we update \( \alpha \) by maximizing the Q function. We have

\[
\alpha = - \frac{\sum_{j=1}^{M} \tilde{\gamma}_j}{\sum_{j=1}^{M} \tilde{\gamma}_j \log p_j}.
\]

We use Newton’s method to update \( b \):

\[
b = b_{old} - H^{-1} g.
\]
where

\[ g = \sum_{j=1}^{M} (-\tilde{\gamma}_j + S(Z_j b)) Z_j, \]

\[ H = \sum_{j=1}^{M} S(Z_j b) S(-Z_j b) Z_j^T Z_j. \]

**Implementation**

Input: \( p, Z, \alpha, b_0 = \log \frac{\pi_1}{1-\pi_1} \), Output: \( \alpha, b, \{\tilde{\gamma}_j\}_{j=1}^{M} \).

- Initialize \( \alpha, b = (b_0, 0, \ldots, 0)^T \).

- E-step: For \( j = 1, \ldots, M \), calculate \( \tilde{\gamma}_j \) as follows

\[ \tilde{\gamma}_j = q(\gamma_j = 1) = \frac{e^{Z_j b} \alpha p_j^{\alpha-1}}{e^{Z_j b} \alpha p_j^{\alpha-1} + 1}. \]

Calculate \( L \):

\[ L = \sum_{j=1}^{M} \tilde{\gamma}_j (\log \alpha + (\alpha - 1) \log p_j + Z_j b - \log \tilde{\gamma}_j) \]

\[ - \sum_{j=1}^{M} (1 - \tilde{\gamma}_j) \log (1 - \tilde{\gamma}_j) + \sum_{j=1}^{M} \log S(-Z_j b). \]

- M-step

\[ \alpha = - \frac{\sum_{j=1}^{M} \pi_j}{\sum_{j=1}^{M} \pi_j \log p_j}, \]

\[ g = \sum_{j=1}^{M} (-\tilde{\gamma}_j + S(Z_j b)) Z_j, \]

\[ H = \sum_{j=1}^{M} S(Z_j b) S(-Z_j b) Z_j^T Z_j, \]

\[ b = b_{old} - H^{-1} g. \]

- Check convergence.
Stage 3: Logistic sparse mixed model

Suppose the latent states $\gamma$ of $M$ SNPs for a given phenotype are given. We consider a logistic mixed model:

$$\log \frac{\Pr (\gamma_j = 1 | Z_j, A_j)}{\Pr (\gamma_j = 0 | Z_j, A_j)} = Z_j b + A_j \beta = \sum_{l=0}^{L} Z_{jl} b_l + \sum_{k=1}^{K} A_{jk} \beta_k,$$

where $Z \in \mathbb{R}^{M \times (L+1)}$, $A \in \mathbb{R}^{M \times K}$, $b = [b_0, b_1, ..., b_L]^T$ is an unknown vector of fixed effects, $\beta = [\beta_1, \beta_2, ..., \beta_K]^T$ is an unknown vector of random effects with a spike-slab prior:

$$\beta_k \sim \begin{cases} N(0, \sigma^2), & \eta_k = 1, \\ \delta_0, & \eta_k = 0, \end{cases}$$

where $\eta_k$ is another latent variable with $\Pr (\eta_k = 1) = \omega$. Here $\eta_k = 1$ means the $k$-th annotation is relevant to this phenotype and $\eta_k = 0$ otherwise.

To handle the Dirac function, we reparametrize the spike-slab prior as $\tilde{\beta}_k \sim N(0, \sigma^2)$, then $\beta_k = \eta_k \tilde{\beta}_k$.

We can use variational EM algorithm to compute the posterior and parameter estimation.

Let $\theta = \{\alpha, b, \sigma^2, \omega\}$ be the collection of model parameters. Using the sigmoid function denoted as $S(x) = \frac{1}{1+e^{-x}}$, the complete data likelihood can be written as

$$\Pr \left( \gamma, \tilde{\beta}, \eta | Z, A; \theta \right) = \Pr \left( \gamma | Z, A, \tilde{\beta}, \eta; b \right) \Pr \left( \tilde{\beta}, \eta | \sigma^2, \omega \right),$$

where

$$\Pr \left( \gamma | Z, A, \tilde{\beta}, \eta; b \right) = \prod_{j=1}^{M} \Pr \left( \gamma_j | Z_j, A_j, \tilde{\beta}, \eta; b \right)$$

$$= \prod_{j=1}^{M} e^{\gamma_j (Z_j b + \sum_k A_{jk} \eta_k \tilde{\beta}_k)} S \left( -Z_j b - \sum_k A_{jk} \eta_k \tilde{\beta}_k \right),$$

$$\Pr \left( \tilde{\beta}, \eta | \sigma^2, \omega \right) = \prod_{k=1}^{K} \Pr \left( \tilde{\beta}_k, \eta_k | \sigma^2, \omega \right) = \prod_{k=1}^{K} N \left( \tilde{\beta}_k | 0, \sigma^2 \right) \omega^{\eta_k} (1 - \omega)^{1-\eta_k}.$$
We can use JJ bound to bound the sigmoid function by

\[ S(x) \geq S(\xi) \exp \left\{ \frac{(x - \xi)}{2} - \lambda(\xi) \left( x^2 - \xi^2 \right) \right\}, \]

where \( \lambda(\xi) = \frac{1}{2\xi} \left[ S(\xi) - \frac{1}{2} \right] \). Using this bound, we have a tractable lower bound of \( \Pr(\gamma|Z, A, \bar{\beta}, \eta; b) \) which is denoted by \( h(\gamma|Z, A, \bar{\beta}, \eta; b, \xi) \):

\[
h\left( \gamma_j | Z_j, A_j, \bar{\beta}, \eta; b, \xi_j \right)
= e^{\gamma_j \left( Z_j b + \sum A_{jk} \eta_k \bar{\beta}_k \right)} S(\xi_j) \exp \left( -\lambda(\xi_j) \left( \left( Z_j b + \sum A_{jk} \eta_k \bar{\beta}_k \right)^2 - \xi_j^2 \right) \right)
+ e^{\gamma_j \left( Z_j b + \sum A_{jk} \eta_k \bar{\beta}_k \right)} S(\xi_j) \exp \left( -\frac{Z_j b + \sum A_{jk} \eta_k \bar{\beta}_k + \xi_j}{2} \right).
\]

Next, Let \( q(\bar{\beta}, \eta) \) be an approximation of the posterior \( \Pr(\bar{\beta}, \eta|Z, A; \theta) \). Then we can obtain a lower bound of the logarithm of the marginal likelihood:

\[
\log \Pr(\gamma|Z, A; \theta)
= \log \sum_{\eta} \int \Pr(\gamma, \bar{\beta}, \eta|Z, A; \theta) \, d\bar{\beta}
= \log \sum_{\eta} \int \Pr(\gamma|Z, A, \bar{\beta}, \eta; b) \Pr(\bar{\beta}, \eta|\sigma^2, \omega) \, d\bar{\beta}
\geq \log \sum_{\eta} \int h\left( \gamma|Z, A, \bar{\beta}, \eta; b, \xi \right) \Pr(\bar{\beta}, \eta|\sigma^2, \omega) \, d\bar{\beta}
\geq \sum_{\eta} \int q(\bar{\beta}, \eta) \log \frac{h\left( \gamma|Z, A, \bar{\beta}, \eta; b, \xi \right) \Pr(\bar{\beta}, \eta|\sigma^2, \omega)}{q(\bar{\beta}, \eta)} \, d\bar{\beta}
= \mathbb{E}_q \left[ \log \frac{h\left( \gamma|Z, A, \bar{\beta}, \eta; b, \xi \right) \Pr(\bar{\beta}, \eta|\sigma^2, \omega)}{q(\bar{\beta}, \eta)} \right] - \log q(\bar{\beta}, \eta)
\triangleq L(q),
\]

where \( L(q) \) is the lower bound. The second inequality follows Jensen’s inequality. We can maximize \( L(q) \) instead of the marginal likelihood to get parameter estimations. To make it feasible to evaluate the lower bound, we assume that \( q(\bar{\beta}, \eta) \) can be
factorized as
\[
q(\tilde{\beta}, \eta) = \prod_{k=1}^{K} q(\tilde{\beta}_k, \eta_k) = \prod_{k=1}^{K} q(\tilde{\beta}_k|\eta_k) q(\eta_k),
\]
where \(q(\eta_k = 1) = \omega_k\).

We can obtain an approximation according to the mean-field method:
\[
\log q(\tilde{\beta}_i, \eta_i) = E_{k \neq i} \left[ \log h(\gamma|Z, A, \tilde{\beta}, \eta, b, \xi) + \log \Pr(\tilde{\beta}, \eta|\sigma^2, \omega) \right],
\]
where the expectation is taken under the distribution \(q(\tilde{\beta}_i, \eta_i) = \prod_{k \neq i} q(\tilde{\beta}_k, \eta_k)\).

Then we have
\[
q(\tilde{\beta}_i, \eta_i) = \left[ \omega_i N(\mu_i, s_i^2) \right]^{\eta_i} \left[ (1 - \omega_i) N(0, \sigma^2) \right]^{1-\eta_i},
\]
where
\[
\mu_i = s_i^2 \sum_{j=1}^{M} \left( \pi_j - \frac{1}{2} - 2\lambda(\xi_j) \left( Z_j b + \sum_{k \neq i} A_{jk} E_k[\eta_k \tilde{\beta}_k] \right) \right) A_{ji},
\]
\[
s_i^2 = \frac{\sigma^2}{1 + 2\sigma^2 \sum_{j=1}^{M} \lambda(\xi_j) A_{ji}^2}.
\]

Then we maximize \(L(q)\) with respect to \(\omega_k\) and \(\xi_j\) and get
\[
\omega_k = \frac{1}{1 + \exp(-u_k)}, \text{ where } u_k = \log \frac{\omega}{1 - \omega} + \frac{1}{2} \log \frac{s_k^2}{\sigma^2} + \frac{\mu_k^2}{2s_k^2},
\]
\[
\xi_j^2 = \left( Z_j b + \sum_{k} A_{jk} \omega_k \mu_k \right)^2 + \sum_{k} A_{jk}^2 \left( \omega_k \left( s_k^2 + \mu_k^2 \right) - \omega_k^2 \mu_k^2 \right).
\]
Now we have evaluate $L(q)$:

$$L(q) = \sum_{j=1}^{M} \left( \gamma_j \left( Z_j b + \sum_k A_{jk} \omega_k \mu_k \right) + \log S(\xi_j) - \left( Z_j b + \sum_k A_{jk} \omega_k \mu_k + \xi_j \right) / 2 \right)$$

$$+ \sum_{j=1}^{M} \left( -\lambda (\xi_j) \left( \left( Z_j b + \sum_k A_{jk} \omega_k \mu_k \right)^2 - \xi_j^2 \right) \right)$$

$$+ \sum_{j=1}^{M} \left( \lambda (\xi_j) \sum_k A_{jk}^2 \omega_k^2 \mu_k^2 - \lambda (\xi_j) \sum_k A_{jk}^2 \omega_k \left( s_k^2 + \mu_k^2 \right) \right)$$

$$- \frac{1}{2\sigma^2} \sum_{k=1}^{K} \left( \omega_k \left( s_k^2 + \mu_k^2 \right) - \omega_k \sigma^2 \right) + \sum_{k=1}^{K} \omega_k \log \omega + \sum_{k=1}^{K} (1 - \omega_k) \log (1 - \omega)$$

$$+ \sum_{k=1}^{K} \left( \frac{1}{2} \omega_k \left( \log s_k^2 - \log \sigma^2 \right) - \omega_k \log \omega_k - (1 - \omega_k) \log (1 - \omega_k) \right).$$

With $q(\gamma, \tilde{\beta}, \eta)$ obtained, we can evaluate the lower bound and then update the model parameters by maximizing $L(q)$.

In the M step, we update $\sigma^2$ and $\omega$ by maximizing $L(q)$. We have

$$\sigma^2 = \frac{\sum_{k=1}^{K} \omega_k \left( s_k^2 + \mu_k^2 \right)}{\sum_{k=1}^{K} \omega_k},$$

$$\omega = \frac{1}{K} \sum_{k=1}^{K} \omega_k.$$

We use Newton’s method to update $b$:

$$b = b_{old} - H^{-1} g,$$

where

$$g = - \sum_{j=1}^{M} Z_j^T \left( \gamma_j - 2\lambda (\xi_j) \left( Z_j b + \sum_k A_{jk} \omega_k \mu_k \right) - \frac{1}{2} \right),$$

$$H = 2 \sum_{j=1}^{M} \lambda (\xi_j) Z_j^T Z_j.$$
Input: $\mathbf{Z}, \mathbf{A}, \{\gamma_j = \tilde{\gamma}_j\}_{j=1,...,M}$, $\mathbf{b}$, Initialize: $\sigma^2 = 1$, $\omega = 0.5$, $\{\omega_k = 0, \mu_k = 0\}_{k=1,...,K}$, $\mathbf{ξ} = \mathbf{Zb}$, Output: $\mathbf{b}$, $\mathbf{ξ}$, $\sigma^2$, $\omega$, $\{\omega_k, \mu_k\}_{k=1,...,K}$.

- Initialize $\mathbf{b}$, $\mathbf{ξ} = \mathbf{Zb}$, $\sigma^2 = 1$, $\omega = 0.5$, $\{\omega_k = 0, \mu_k = 0\}_{k=1,...,K}$. Let $\tilde{y} = \sum_k A_{jk} \omega_k \mu_k$.

- E-step: For $i = 1,...,K$, first obtain $\tilde{y}_i = \tilde{y} - A_{ji} \omega_i \mu_i$, and then update $\mu_i, s_i^2, \omega_i$ and $\tilde{y}$ as follows

$$s_i^2 = \frac{\sigma^2}{1 + 2\sigma^2 \sum_{j=1}^M \lambda(\xi_j) A_{ji}^2},$$

$$\mu_i = s_i^2 \sum_{j=1}^M \left( \left( \gamma_j - \frac{1}{2} - 2\lambda(\xi_j) (\mathbf{Z}_j \mathbf{b} + \tilde{y}_i) \right) A_{ji} \right),$$

$$\omega_i = \frac{1}{1 + \exp(-u_i)}, \text{ where } u_i = \log \frac{\omega}{1 - \omega} + \frac{1}{2} \log \frac{s_i^2}{\sigma^2} + \frac{\mu_i^2}{2s_i^2},$$

$$\tilde{y} = \tilde{y}_i + A_{ji} \omega_i \mu_i.$$

Then for $j = 1,...,M$, update $\xi_j$ as follows

$$\xi_j^2 = (\mathbf{Z}_j \mathbf{b} + \tilde{y})^2 + \sum_k A_{jk}^2 \left( \omega_k \left( s_k^2 + \mu_k^2 \right) - \omega_k^2 \mu_k^2 \right).$$

Calculate $L(q)$:

$$L(q) = \sum_{j=1}^M \left( \gamma_j (\mathbf{Z}_j \mathbf{b} + \tilde{y}) + \log S(\xi_j) - \frac{\mathbf{Z}_j \mathbf{b} + \tilde{y} + \xi_j}{2} \right)$$

$$- \frac{1}{2\sigma^2} \sum_{k=1}^K \left( \omega_k \left( s_k^2 + \mu_k^2 \right) - \omega_k \sigma^2 \right) + \sum_{k=1}^K \omega_k \log \omega + \sum_{k=1}^K (1 - \omega_k) \log (1 - \omega)$$

$$+ \sum_{k=1}^K \left[ \frac{1}{2} \omega_k \left( \log s_k^2 - \log \sigma^2 \right) - \omega_k \log \omega_k - (1 - \omega_k) \log (1 - \omega_k) \right].$$
• M-step

\[ g = - \sum_{j=1}^{M} Z_j^T \left( \pi_j - 2\lambda (\xi_j) (Z_j b + \tilde{y}) - \frac{1}{2} \right), \]

\[ H = 2 \sum_{j=1}^{M} \lambda (\xi_j) Z_j^T Z_j, \]

\[ b = b_{old} - H^{-1} g, \]

\[ \sigma^2 = \frac{\sum_{k=1}^{K} \omega_k (s^2_k + \mu^2_k)}{\sum_{k=1}^{K} \omega_k}, \]

\[ \omega = \frac{1}{K} \sum_{k=1}^{K} \omega_k. \]

• Check convergence.

Stage 4: LSMM

Implementation

Input: \( p, Z, A, \alpha, b, \xi, \sigma^2, \omega, \{\omega_k, \mu_k\}_{k=1,...,K} \), Initialize: \( \{\pi_j = \tilde{\gamma}_j\}_{j=1,...,M} \), Output: \( \alpha, b, \sigma^2, \omega, \{\omega_k, \beta_k = \mu_k \omega_k\}_{k=1,...,K}, \{\pi_j\}_{j=1,...,M} \).

• Initialize \( \alpha, \sigma^2, \omega, b, \{\omega_k, \mu_k\}_{k=1,...,K}, \{\xi_j, \pi_j\}_{j=1,...,M} \). Let \( \bar{y} = \sum_k A_{jk} \omega_k \mu_k \).

• E-step: For \( i = 1, ..., K \), first obtain \( \tilde{y}_i = \bar{y} - A_{ji} \omega_i \mu_i \), and then update \( \mu_i, s^2_i, \omega_i \) and \( \tilde{y} \) as follows

\[ s^2_i = \frac{\sigma^2}{1 + 2\sigma^2 \sum_{j=1}^{M} \lambda (\xi_j) A^2_{ji}}, \]

\[ \mu_i = s^2_i \sum_{j=1}^{M} \left( \left( \frac{\pi_j}{2} - 2\lambda (\xi_j) (Z_j b + \tilde{y}) \right) A_{ji} \right), \]

\[ \omega_i = \frac{1}{1 + \exp(-u_i)}, \text{ where } u_i = \log \frac{\omega}{1 - \omega} + \frac{1}{2} \log \frac{s^2_i}{\sigma^2} + \frac{\mu^2_i}{2s^2_i}, \]

\[ \tilde{y} = \bar{y} + A_{ji} \omega_i \mu_i. \]
Then for $j = 1, \ldots, M$, update $\pi_j, \xi_j$ as follows

\[
\pi_j = \frac{1}{1 + \exp(-v_j)}, \quad \text{where } v_j = \log \alpha + (\alpha - 1) \log p_j + Z_j b + \bar{y},
\]

\[
\xi_j^2 = (Z_j b + \bar{y})^2 + \sum_k A_{jk}^2 (\omega_k (s_k^2 + \mu_k^2) - \omega_k^2 \mu_k^2).
\]

Calculate $L(q)$:

\[
L(q) = \sum_{j=1}^M \pi_j (\log \alpha + (\alpha - 1) \log p_j) - \sum_{j=1}^M (\pi_j \log \pi_j + (1 - \pi_j) \log (1 - \pi_j)) \]

\[
+ \sum_{j=1}^M \left( \pi_j (Z_j b + \bar{y}) + \log S(\xi_j) - \frac{Z_j b + \bar{y} + \xi_j}{2} \right) \]

\[- \frac{1}{2\sigma^2} \sum_{k=1}^K \left( \omega_k (s_k^2 + \mu_k^2) - \omega_k \sigma^2 \right) + \sum_{k=1}^K \omega_k \log \omega + \sum_{k=1}^K (1 - \omega_k) \log (1 - \omega) \]

\[+ \sum_{k=1}^K \left( \frac{1}{2} \omega_k (\log s_k^2 - \log \sigma^2) - \omega_k \log \omega_k - (1 - \omega_k) \log (1 - \omega_k) \right). \]

- **M-step**

\[
\alpha = - \frac{\sum_{j=1}^M \pi_j}{\sum_{j=1}^M \pi_j \log p_j},
\]

\[
\sigma^2 = \frac{\sum_{k=1}^K \omega_k (s_k^2 + \mu_k^2)}{\sum_{k=1}^K \omega_k},
\]

\[
\omega = \frac{1}{K} \sum_{k=1}^K \omega_k,
\]

\[
g = - \sum_{j=1}^M Z_j^T \left( \pi_j - 2\lambda(\xi_j) (Z_j b + \bar{y}) - \frac{1}{2} \right),
\]

\[
H = 2 \sum_{j=1}^M \lambda(\xi_j) Z_j^T Z_j,
\]

\[
b = b_{old} - H^{-1} g.
\]

- Evaluate $L(q)$ to track the convergence of the algorithm.
2.4 Inference

After the convergence of the variational EM algorithm, the approximated posterior of latent variables $\gamma$ and $\eta$ can be obtained. Using this information, we are able to prioritize risk SNPs and relevant cell-type specific functional annotations.

2.4.1 Identification of Risk SNPs

Risk SNPs are identified based on $q(\gamma_j = 1)$, an approximation of the posterior probability that the $j$-th SNP is associated with this phenotype. Accordingly, we can calculate the approximated local false discovery rate $fdr_j = 1 - q(\gamma_j = 1)$. To control the global false discovery rate (FDR), we sort SNPs by $fdr$ from the smallest to the largest and regard the $j$-th re-ordered SNP as a risk SNP if

$$FDR_{(j)} = \frac{\sum_{i=1}^{j} fdr_{(i)}}{j} \leq \tau,$$

where $fdr_{(i)}$ is the $i$-th ordered $fdr$, $FDR_{(j)}$ is the corresponding global FDR, and $\tau$ is the threshold of global FDR. In simulations, we chose $\tau = 0.1$.

2.4.2 Detection of Relevant Cell-type Specific Functional Annotations

Relevant cell-type specific functional annotations are inferred from $q(\eta_k = 1)$, an approximation of the posterior probability that annotation $k$ is relevant to this phenotype. Similarly, we can calculate the approximated local false discovery rate $fdr_k = 1 - q(\eta_k = 1)$ and convert it into the global false discovery rate. We can either control the local false discovery rate (e.g., $fdr_k \leq 0.1$) or global false discovery rate with $\tau = 0.1$. 

29
2.5 Simulation

2.5.1 Ideal Case

Simulation setup

We conducted simulations to evaluate the performance of the proposed LSMM. The simulation data was generated from the generative model as follows. The numbers of SNPs, fixed effects (genic category annotations) and random effects (cell-type specific functional annotations) were set to be $M = 100,000$, $L = 10$ and $K = 500$ respectively. The entries in design matrices $Z_{jl}$ and $A_{jk}$ were generated from $Bernoulli (0.1)$, $j = 1, ..., M$, $l = 1, ..., L$ and $k = 1, ..., K$. Given the proportion of relevant cell-type specific functional annotations $\omega$, $\eta_k$ was drawn from $Bernoulli (\omega)$ and the corresponding nonzero entries of random effects $\beta$ were simulated from $N (0, 1)$. The first entry of the coefficients of fixed effects $b$, i.e., the intercept in the logistic model, was fixed at $-2$ and other entries were generated from $N (0, 1)$ and then kept fixed in multiple replications. After that, we simulated $\gamma_j$ from Bernoulli distribution with probability $S (Z_j b + A_j \beta)$, and then generated $p_j$ from $U [0, 1]$ if $\gamma_j = 0$ and $Beta (\alpha, 1)$ otherwise.

Performance in the identification of risk SNPs

We compared LSMM with two special cases, LFM (with fixed effects only) and TGM (without fixed effects and random effects). After prioritizing the risk SNPs using these methods, we made a comparison upon their empirical FDR, power, area under the receiver operating characteristic curve (AUC) and partial AUC. We varied the proportion of relevant random effects $\omega$ at $\{0, 0.01, 0.05, 0.1, 0.2\}$ and controlled global FDR at 0.1 to evaluate empirical FDR and power.

Figure 2.1-2.3 shows the performance of these three models with $\alpha \in \{0.2, 0.4, 0.6\}$ and $K \in \{100, 500, 1000\}$ based on 50 replications. The results show that the empirical FDRs are indeed controlled at the nominal level ($\tau = 0.1$) for all these models. For TGM and LFM, the powers increase as the proportion of relevant functional annotations $\omega$ increases. This is because a larger $\omega$ could result in an increasing proportion of non-null group for SNPs. However, the AUC and partial AUC of LFM slightly
decrease because the estimates of fixed effects using LFM would become less accurate when the impact of functional annotations becomes larger. LSMM can adaptively select relevant functional annotations to improve its performance. As expected, it outperforms both TGM and LFM in terms of the power, AUC and partial AUC.

![Figure 2.1: Performance of LSMM, LFM and TGM for identification of risk SNPs with α = 0.2 and K = 100 (upper), K = 500 (middle), K = 1000 (lower).](image)

One may wonder what if we do not do variable selection and simply treat the effects of all covariates as fixed effects. We evaluated this approach and found that, without variable selection, the FDR would be inflated when the GWAS signal is relatively weak (see Figure 2.4).

**Performance in the detection of relevant cell-type specific functional annotations**

We evaluated the performance of LSMM in the detection of relevant cell-type specific functional annotations in terms of the FDR, power, AUC and partial AUC. We varied the proportion of relevant cell-type specific functional annotations $\omega$ at \{0.01, 0.05, 0.1, 0.2\} and controlled global FDR at 0.1 to evaluate empirical FDR and
Figure 2.2: Performance of LSMM, LFM and TGM for identification of risk SNPs with $\alpha = 0.4$ and $K = 100$ (upper), $K = 500$ (middle), $K = 1000$ (lower).

Figure 2.3: Performance of LSMM, LFM and TGM for identification of risk SNPs with $\alpha = 0.6$ and $K = 100$ (upper), $K = 500$ (middle), $K = 1000$ (lower).
Figure 2.4: FDR of LSMM and LSMM (treat the effects of all covariates as fixed effects) for identification of risk SNPs with $K = 500$.

The results based on 50 replications are given in Figure 2.5–2.7. The empirical FDR is controlled at 0.1 with conservativeness. This is because the variational approach is adopted to approximate the posterior, e.g., the JJ bound and mean-field approximation. When the signal of the GWAS data is relatively strong, i.e., $\alpha$ is relatively small, LSMM has a very good performance of detecting relevant functional annotations, as indicated by power, AUC and partial AUC.

When the number of SNPs becomes larger (e.g., $M = 500,000$), for a fixed signal strength, the empirical FDR becomes less conservative and the performance becomes better (See Figure 2.8).

Estimation of parameters

Regarding parameter estimation, LSMM provides a satisfactory estimate of $\alpha$, the parameter in Beta distribution (see Figure 2.9). When the signal strength of GWAS data is not very weak, the estimated fixed effects $b$ (Figures 2.10–2.20) and the proportion of non-zero random effects $\omega$ (Figure 2.21) are relatively accurate.
Figure 2.5: Performance of LSMM for detection of relevant annotations with $\alpha = 0.2$ and $K = 100$ (upper), $K = 500$ (middle), $K = 1000$ (lower).

Figure 2.6: Performance of LSMM for detection of relevant annotations with $\alpha = 0.4$ and $K = 100$ (upper), $K = 500$ (middle), $K = 1000$ (lower).
Figure 2.7: Performance of LSMM for detection of relevant annotations with $\alpha = 0.6$ and $K = 100$ (upper), $K = 500$ (middle), $K = 1000$ (lower).

Figure 2.8: Performance of LSMM for detection of relevant annotations with $K = 100$ and $\alpha = 0.2$ (upper), $\alpha = 0.4$ (middle), $\alpha = 0.6$ (lower).
Figure 2.9: Performance in estimation of parameter $\alpha$ when the true $\alpha = 0.2$ (upper), $\alpha = 0.4$ (middle), $\alpha = 0.6$ (lower).
Figure 2.10: Performance in estimation of parameter $b_0$.

Figure 2.11: Performance in estimation of parameter $b_1$
Figure 2.12: Performance in estimation of parameter $b_2$.

Figure 2.13: Performance in estimation of parameter $b_3$.  

38
Figure 2.14: Performance in estimation of parameter $b_4$.

Figure 2.15: Performance in estimation of parameter $b_5$. 
Figure 2.16: Performance in estimation of parameter $b_6$.

Figure 2.17: Performance in estimation of parameter $b_7$. 
Figure 2.18: Performance in estimation of parameter $b_8$.

Figure 2.19: Performance in estimation of parameter $b_9$. 

41
Figure 2.20: Performance in estimation of parameter $b_{10}$.

Figure 2.21: Performance in estimation of parameter $\omega$. 
Initial parameter specification

The parameters which need initialization in LSMM are only $\alpha$ and $\pi_1$ in the first stage (Two-group model, in short, TGM), and their estimates naturally give the starting point of the second stage. Here we used the TGM to generate data such that we can evaluate whether the estimates converge to their true values. In the simulation, we set the number of SNPs $M = 100,000$ and varied the true value of $\pi_1 \in \{0.01, 0.05, 0.1, 0.15, 0.2\}$. To check whether LSMM could give accurate estimates when using different initial values, we considered two cases, default setting and random setting. In the default setting, both $\alpha$ and $\pi_1$ are initialized at 0.1. In the random setting, we randomly generated the initial values of $\alpha$ from $U[0.1, 0.6]$ and the initial values of $\pi_1$ from $U[0, 0.3]$.

The results of the estimation $\hat{\pi}_1$ using LSMM (default setting and random setting) based on 50 replications are shown in the upper panel of Figure 2.22. We controlled global FDR at 0.1 to evaluate empirical FDR. The true values are indicated by dotted lines with different colors. Comparing the performance of difference initial value settings, default setting and random setting, we note that LSMM is not sensitive to initial parameter specification in most situations except when the true proportion of risk SNPs is small and the signal of GWAS data is weak (e.g., $\pi_1 = 0.01$ for $\alpha = 0.6$). However, LSMM can still provide a valid FDR control which is shown in the lower panel of Figure 2.22. In the context of GWAS, the proportion of risk variants is not very small due to the polygenic effect. Therefore, we believe LSMM with default setting will work well in practice, without suffering much from initialization.

Computational time

The computational time of LSMM depends on the strength of GWAS signal, the number of SNPs and the number of random effects. The left panel of Figure 2.23 shows that the computational time is nearly linear with respect to $M$ and $K$ with $\alpha = 0.2$. In the right panel, we fixed $M = 100,000$ and varied $K$ and $\alpha$. When the GWAS signal is relatively weak, e.g., $\alpha = 0.6$, the timings of LSMM remain the same for different scales of random effects. This is because LSMM adopts a warm-start strategy and its last two stages start from the estimates at the second stage (i.e., fixed
Figure 2.22: Upper panel: parameter estimation ($\hat{\pi}_1$ v.s. true $\pi_1$) using LSMM (default setting and random setting). Lower panel: FDR for identification of risk SNPs using LSMM (default setting and random setting).

effects only) and converge in a few iterations when the GWAS signal is too weak to provide information for updating the random effects.

Figure 2.23: Computational time of LSMM. Left panel: We varied the number of SNPs $M$ and the number of random effects $K$, with $\alpha = 0.2$. Right panel: We varied the number of random effects $K$ and the strength of GWAS signal $\alpha$ with $M = 100,000$. The results are summarized from 10 replications.
2.5.2 More simulations

We conducted simulations to test the robustness of LSMM under some special cases or under the situation that some assumptions in our proposed model are violated.

Performance of LSMM when the number of SNPs is small

We conducted simulations with the number of SNPs $M$ varied from 1,000 to 100,000 to evaluate the performance of LSMM. In the simulation, we set $L = 10$, $K = 100$, $\alpha = 0.2$ and $\omega = 0.2$. To easily control signal-noise ratio, we used the probit model:

$$y_j = Z_j b + A_j \beta + e_j,$$

where $e_j \sim N(0, \sigma^2_e)$. And we set $\gamma_j = 1$ if $y_j > 0$, $\gamma_j = 0$ if $y_j \leq 0$. The first entry of the coefficients of fixed effects $b$, i.e. the intercept term, was fixed at $-1$ and other entries were generated from $N(0,1)$ and fixed during multiple replications. We varied the signal-noise ratio $r = \frac{\text{var}(Zb + A\beta)}{\text{var}(e)} = \{4:1,1:1,1:4\}$ and controlled global FDR at 0.1 to evaluate empirical FDR and power.

The results based on 50 replications are given in Figures 2.24 and 2.25. As the number of SNPs becomes smaller, the performance of LSMM for both identification of risk SNPs and detection of relevant annotations become worse, as indicated by power, AUC and partial AUC. With a large signal-noise ratio, the performance of LSMM becomes better, especially when the number of SNPs is small. Therefore, the performance of LSMM in the detection of relevant annotations is influenced by the signal strength of the GWAS data, the number of SNPs and the importance of annotations. In order to obtain reliable results using LSMM, the number of SNPs should not be very small. To summarize, LSMM could be applied to a subset of SNPs when the number of SNPs is not too small and signals from annotations are not too weak.
Figure 2.24: Performance of LSMM for identification of risk SNPs based on probit model.

Figure 2.25: Performance of LSMM for detection of relevant annotations based on probit model.

**Performance in identification of relevant annotations when the annotations in design matrix of fixed effects and random effects are not independent**

We also conducted the following simulations to examine the role of adjusting covariates (i.e., genic category annotations) using fixed effects for detecting relevant cell-type specific annotations. We consider the case that genic category annotations and some cell-type specific annotations are correlated and \( \mathbf{b} \), the vector of coefficients corresponding to genic category annotations, is nonzero. Without adjusting genic category annotations, some irrelevant cell-type specific annotations will be falsely included in the model due to their correlation with genic category annotations.

To verify this, we simulated a case that 10 genic category annotations and first 50 cell-type specific annotations are correlated with correlation coefficient varied at \( \{0, 0.2, 0.4, 0.6, 0.8\} \) and the remaining annotations are generated independently. To simulate the design matrices for genic category and cell-type specific annotations, we first simulated \( M \) samples from a multivariate normal distribution with the correlation matrix among annotations and then made a cutoff so that 10% of the entries would be 1 and the others be 0. We controlled global FDR at 0.1 to evaluate em-
porical FDR. In the presence of correlation, as expected, a larger FDR of detecting relevant cell-type specific annotations is observed without adjusting genic category annotations (see Figure 2.26).

![FDR Diagram](image)

Figure 2.26: FDR of LSMM and LSMM without fixed effects for detection of relevant annotations with $\alpha = 0.2$ and $K = 100$.

Simulations based on individual-level data

To provide a reference for the relationship between heritability and $\alpha$, here we conducted simulations when the $P$-values for SNPs are obtained from individual-level data instead of directly simulating from the generative model 2.1. The simulation data was generated as follows. To simulate the genotype matrix $X$ for $N$ individuals with $M$ independent SNPs, we first draw the minor allele frequencies (MAFs) of these SNPs from $U[0,1]$. Based on the MAFs, the entries in the genotype matrix $X$, which were encoded by $\{0, 1, 2\}$, were generated according to the Hardy-Weinberg principle. Given $\gamma$, which was simulated as what we described in the paper, the corresponding nonzero entries of effect sizes $\beta_{SNP}$ were simulated from $N(0,1)$. The noise level $\sigma^2_e$ was specified to control heritability $h^2 = \frac{\text{var}(X\beta_{SNP})}{\text{var}(X\beta_{SNP}) + \sigma^2_e}$ at given levels. The phenotype data $y$ was generated based on $y = X\beta_{SNP} + e$, where $e_i \sim N(0,\sigma^2_e)$ for $i = 1,\ldots,N$. Then we conducted univariate linear regression to obtain the summary statistics ($P$-value) for each SNP.

In the simulation, we set $M = 20,000$, $L = 10$, $K = 100$ and $\omega = 0.1$. We varied
heritability $h^2 \in \{0.2, 0.4, 0.6, 0.8\}$ and the sample size $N \in \{10,000, 5,000\}$. Figure 2.27 shows the estimation of $\alpha$ using LSMM based on 50 replications, indicating that the value of $\alpha$ is determined by both heritability and sample size. When $N = 10,000$, heritability $h^2 = 0.6$ and $h^2 = 0.2$ are approximately corresponding to $\alpha = 0.4$ and $\alpha = 0.6$, respectively. When the sample size reduces to $N = 5,000$, the corresponding estimation of $\alpha$ becomes larger. To conclude, given fixed sample sizes and nonzero proportion, smaller $\alpha$ corresponds to larger heritability. Hence, we used $\alpha$ to indicate the strength of GWAS in our paper.

![Box plots showing the estimation of $\alpha$ for different heritabilities and sample sizes](image)

Figure 2.27: The estimation of parameter $\alpha$ using individual-level data.

**Simulations if $P$-values are not from beta distribution**

LSMM assumes that the $P$-values of non-null SNPs follow the Beta distribution. Here we simulated the underlying distribution of $P$-values in non-null group from other distributions. We first generated $z$-scores and then converted them to $P$-values. Here $z$-values from the null group follow the standard normal distribution and $z$-values from the non-null group follow the alternative distributions in Table 2.1. In these simulations, the $P$-values in non-null group converted from $z$-scores will not from Beta distribution. Instead of using generative model 2.2, we conducted simulations based on probit model:

$$y_j = Z_j b + A_j \beta + e_j,$$

where $e_j \sim N(0, \sigma_e^2)$. And we set $\gamma_j = 1$ if $y_j > 0$, $\gamma_j = 0$ if $y_j \leq 0$. The first entry of the coefficients of fixed effects $b$, i.e. the intercept term, was fixed at $-1$ and other entries were generated from $N(0, 1)$ and fixed during multiple replications.
We set $\alpha = 0.2, \omega = 0.2$ and varied the signal-noise ratio $r = \{4 : 1, 1 : 1, 1 : 4\}$. We controlled global FDR at 0.1 to evaluate empirical FDR. The results based on 50 replications are shown in Figures 2.28, indicating that the FDR of LSMM is still well controlled at the nominal level.

Table 2.1: Alternative distributions for $z$-scores.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>spiky</td>
<td>$0.4N(0, 0.25^2) + 0.2N(0, 0.5^2) + 0.2N(0, 1^2) + 0.2N(0, 2^2)$</td>
</tr>
<tr>
<td>near normal</td>
<td>$\frac{2}{3}N(0, 1^2) + \frac{1}{3}N(0, 2^2)$</td>
</tr>
<tr>
<td>skew</td>
<td>$\frac{1}{2}N(-2, 2^2) + \frac{1}{2}N(-1, 1.5^2) + \frac{1}{2}N(0, 1^2) + \frac{1}{6}N(1, 1^2)$</td>
</tr>
<tr>
<td>big-normal</td>
<td>$N(0, 4^2)$</td>
</tr>
</tbody>
</table>

Simulation study for evaluating the LD effects on LSMM

We assume independence among SNPs, which greatly facilitates the computation and inference of LSMM. To study the influence of LD effects on our LSMM, we used the observed genotype data (1,500 individuals from the 1958 British Birth Cohort (58C)) from WTCCC (The Wellcome Trust Case Control Consortium 2007). For simplicity, we only consider 23874 SNPs in chromosome 1 after quality control. We simulated a risk SNP every 1000 SNPs. So we had 24 risk SNPs. We assumed the 24 risk SNPs can explain 5% phenotypic variance. We used GCTA to simulation phenotypes and used PLINK to get $P$-values for SNPs. Then we applied LSMM and detect risk SNPs.

As the presence of LD effects, SNPs in a local genomic region would be correlated and detection of risk SNPs would be difficult. Because GWAS only aim to identify the local genomic region in LD with true risk genetic variants, it is reasonable to consider the identified SNPs not as false positives if they are in the flanking region of the true risk SNPs. Here we used different distance threshold to define the region around true risk SNPs. The identified risk SNPs which in the region of true risk SNPs were considered as true positives.

We considered four cases. The first case, no effects, means we only use the $P$-values and don’t use fixed effects and random effects. In the second case, fixed effects, we only add 10 fixed effects. In the fixed effects, SNPs within 1Mb of true risk SNPs are annotated with a probability of 0.6. In the third case, fixed + random effects,
Figure 2.28: FDR of LSMM, LFM and TGM with $K = 100$ (upper), $K = 500$ (middle), $K = 1000$ (lower) when P-values are not from beta distribution.
we further add 100 random effects in which SNPs are annotated randomly. In the fourth case, fixed + relevant random effects, we assume 20% of random effects are relevant to the phenotype and SNPs within 1Mb of true risk SNPs are annotated with a probability of 0.6 in the relevant random effects. The results of observed FDR were shown in Figure 2.29 based on 50 replications. In the first case, when we used no effects, the observed FDR was quite stable at 0.1. When we added fixed effects and random effects, the observed FDR was just inflated a little with the smallest distance threshold and became conservative as the distance threshold increased. As a result, we believe that LSMM can provide a satisfactory FDR control in detecting a local genomic region of risk SNPs.

Performance of LSMM when the proportion of risk SNPs $\pi_1$ is extremely small

We assume the proportion of risk variants is not very small due to the polygenic effect in the context of GWAS. In the simulation, we used the TGM to generate data such that we can evaluate whether the estimates converge to their true values. We set the number of SNPs $M = 100,000$ and varied the true value of the proportion of risk SNPs $\pi_1 \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.15, 0.2\}$. We also used Higher Criticism to estimate the proportion of non-null effects as a comparison. The software for Higher Criticism was downloaded from http://www.stat.cmu.edu/~jiashun/Research/software/NullandProp/. We controlled global FDR at 0.1 to evaluate empirical FDR.

The results based on 50 replications (Figure 2.30) show that when the true proportion of risk SNPs is extremely small (e.g., $\pi_1 \leq 0.001$ for $\alpha = 0.4$) and the signal of GWAS data is weak (e.g., $\pi_1 \leq 0.01$ for $\alpha = 0.6$), the estimation using LSMM is not very accurate. However, LSMM can still provide a valid FDR control. The performance of Higher Criticism is quite opposite. Although it can provide stable estimation when the true proportion of risk SNPs is small ($\pi_1 \leq 0.01$), its performance is not as well as LSMM when $\pi_1$ is relatively large, e.g., $\pi_1 \geq 0.05$. In the context of GWAS, the proportion of risk variants is not very small due to the polygenic effect. Therefore, we believe LSMM will work well in practice.
Figure 2.29: FDR of LSMM for identification of risk SNPs with different distance thresholds. The red line indicates the threshold of global FDR $\tau = 0.1$. 
Simulations based on probit model

In LSMM, we use generative model (2.2) to integrate functional annotations. Here we conducted simulations based on probit model:

\[ y_j = Z_j b + A_j \beta + e_j, \]

where \( e_j \sim N(0, \sigma^2_e) \). And we set \( \gamma_j = 1 \) if \( y_j > 0 \), \( \gamma_j = 0 \) if \( y_j \leq 0 \). The first entry of the coefficients of fixed effects \( b \), i.e. the intercept term, was fixed at \(-1\) and other entries were generated from \( N(0, 1) \) and fixed during multiple replications. We set \( \alpha = 0.2 \) and varied the signal-noise ratio \( r = \{4 : 1, 1 : 1, 1 : 4\} \). We controlled global FDR at 0.1 to evaluate empirical FDR and power.

The results based on 50 replications are provided in Figures 2.31-2.32. Noting that FDRs are all well-controlled at the nominal level. LSMM shows the best performance in power, AUC and partial AUC in identification of risk SNPs, and the advantages of LSMM over LFM and TGM are more noticeable as the signal-noise ratio increases.
Figure 2.31: Performance of LSMM, LFM and TGM for identification of risk SNPs based on probit model with $K = 100$ (upper), $K = 500$ (middle), $K = 1000$ (lower).
Figure 2.32: Performance of LSMM for detection of relevant annotations based on probit model with $K = 100$ (upper), $K = 500$ (middle), $K = 1000$ (lower).
Performance of LSMM when the random effects are correlated

The correlation among random effects is ignored in LSMM. Here we generated $\beta$ from a multivariate normal distribution $MVN(0, \Sigma)$, where $\Sigma$ is an autocorrelation matrix with $\rho$ varied at \{0, 0.2, 0.4, 0.6, 0.8\}. We set $\omega = 0.2$ and controlled global FDR at 0.1 to evaluate empirical FDR and power. The results based on 50 replications are shown in Figure 2.33 and Figure 2.34, indicating the robustness of LSMM.

Figure 2.33: Performance of LSMM, LFM and TGM for identification of risk SNPs when random effects are correlated with $\alpha = 0.2$ and $K = 500$.

Figure 2.34: Performance of LSMM for detection of relevant annotations when random effects are correlated with $\alpha = 0.2$ and $K = 500$.

Performance of LSMM when the random effects don’t share the same variance

LSMM assumes a common variance, $\sigma^2$, for random effects. Here we generated the variance for each random effect from $U[1, 10]$. We controlled global FDR at 0.1 to evaluate empirical FDR and power. The performance of LSMM based on 50 replications (Figures 2.35-2.36) is still comparable to the ideal case shown in Figure 2.1 and Figure 2.5.

In summary, the above simulation results suggest the robustness of LSMM and its potentially wide usage.
2.5.3 Comparison with other methods

Comparison with GPA

We compared LSMM with GPA in the identification of risk variants and detection of cell-type specific annotations. As LSMM can integrate both genic category and functional annotations, we compared GPA with LSMM without fixed effects (integrate functional annotations only) for a fair comparison. From the model setup, one main difference between GPA and LSMM is that GPA assumes conditional independence among annotations, whereas in LSMM we do not make this assumption.

To check the influence of correlated functional annotations, we simulated a case that the first 10 functional annotations were correlated and all the others were independent. We set $\alpha = 0.2$ and varied the correlation among annotations at $\{0, 0.2, 0.4, 0.6, 0.8\}$. To simulate the design matrices for correlated functional annotations, we first simulated $M$ samples from a multivariate normal distribution with the correlation matrix among annotations and then made a cutoff so that 10% of the entries would be 1 and the others be 0.

The results based on 50 replications are shown in Figure 2.37. We observe that
the empirical FDRs of LSMM and LSMM without fixed effects are indeed controlled at 0.1, but the FDR of GPA inflates very much when annotations are correlated. As the FDR of GPA is not controlled, the power of GPA is not comparable to the other two models. According to the AUC and partial AUC, the performance of GPA becomes worse as the correlation among annotations increase, while the performance of LSMM is still stable and outstanding. It implies that LSMM is able to identify true relevant annotations among correlated misleading ones.

Comparison with cmfdr

We also conducted simulations to compare LSMM with cmfdr [Zablocki et al., 2014] is a fully Bayesian approach to incorporate genic category annotations in GWAS using MCMC sampling algorithm. As cmfdr is not able to handle a large number of annotations and the MCMC sampling algorithm is very time-consuming, we set $M = 5000$, $L = 5$, $K = 5$ and run 2500 iterations with 2000 retained draws for cmfdr. Besides the computational time, we observe the empirical FDR of cmfdr is slightly inflated and its performance for prioritization of risk variants is inferior to LSMM in terms of AUC and partial AUC (See Figure 2.38).

Comparison with GenoWAP

As a comparison, we also used GenoWAP, a GWAS signal prioritization method that integrates genomic functional annotation and GWAS test statistics, to prioritize SNPs in our simulation. As GenoWAP can only integrate one annotation at a time, in the simulation we set $L = 1$ and let $Z$ be the functional annotation integrated using GenoWAP.

The performance of LSMM, LFM and GenoWAP for identification of risk SNPs are shown in Figure 2.39. We observe that for GenoWAP, the empirical FDRs are very conservative and its power, AUC and pAUC are all very low. That is because the SNPs that GenoWAP detects are disease-specific functional, i.e., only SNPs which are annotated to be functional have a chance be detected. In our simulation, 90% SNPs are not functional in the annotation category and thus are not identified by GenoWAP.
Figure 2.37: Performance of LSMM, LSMM without fixed effects and GPA for identification of risk SNPs with $K = 500$, $K = 100$, $K = 50$ and $K = 10$ (from the uppermost panel to the lowermost panel).
To evaluate the influence of the functional proportion on GenoWAP, we conducted the following simulations with no tissue-specific functional annotations. We set $L = 1$ and generated data from LFM. The results for different functional proportions are shown in Figure 2.40. The empirical FDRs of GenoWAP are still very small. Each of the power, AUC and pAUC of GenoWAP shows an increasing trend as the functional proportion increases, indicating that the performance of GenoWAP is influenced by the quality of annotation.


2.6 Real Data Analysis

We applied LSMM to analyze 30 GWASs of complex phenotypes. The source of the 30 GWASs is given in Table 2.2. We used ANNOVAR [Wang et al., 2010] to provide the genic category annotations: upstream, downstream, exonic, intergenic, intronic, ncRNA_exonic, ncRNA_intronic, UTR3 and UTR5, where ncRNA means variant overlaps a transcript without coding annotation in the gene definition. We obtained 127 cell-type specific functional annotations from GenoSkylinePlus [Lu et al., 2017b] (http://genocanyon.med.yale.edu/GenoSkyline). To avoid unusually large GWAS signals in the MHC region (Chromosome 6, 25Mb - 35Mb), we excluded the SNPs in this region.

We compared the number of identified risk SNPs using TGM, LFM and LSMM for 30 GWASs. Using LSMM as a reference, we calculated the ratio of the number of risk SNPs each method identified to that from LSMM under FDR thresholds $\tau = 0.05$ and $\tau = 0.1$. The results are shown in Figure 2.41. For detecting the relevant cell-type specific functional annotations, we controlled the local fdr at 0.1. Figure 2.42 shows the approximated posterior probability for annotations and phenotypes, where the darkness of the red entry implies the level of relevance between the corresponding cell-type specific functional annotation and the phenotype, the darker the more relevant.

Figure 2.41 shows that LSMM can identify more risk variants than TGM and LFM, under the same level of FDR control. The differences between TGM and LFM are due to the impact of genic category annotations and the differences between LFM and LSMM can be attributed to cell-type specific functional annotations. For HIV [McLaren et al., 2013] and bipolar disorder [Psychiatric GWAS Consortium Bipolar Disorder Working Group, 2011], a clear improvement in the identification of risk SNPs can be found from TGM to LFM, reflecting a large enrichment of genic category annotations. The contribution of cell-type specific annotations can be clearly seen with the improvement from LFM to LSMM in several GWAS analyses, such as multiple sclerosis [Sawcer et al., 2011] and coronary artery disease (CAD) [Schunkert et al., 2011]. For multiple sclerosis, genic category annotations do not show huge contributions, however, the contributions of cell-type specific annotations are substantial. As shown in Figure 2.42 its relevant cell-type specific annotations
Table 2.2: The source of the 30 GWASs.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Study/Group</th>
<th>Year</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alzheimer</td>
<td>Lambert et al.</td>
<td>2013</td>
<td>Nature Genetics</td>
</tr>
<tr>
<td>BMI</td>
<td>Speliotes et al.</td>
<td>2010</td>
<td>Nature Genetics</td>
</tr>
<tr>
<td>Bipolar Disorder</td>
<td>Psychiatric GWAS Consortium Bipolar Disorder Working Group</td>
<td>2011</td>
<td>Nature Genetics</td>
</tr>
<tr>
<td>Coronary Artery Disease</td>
<td>Schunkert et al.</td>
<td>2011</td>
<td>Nature Genetics</td>
</tr>
<tr>
<td>Crohn’s Disease</td>
<td>Jostins et al.</td>
<td>2012</td>
<td>Nature</td>
</tr>
<tr>
<td>Height</td>
<td>Wood et al.</td>
<td>2014</td>
<td>Nature Genetics</td>
</tr>
<tr>
<td>High-density Lipoprotein</td>
<td>Global Lipids Genetics Consortium</td>
<td>2013</td>
<td>Nature Genetics</td>
</tr>
<tr>
<td>HIV</td>
<td>McLaren et al.</td>
<td>2013</td>
<td>PLoS Pathogens</td>
</tr>
<tr>
<td>Inflammatory Bowel Disease</td>
<td>Jostins et al.</td>
<td>2012</td>
<td>Nature</td>
</tr>
<tr>
<td>Low-density Lipoprotein</td>
<td>Global Lipids Genetics Consortium</td>
<td>2013</td>
<td>Nature Genetics</td>
</tr>
<tr>
<td>Lupus</td>
<td>Bentham et al.</td>
<td>2015</td>
<td>Nature Genetics</td>
</tr>
<tr>
<td>Mean Cell Haemoglobin</td>
<td>Pickrell</td>
<td>2014</td>
<td>The American Journal of Human Genetics.</td>
</tr>
<tr>
<td>Menopause</td>
<td>Day et al.</td>
<td>2015</td>
<td>Nature Genetics</td>
</tr>
<tr>
<td>Multiple Sclerosis</td>
<td>Sawcer et al.</td>
<td>2011</td>
<td>Nature</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>Okbay et al.</td>
<td>2016a</td>
<td>Nature Genetics</td>
</tr>
<tr>
<td>Primary Biliary Cirrhosis</td>
<td>Cordell et al.</td>
<td>2015</td>
<td>Nature Communications.</td>
</tr>
<tr>
<td>Rheumatoid Arthritis</td>
<td>Okada et al.</td>
<td>2014</td>
<td>Nature</td>
</tr>
<tr>
<td>Schizophrenia1 (SCZ subset)</td>
<td>Cross-Disorder Group of the Psychiatric Genomics Consortium</td>
<td>2013b</td>
<td>The Lancet.</td>
</tr>
<tr>
<td>Schizophrenia2 (SCZ1)</td>
<td>Schizophrenia Psychiatric GWAS Consortium</td>
<td>2011</td>
<td>Nature Genetics</td>
</tr>
<tr>
<td>Schizophrenia3 (Sweden+SCZ1)</td>
<td>Ripke et al.</td>
<td>2013</td>
<td>Nature Genetics</td>
</tr>
<tr>
<td>Schizophrenia4 (SCZ2)</td>
<td>Ripke et al.</td>
<td>2014</td>
<td>Nature</td>
</tr>
<tr>
<td>Total Cholesterol</td>
<td>Global Lipids Genetics Consortium</td>
<td>2013</td>
<td>Nature Genetics</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>Global Lipids Genetics Consortium</td>
<td>2013</td>
<td>Nature Genetics</td>
</tr>
<tr>
<td>Type 1 Diabetes</td>
<td>Bradfield et al.</td>
<td>2011</td>
<td>PLoS Genetics</td>
</tr>
<tr>
<td>Type 2 Diabetes</td>
<td>Morris et al.</td>
<td>2012</td>
<td>Nature Genetics</td>
</tr>
<tr>
<td>Ulcerative Colitis</td>
<td>Jostins et al.</td>
<td>2012</td>
<td>Nature</td>
</tr>
<tr>
<td>Years of Education1</td>
<td>Rietveld et al.</td>
<td>2013</td>
<td>Science.</td>
</tr>
<tr>
<td>Years of Education2</td>
<td>Okbay et al.</td>
<td>2016b</td>
<td>Nature</td>
</tr>
</tbody>
</table>
Figure 2.41: The number of risk variants identified by TGM, LFM and LSMM for 30 GWASs, under the same level of global FDR control (0.05 and 0.1). For visualization purpose, these numbers are normalized by dividing the corresponding number of variants identified by LSMM.
Figure 2.42: Relevant cell-type specific functional annotations for 30 GWASs.
are related with immune system, GM12878 lymphoblastoid cells and primary B cells from peripheral blood. For CAD, both enrichment of genic category and cell-type specific annotations are estimated and its relevant cells are from a few different tissues, including blood, heart, lung and skin (See Figure 2.42). As a cardiovascular disease, it is reasonable to discover the relevance of these cells to CAD, and Fernández-Ruiz 2016 has shown its relationship with immune system. The annotations in lung and skin we detected may provide some new insights about the disease.

Among the 30 GWASs, we analyzed four GWASs of schizophrenia with different sample sizes, Schizophrenia1 (9,379 cases and 7,736 controls) [Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013], Schizophrenia2 (9,394 cases and 12,462 controls) [Schizophrenia Psychiatric GWAS Consortium, 2011], Schizophrenia3 (13,833 cases and 18,310 controls) [Ripke et al., 2013] and Schizophrenia4 (36,989 cases and 113,075 controls) [Ripke et al., 2014]. The detailed results are summarized in Table 2.3. The Manhattan plots using TGM and LSMM are provided in Figure 2.43. Clearly, LSMM steadily improves over TGM and LFM in the analysis of schizophrenia, a highly polygenic trait, with different sample sizes. In particular, for Schizophrenia3, LSMM identified 1,492 risk variants which could not be identified by TGM. Interestingly, the majority of them (872 variants) can be re-identified in Schizophrenia4 using TGM. This indicates that LSMM has a better power in prioritizing risk variants than TGM. For Schizophrenia4, four cell-type specific functional annotations are detected. In our analysis, both genetic variants related to functions of brain cells (brain angular gyrus) and blood cells (K562 leukemia cells) are detected to be relevant. This evidence not only connects Schizophrenia with brain, but also suggests the biological link between Schizophrenia and immune system [Ripke et al., 2014]. To make a comparison, we also used GenoWAP to analyze Schizophrenia3 and Schizophrenia4 by integrating each of the 9 genic category annotations. The results are shown in Table 2.4. With the nominal local FDR controlled at 0.1, even we collected the risk SNPs identified by integrating every annotation using GenoWAP, the total number is still much less than TGM, LFM and LSMM, suggesting GenoWAP is too conservative for real data analysis. We also analyzed two GWAS of years of education, Years of Education 1 [Rietveld et al., 2013] and Years of Education 2
Okbay et al. 2016b. Compared with Years of Education 1, the GWAS data set for Years of Education 2 is based on a larger sample size, and thus it enables LSMM to detect relevant functional annotations in brain and immune system. Our results are consistent with Finucane et al. 2015.

Table 2.3: Summary of results for Schizophrenia.

<table>
<thead>
<tr>
<th></th>
<th>( \hat{\alpha} )</th>
<th>No. of risk SNPs</th>
<th>Bonferroni correction</th>
<th>TGM</th>
<th>LFM</th>
<th>LSMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schizophrenia1</td>
<td>0.677</td>
<td>2</td>
<td>470</td>
<td>527</td>
<td>527</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia2</td>
<td>0.633</td>
<td>7</td>
<td>2,107</td>
<td>2,404</td>
<td>2,405</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia3</td>
<td>0.562</td>
<td>126</td>
<td>6,811</td>
<td>7,541</td>
<td>7,545</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia4</td>
<td>0.413</td>
<td>1110</td>
<td>48,802</td>
<td>50,481</td>
<td>50,990</td>
<td></td>
</tr>
</tbody>
</table>

a. The estimate \( \hat{\alpha} \) is obtained using LSMM.
b. The number of risk SNPs is reported based on global \( FDR \leq 0.1 \).

More findings about the relevance between cell-type specific annotations and GWAS are shown in Figure 2.42. Some are concordant with previous GWAS analyses. For example, we detect the functional annotation in liver to be relevant to the lipid-related phenotypes, including low-density lipoprotein, high-density lipoprotein, triglycerides and total cholesterol [Global Lipids Genetics Consortium, 2013]. Similar functional enrichment has been found by Finucane et al. 2015; Kundaje et al. 2015 and Lu et al. 2017b. For height Wood et al. 2014, more than 40 cell-type spe-
specific functional annotations are detected to be relevant using LSMM, which reflects its highly polygenic genetic architecture. These relevant annotations include cells in bone, vascular and skeletal muscle which were also shown significant enrichments for height by Finucane et al. 2015. Recent research has linked some neurodegenerative diseases, which were believed to be more related to brain and neural system, to the immune system, such as Alzheimer’s disease Sims et al. 2017 and Parkinson’s disease Sulzer et al. 2017. For Alzheimer’s disease Lambert et al. 2013, similar results have been found using LSMM. The relevant functional annotations are from blood cells, including monocytes-CD14+ and K562 leukemia cells. For autoimmune diseases including Crohn’s disease Jostins et al. 2012, ulcerative colitis Jostins et al. 2012, inflammatory bowel disease Jostins et al. 2012, rheumatoid arthritis Okada et al. 2014, lupus Bentham et al. 2015, menopause Day et al. 2015, multiple sclerosis Sawcer et al. 2011 and primary biliary cirrhosis Cordell et al. 2015, the detected relevant functional annotations are mainly from the immune system and have many overlaps. Our results also provide the genomic level supports to previous medical literature, such as the relevance between spleen and inflammatory bowel disease Muller et al. 1993, between liver and menopause Mucci et al. 2001. The result also provides several new insights. Lipid-related phenotypes including high-density lipoprotein and total cholesterol are also relevant to functional annota-
tions in immune system and brain. Additionally, annotations in immune system are considered relevant to blood-related phenotypes including red cell count, mean cell haemoglobin and mean cell volume [Pickrell, 2014]. The foreskin fibroblast primary cells in skin are relevant to ulcerative colitis, four lipid-related phenotypes and red cell count.

Regarding the computational time, LSMM takes less than six minutes to handle each of the 30 GWAS datasets. We also recorded timings of cmfdr as a comparison. As cmfdr is not scalable to a large number of covariates, we only integrated the 9 genic category annotations in cmfdr. The MCMC algorithm was suggested [Zablocki et al., 2014] to run with 5,000 burn-in and 20,000 main iterations. According to our estimates, cmfdr takes more than ten days for most phenotypes. The detailed timing results are shown in Figure 2.44.

Figure 2.44: Computational time using LSMM and cmfdr for 30 GWASs.

If we did not adjust the genic category annotation, more relevant cell-type specific functional annotations would be detected (see Figure 2.45). This indicates that LSMM could adjust covariates’ effects and provide a more reliable identification of relevant functional annotations.
Figure 2.45: Relevant functional annotations for 30 GWASs without integrating genic category annotations.
Chapter 3

LPM: a latent probit model to characterize relationship among complex traits using summary statistics from multiple GWASs and functional annotations

3.1 Introduction

In this chapter, we propose a latent probit model (LPM) to characterize relationship among complex traits by integrating summary statistics from multiple GWASs and genic category annotations. LPM can not only fulfill the three goals (characterizing relationship, prioritizing SNPs and integrating functional annotations) under a unified framework but also achieve a better performance than conventional methods.

We first described the model in Section 3.2, introduced the efficient algorithm based on PX-EM in Section 3.3, methods of statistical inference in Section 3.4 and related theorem in Section 3.5. Then we conducted comprehensive simulations to investigate the properties of LPM in Section 3.6. We also analyzed 44 GWASs of complex traits with nine genic category annotations using LPM in Section 3.7.
3.2 Model

Suppose we have the summary statistics ($p$-values) of $M$ SNPs from $K$ GWASs. In this paper, we use $j = 1, \ldots, M$ to index SNPs and $k = 1, \ldots, K$ to index GWAS data sets. For each GWAS, we consider the $p$-values follow the two-groups model, i.e., a mixture of null and non-null distribution, and introduce a latent variable $\eta_{jk}$ to indicate which group the $j$-th SNP belongs to for the $k$-th GWAS. Here $\eta_{jk} = 0$ and $\eta_{jk} = 1$ indicates the $j$-th SNP is un-associated (in the null group) and associated (in the non-null group) with the $k$-th trait, respectively. We assume the $p$-values from the $k$-th GWAS follow the following distribution:

$$P_{jk} \sim \begin{cases} U[0, 1], & \eta_{jk} = 0, \\ Beta(\alpha_k, 1), & \eta_{jk} = 1, \end{cases}$$

where $U[0, 1]$ is the uniform distribution on $[0, 1]$ and $Beta(\alpha_k, 1)$ is the beta distribution with the constrain $0 < \alpha_k < 1$. This model is designed to capture the pattern that $p$-values from the non-null group are more likely to be close to 0.

To adjust the effect of genic category annotations and model the relationship among traits, we consider the following latent probit model (LPM):

$$\eta_{jk} = \begin{cases} 1, & \text{if } Z_{jk} > 0, \\ 0, & \text{if } Z_{jk} \leq 0, \end{cases}$$

(3.1)

with

$$Z_j = \beta X_j + \epsilon_j,$$

(3.2)

and

$$\epsilon_j \sim N(0, R),$$

(3.3)

where $Z \in \mathbb{R}^{M \times K}$ is the latent variable in probit model, $X \in \mathbb{R}^{M \times (D+1)}$ is the design matrix of functional annotations, comprised of an intercept and $D$ annotations, $\beta \in \mathbb{R}^{K \times (D+1)}$ is a matrix of the coefficients. For the $j$-th SNP, $Z_j^T$, $X_j^T$ and $\epsilon_j^T$ are correspond to the $j$-th row of $Z$, $X$ and $\epsilon$, respectively. $R \in \mathbb{R}^{K \times K}$ is the correlation matrix measuring the relationship among the $K$ traits. When the relationship be-
tween any two traits exists, the corresponding entry in \( R \) is expected to differ from 0. Let \( \theta = \{ \alpha, \beta, R \} \) be the collection of model parameters.

We analyze the GWASs pairwise. In this case, \( K = 2 \) and we denote this model as bLPM (bivariate LPM).

\[
\tilde{P}_{jk} \sim \begin{cases} 
U[0,1], & \tilde{\eta}_{jk} = 0, \\
Beta(\tilde{\alpha}_k, 1), & \tilde{\eta}_{jk} = 1,
\end{cases}
\]

\[
\tilde{\eta}_{jk} = \begin{cases} 
1, & \text{if } \tilde{Z}_{jk} > 0, \\
0, & \text{if } \tilde{Z}_{jk} \leq 0,
\end{cases}
\]

with

\[
\tilde{Z}_j = \tilde{\beta}X_j + \tilde{\epsilon}_j, \tag{3.4}
\]

and

\[
\tilde{\epsilon}_j \sim N(0, \tilde{R}), \tag{3.5}
\]

where \( k = 1, 2, \tilde{\beta} \in \mathbb{R}^{2 \times (D+1)} \) and \( \tilde{R} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \).

Let \( \hat{\theta} = \{ \hat{\alpha}, \tilde{\beta}, \tilde{R} \} \) be the collection of model parameters. The logarithm of the marginal likelihood can be written as

\[
\log \Pr \left( \tilde{P} | X; \hat{\theta} \right) = \log \sum_{\eta} \int \Pr \left( \tilde{P}, \tilde{\eta}, \tilde{Z} | X; \theta \right) d\tilde{Z}, \tag{3.6}
\]

where

\[
\Pr \left( \tilde{P}, \tilde{\eta}, \tilde{Z} | X; \hat{\theta} \right) = \Pr \left( \tilde{P} | \tilde{\eta}; \hat{\alpha} \right) \Pr \left( \tilde{Z} | X; \tilde{\beta}, \tilde{R} \right)
\]

and

\[
\Pr \left( \tilde{P} | \tilde{\eta}; \hat{\alpha} \right) = \prod_{j=1}^{M} \prod_{k=1}^{2} \left( \hat{\alpha}_k \tilde{P}_{jk}^{\hat{\alpha}_k - 1} \right)^{\tilde{\eta}_{jk}}
\]

\[
\Pr \left( \tilde{Z} | X; \tilde{\beta}, \tilde{R} \right) = \prod_{j=1}^{M} N(\tilde{Z}_j; \tilde{\beta}X_j, \tilde{R})
\]
Our goal is to obtain the estimation of $\tilde{\theta}$ which maximizes the marginal likelihood (3.6) for each pair of GWASs and calculate the estimation of $\theta$ in LPM. Under the estimation $\hat{\theta}$, we can compute the following posterior:

$$\Pr(\eta|P,X;\hat{\theta}) = \frac{\Pr(P,\eta|X;\hat{\theta})}{\Pr(P|X;\hat{\theta})}.$$ 

Then we can make statistical inference on the association of SNPs and the relationship among traits.

The details of the algorithm are provided in Section 3.3.

### 3.3 Algorithm

#### 3.3.1 The PX-EM Algorithm

Instead of using the EM algorithm, we propose a PX-EM algorithm, which dominates EM in global rate of convergence, for parameter estimation in bivariate LPM. We expand the parameter to $\Theta = \{\tilde{\alpha}, \gamma, \Sigma\}$ and accordingly model (3.4) and (3.5) are expanded to

$$\tilde{Z}_j = \gamma X_j + \tilde{\epsilon}_j,$$

and

$$\tilde{\epsilon}_j \sim N(0, \Sigma),$$

where $\gamma = D\tilde{\beta}$, $\Sigma = D\tilde{\Sigma}D = \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix}$ and $D = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}$ is the auxiliary parameter whose value is fixed at $I_2$ in the original model.
For the expanded model, the complete-data log-likelihood can be written as

\[
\log \Pr \left( \tilde{P}, \tilde{\eta}, \tilde{Z} | X; \Theta \right)
= \log \Pr \left( \tilde{P} | \tilde{\eta}; \tilde{\alpha} \right) + \log \Pr \left( \tilde{Z} | X; \gamma, \Sigma \right)
= \sum_{j=1}^{M} \sum_{k=1}^{2} \tilde{\eta}_{jk} \left( \log \tilde{\alpha}_k + (\tilde{\alpha}_k - 1) \log \tilde{P}_{jk} \right)
+ \sum_{j=1}^{M} \left[ -\frac{1}{2} \log (2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} \left( \tilde{Z}_j - \gamma X_j \right)^T \Sigma^{-1} \left( \tilde{Z}_j - \gamma X_j \right) \right].
\]

The Q function is evaluated as

\[
Q = E_{\tilde{\eta}, \tilde{Z}} \log \Pr \left( \tilde{P}, \tilde{\eta}, \tilde{Z} | X; \Theta \right)
= \sum_{j=1}^{M} \sum_{k=1}^{2} E \left[ \tilde{\eta}_{jk} \right] \left( \log \tilde{\alpha}_k + (\tilde{\alpha}_k - 1) \log \tilde{P}_{jk} \right)
+ \sum_{j=1}^{M} \left( -\frac{1}{2} \log (2\pi) - \frac{1}{2} \log |\Sigma| \right)
+ \sum_{j=1}^{M} \left[ -\frac{1}{2} \text{trace} \left( E \left[ \tilde{Z}_j \tilde{Z}_j^T \right] \Sigma^{-1} \right)
+ E \left[ \tilde{Z}_j^T \right] \Sigma^{-1} \gamma X_j
- \frac{1}{2} X_j^T \gamma^T \Sigma^{-1} \gamma X_j \right],
\]

where the expectation is calculated based on the current \( \tilde{\theta} \) in the original model.

In the PX-E step, we calculate \( E \left[ \tilde{\eta} \right] \) and \( E \left[ \tilde{Z} \right] \) based on the original model.

To obtain the expectation of \( \tilde{\eta} \), we calculate the posterior of \( \tilde{\eta} \) in the original model:

\[
\pi_{j11} = \Pr \left( \tilde{\eta}_{j1} = 1, \tilde{\eta}_{j2} = 1 | \tilde{P}, X; \tilde{\theta} \right) = \frac{\Phi_{j11} \tilde{\alpha}_1 \tilde{P}_{\tilde{\alpha}_1 - 1} \tilde{P}_{\tilde{\alpha}_2 - 1}}{\Phi_j},
\]

\[
\pi_{j10} = \Pr \left( \tilde{\eta}_{j1} = 1, \tilde{\eta}_{j2} = 0 | \tilde{P}, X; \tilde{\theta} \right) = \frac{\Phi_{j10} \tilde{\alpha}_1 \tilde{P}_{\tilde{\alpha}_1 - 1}}{\Phi_j},
\]

\[
\pi_{j01} = \Pr \left( \tilde{\eta}_{j1} = 0, \tilde{\eta}_{j2} = 1 | \tilde{P}, X; \tilde{\theta} \right) = \frac{\Phi_{j01} \tilde{\alpha}_2 \tilde{P}_{\tilde{\alpha}_2 - 1}}{\Phi_j},
\]

\[
\pi_{j00} = \Pr \left( \tilde{\eta}_{j1} = 0, \tilde{\eta}_{j2} = 0 | \tilde{P}, X; \tilde{\theta} \right) = \frac{\Phi_{j00}}{\Phi_j},
\]
where

\[ \Phi_{j11} = \Phi_2 \left( \begin{pmatrix} \hat{\beta}_1^T X_j \\ \hat{\beta}_2^T X_j \end{pmatrix}, \rho \right), \]

\[ \Phi_{j10} = \Phi_2 \left( \begin{pmatrix} \hat{\beta}_1^T X_j \\ -\hat{\beta}_2^T X_j \end{pmatrix}, -\rho \right) = -\Phi_{j11} + \Phi \left( \begin{pmatrix} \hat{\beta}_1^T X_j \end{pmatrix}, \right), \]

\[ \Phi_{j01} = \Phi_2 \left( \begin{pmatrix} -\hat{\beta}_1^T X_j \\ \hat{\beta}_2^T X_j \end{pmatrix}, -\rho \right) = -\Phi_{j11} + \Phi \left( \begin{pmatrix} \hat{\beta}_2^T X_j \end{pmatrix}, \right), \]

\[ \Phi_{j00} = \Phi_2 \left( \begin{pmatrix} -\hat{\beta}_1^T X_j \\ -\hat{\beta}_2^T X_j \end{pmatrix}, \rho \right) = 1 + \Phi_{j11} - \Phi \left( \begin{pmatrix} \hat{\beta}_1^T X_j \end{pmatrix}, \right) - \Phi \left( \begin{pmatrix} \hat{\beta}_2^T X_j \end{pmatrix}, \right), \]

\[ \Phi_j = \Phi_{j11} \tilde{\alpha}_1 \tilde{P}_{j1} \tilde{\alpha}_1^{-1} \tilde{\alpha}_2 \tilde{P}_{j2} \tilde{\alpha}_2^{-1} + \Phi_{j10} \tilde{\alpha}_1 \tilde{P}_{j1} \tilde{\alpha}_1^{-1} + \Phi_{j01} \tilde{\alpha}_2 \tilde{P}_{j2} \tilde{\alpha}_2^{-1} + \Phi_{j00}, \]

and \( \Phi (\cdot) \) is the univariate standard normal cumulative distribution function,

\[ \Phi_2 \left( \begin{pmatrix} \hat{\beta}_1^T X_j \\ \hat{\beta}_2^T X_j \end{pmatrix}, \rho \right) = \int_{-\infty}^{\hat{\beta}_1^T X_j} \int_{-\infty}^{\hat{\beta}_2^T X_j} \phi (x_1, x_2, \rho) \, dx_1 \, dx_2, \]

\[ \phi (x_1, x_2, \rho) = \frac{\exp \left[ -\frac{x_1^2 + x_2^2 - 2\rho x_1 x_2}{2(1-\rho^2)} \right]}{2\pi \sqrt{1-\rho^2}}. \]

Therefore, we have

\[ E [\tilde{\eta}_{j1}] = \frac{\Phi_{j11} \tilde{\alpha}_1 \tilde{P}_{j1} \tilde{\alpha}_1^{-1} \tilde{\alpha}_2 \tilde{P}_{j2} \tilde{\alpha}_2^{-1} + \Phi_{j10} \tilde{\alpha}_1 \tilde{P}_{j1} \tilde{\alpha}_1^{-1}}{\Phi_j}, \]

\[ E [\tilde{\eta}_{j2}] = \frac{\Phi_{j11} \tilde{\alpha}_1 \tilde{P}_{j1} \tilde{\alpha}_1^{-1} \tilde{\alpha}_2 \tilde{P}_{j2} \tilde{\alpha}_2^{-1} + \Phi_{j01} \tilde{\alpha}_2 \tilde{P}_{j2} \tilde{\alpha}_2^{-1}}{\Phi_j}. \]

The logarithm of the likelihood can be evaluated by

\[ L = \sum_{j=1}^{M} \log \Phi_j. \]

The posterior of \( \tilde{Z}_j \) given \( \tilde{\eta}_j \) is a truncated bivariate normal distribution. Using the properties of truncated bivariate normal distribution, we can get

\[ E \left( \tilde{Z}_{j1} \right) = \tilde{\beta}_1^T X_j + \frac{\Phi_{j11}}{\Phi_j} E \left( \tilde{\epsilon}_{j1} | \tilde{\eta}_{j1} = 1, \tilde{\eta}_{j2} = 1, X_j; \hat{\theta} \right) \]

\[ + \frac{\phi \left( \tilde{\beta}_1^T X_j \right)}{\Phi_j} \left( \hat{\alpha}_1 \tilde{P}_{j1} \hat{\alpha}_1^{-1} - 1 \right) + \frac{\rho \phi \left( \tilde{\beta}_2^T X_j \right)}{\Phi_j} \left( \hat{\alpha}_2 \tilde{P}_{j2} \hat{\alpha}_2^{-1} - 1 \right). \]
\[ E\left(\tilde{Z}_{j2}\right) = \tilde{\beta}_2^T X_j + \frac{\Phi_{j11}}{\Phi_j} \xi_j E\left(\tilde{\varepsilon}_{j2|\bar{\eta}_{j1}} = 1, \bar{\eta}_{j2} = 1, X_j; \tilde{\theta}\right) \]
\[ + \frac{\rho j_{11} \xi_j}{\Phi_j} \left( E\left(\tilde{\varepsilon}_{j1|\bar{\eta}_{j1}} = 1, \bar{\eta}_{j2} = 1, X_j; \tilde{\theta}\right) - 1 \right) \]
\[ + \frac{\tilde{\beta}_1^T X_j \phi \left( \tilde{\beta}_1^T X_j \right)}{\Phi_j} \left( \tilde{\alpha}_1 \hat{P}_{\tilde{\alpha}_1}^{-1} - 1 \right) \]
\[ + \frac{\rho \tilde{\beta}_2^T X_j \phi \left( \tilde{\beta}_2^T X_j \right)}{\Phi_j} \left( \tilde{\alpha}_2 \hat{P}_{\tilde{\alpha}_2}^{-1} - 1 \right) \].

\[ E\left(\tilde{Z}_{j1} \tilde{Z}_{j2}\right) = - \tilde{\beta}_1^T X_j \tilde{\beta}_2^T X_j + \tilde{\beta}_1^T X_j E\left(\tilde{Z}_{j2}\right) + \tilde{\beta}_2^T X_j E\left(\tilde{Z}_{j1}\right) + \rho \]
\[ + \frac{\Phi_{j11} \xi_j}{\Phi_j} \left( E\left(\tilde{\varepsilon}_{j1|\bar{\eta}_{j1}} = 1, \bar{\eta}_{j2} = 1, X_j; \tilde{\theta}\right) - \rho \right) \]
\[ - \frac{\rho \tilde{\beta}_1^T X_j \phi \left( \tilde{\beta}_1^T X_j \right)}{\Phi_j} \left( \tilde{\alpha}_1 \hat{P}_{\tilde{\alpha}_1}^{-1} - 1 \right) \]
\[ - \frac{\rho \tilde{\beta}_2^T X_j \phi \left( \tilde{\beta}_2^T X_j \right)}{\Phi_j} \left( \tilde{\alpha}_2 \hat{P}_{\tilde{\alpha}_2}^{-1} - 1 \right) \].

where

\[ \xi_j = \tilde{\alpha}_1 \hat{P}_{\tilde{\alpha}_1}^{-1} \tilde{\alpha}_2 \hat{P}_{\tilde{\alpha}_2}^{-1} - \tilde{\alpha}_1 \hat{P}_{\tilde{\alpha}_1}^{-1} - \tilde{\alpha}_2 \hat{P}_{\tilde{\alpha}_2}^{-1} + 1, \]

\[ E\left(\tilde{\varepsilon}_{j1|\bar{\eta}_{j1}} = 1, \bar{\eta}_{j2} = 1, X_j; \tilde{\theta}\right) = \frac{1}{\Phi_{j11}} \phi \left( \tilde{\beta}_1^T X_j \right) \Phi \left( \left( \tilde{\beta}_2^T X_j - \rho \tilde{\beta}_1^T X_j \right) c \right) \]
\[ + \frac{1}{\Phi_{j11}} \rho \phi \left( \tilde{\beta}_2^T X_j \right) \Phi \left( \left( \tilde{\beta}_1^T X_j - \rho \tilde{\beta}_2^T X_j \right) c \right), \]
\[ E(\tilde{\epsilon}_{j2}|\tilde{\eta}_{j1} = 1, \tilde{\eta}_{j2} = 1, X_j; \tilde{\theta}) = \frac{1}{\phi_{j11}} \phi(\tilde{\beta}_2^T X_j) \Phi\left( (\tilde{\beta}_2^T X_j - \rho \tilde{\beta}_1^T X_j) c \right) \]
\[ + \frac{1}{\phi_{j11}} \rho \phi(\tilde{\beta}_1^T X_j) \Phi\left( (\tilde{\beta}_1^T X_j - \rho \tilde{\beta}_1^T X_j) c \right), \]

\[ E(\tilde{\epsilon}_{j1}|\tilde{\eta}_{j1} = 1, \tilde{\eta}_{j2} = 1, X_j; \tilde{\theta}) = 1 - \frac{1}{\phi_{j11}} \tilde{\beta}_1^T X_j \phi(\tilde{\beta}_1^T X_j) \Phi\left( (\tilde{\beta}_2^T X_j - \rho \tilde{\beta}_1^T X_j) c \right) \]
\[ - \frac{1}{\phi_{j11}} \rho^2 \tilde{\beta}_2^T X_j \phi(\tilde{\beta}_2^T X_j) \Phi\left( (\tilde{\beta}_2^T X_j - \rho \tilde{\beta}_1^T X_j) c \right) \]
\[ + \frac{1}{\phi_{j11}} c^{-1} \rho \phi(\tilde{\beta}_2^T X_j) \phi\left( (\tilde{\beta}_2^T X_j - \rho \tilde{\beta}_2^T X_j) c \right), \]

\[ E(\tilde{\epsilon}_{j2}|\tilde{\eta}_{j1} = 1, \tilde{\eta}_{j2} = 1, X_j; \tilde{\theta}) = 1 - \frac{1}{\phi_{j11}} \tilde{\beta}_2^T X_j \phi(\tilde{\beta}_2^T X_j) \Phi\left( (\tilde{\beta}_2^T X_j - \rho \tilde{\beta}_1^T X_j) c \right) \]
\[ - \frac{1}{\phi_{j11}} \rho^2 \tilde{\beta}_1^T X_j \phi(\tilde{\beta}_1^T X_j) \Phi\left( (\tilde{\beta}_2^T X_j - \rho \tilde{\beta}_2^T X_j) c \right) \]
\[ + \frac{1}{\phi_{j11}} c^{-1} \rho \phi(\tilde{\beta}_1^T X_j) \phi\left( (\tilde{\beta}_1^T X_j - \rho \tilde{\beta}_1^T X_j) c \right), \]

\[ \rho = \frac{1}{\sqrt{1 - \rho^2}}, \]

and \( \phi(\cdot) \) is the standard univariate normal density.

In the PX-M step, we maximize the Q function with respect to \( \Theta \) and obtain the updating equations

\[ \hat{\alpha}_k = - \frac{\sum_{j=1}^{M} E[\tilde{\eta}_{jk}]}{\sum_{j=1}^{M} E[\tilde{\eta}_{jk}] \log P_{jk}}, \]
\[ \gamma = E\left[ \tilde{Z}^T \right] X \left( X^T X \right)^{-1}, \]
\[ \Sigma = \frac{1}{M} \left[ \left( \sum_{j=1}^{M} E[\tilde{Z}_j \tilde{Z}_j^T] \right) - \gamma X^T E \left[ \tilde{Z} \right] \right]. \]
Then we apply the reduction function to obtain the original parameters:

\[
\tilde{\beta} = D^{-1}\gamma, \\
\tilde{R} = D^{-1}\Sigma D^{-1}.
\]

When the correlation coefficient \( \rho \) is zero, we can analyze the phenotypes independently, which provides warm starts for generating our three-stage algorithm. In the first stage, we suppose all the coefficients in \( \tilde{\beta} \) (except the intercept term) and the correlation coefficient \( \rho \) are zero and run an EM algorithm to obtain the parameter \( \tilde{\alpha} \) and \( \tilde{\beta}_0 \). Then in the second stage, we use the estimated parameters as the starting point to obtain \( \tilde{\alpha} \) and \( \tilde{\beta} \) using a PX-EM algorithm. Finally, in the third stage, we run the above PX-EM algorithm with the obtained parameters in the second stage and update \( \tilde{\alpha}, \tilde{\beta} \) and \( \rho \) simultaneously until convergence. Since our algorithm is based on the framework of EM and PX-EM algorithm, the logarithm of the likelihood is guaranteed to increase at each iteration. The details of the three-stage algorithm are provided in the Section 3.3.2.

### 3.3.2 Details of the Three-stage Algorithm

**Stage 1**

Suppose all the coefficients in \( \tilde{\beta} \) (except the intercept term) and the correlation coefficient \( \rho \) are zero, then the model becomes

\[
\tilde{P}_{jk} \sim \begin{cases} 
U[0, 1], & \tilde{\eta}_{jk} = 0, \\
\text{Beta} (\tilde{\alpha}_k, 1), & \tilde{\eta}_{jk} = 1,
\end{cases}
\]

\[
\pi_{1k} = \Pr (\tilde{\eta}_{jk} = 1),
\]

where \( \pi_{1k} = 1 - \Phi (\bar{\beta}_{k0}) \), \( \bar{\beta}_{k0} \) is the intercept term in the probit model.

We can use EM algorithm to estimate the parameters \( \tilde{\alpha}_k \) and \( \pi_{1k} \).
The complete-data log-likelihood is
\[
\log \Pr (\tilde{P}, \tilde{\eta} | \tilde{\alpha}, \pi_1)
\]
\[
= \log \Pr \left( \tilde{P} | \tilde{\eta}; \tilde{\alpha} \right) + \log \Pr \left( \tilde{\eta} | \pi_1 \right)
\]
\[
= \sum_{j=1}^{M} \sum_{k=1}^{2} \left[ \tilde{\eta}_{jk} \left( \log \hat{\alpha}_k + (\hat{\alpha}_k - 1) \log \tilde{P}_{jk} + \pi_{1k} \right) + (1 - \tilde{\eta}_{jk}) \log (1 - \pi_{1k}) \right].
\]

In the E step, we compute the posterior
\[
\tilde{\eta}_{jk} = \Pr (\tilde{\eta}_{jk} = 1 | \tilde{P}, \tilde{\alpha}, \pi_1) = \frac{\pi_{1k} \hat{\alpha}_k \tilde{P}_{jk}^{\hat{\alpha}_k-1}}{\pi_{1k} \hat{\alpha}_k \tilde{P}_{jk}^{\hat{\alpha}_k-1} + 1 - \pi_{1k}}
\]
and obtain the Q function
\[
Q = \sum_{j=1}^{M} \sum_{k=1}^{2} \left[ \tilde{\eta}_{jk} \left( \log \hat{\alpha}_k + (\hat{\alpha}_k - 1) \log \tilde{P}_{jk} + \pi_{1k} \right) + (1 - \tilde{\eta}_{jk}) \log (1 - \pi_{1k}) \right].
\]

The incomplete log-likelihood can be evaluated as
\[
L = \sum_{j=1}^{M} \sum_{k=1}^{2} \log \left( \frac{\pi_{1k} \hat{\alpha}_k \tilde{P}_{jk}^{\hat{\alpha}_k-1} + 1 - \pi_{1k}}{\pi_{1k} \hat{\alpha}_k \tilde{P}_{jk}^{\hat{\alpha}_k-1} + 1 - \pi_{1k}} \right).
\]

In the M step, we update \( \hat{\alpha}_k \) and \( \pi_{1k} \) by maximizing the Q function. We have
\[
\hat{\alpha}_k = -\frac{\sum_{j=1}^{M} \tilde{\eta}_{jk}}{\sum_{j=1}^{M} \tilde{\eta}_{jk} \log \tilde{P}_{jk}},
\]
\[
\pi_{1k} = \frac{1}{M} \sum_{j=1}^{M} \tilde{\eta}_{jk}.
\]

**Implementation**

Input: \( \tilde{P} \). Initialize: \( \tilde{\alpha} = c(0.1, 0.1), \pi_1 = c(0.1, 0.1) \). Output: \( \tilde{\alpha}, \pi_1 \).

- Initialize \( \tilde{\alpha} = c(0.1, 0.1), \pi_1 = c(0.1, 0.1) \).
- E-step: For \( j = 1, \ldots, M \) and \( k = 1, 2 \), calculate
  \[
  \tilde{\eta}_{jk} = \frac{\pi_{1k} \hat{\alpha}_k \tilde{P}_{jk}^{\hat{\alpha}_k-1}}{\pi_{1k} \hat{\alpha}_k \tilde{P}_{jk}^{\hat{\alpha}_k-1} + 1 - \pi_{1k}}.
  \]
Evaluate $L$

\[
L = \sum_{j=1}^{M} \sum_{k=1}^{2} \log \left( \pi_{1k} \tilde{\alpha}_k \tilde{P}_{jk}^{\alpha_k - 1} + 1 - \pi_{1k} \right).
\]

- M-step: For $k = 1, 2$, update

\[
\tilde{\alpha}_k = -\frac{\sum_{j=1}^{M} \tilde{\eta}_{jk}}{\sum_{j=1}^{M} \tilde{\eta}_{jk} \log \tilde{P}_{jk}},
\]

\[
\pi_{1k} = \frac{1}{M} \sum_{j=1}^{M} \tilde{\eta}_{jk}.
\]

- Check the convergence of $L$.

**Stage 2**

Suppose the correlation coefficient $\rho$ is zero, then the model becomes

\[
\tilde{P}_{jk} \sim \begin{cases} 
U[0, 1], & \tilde{\eta}_{jk} = 0, \\
\text{Beta}(\tilde{\alpha}_k, 1), & \tilde{\eta}_{jk} = 1,
\end{cases}
\]

\[
\tilde{\eta}_{jk} = \begin{cases} 
1, & \text{if } \tilde{Z}_{jk} > 0, \\
0, & \text{if } \tilde{Z}_{jk} \leq 0,
\end{cases}
\]

\[
\tilde{Z}_{jk} = \tilde{\beta}_k^T X_j + \tilde{\epsilon}_{jk},
\]

\[
\tilde{\epsilon}_{jk} \sim N(0, 1),
\]

where $\tilde{Z} \in \mathbb{R}^{M \times 2}$ is the latent variable in probit model, $X \in \mathbb{R}^{M \times (D+1)}$ is the design matrix of functional annotations, comprised of an intercept and $D$ annotations, $\tilde{\beta} \in \mathbb{R}^{2 \times (D+1)}$ is a matrix of the coefficients, $\tilde{\beta}_k$ is a vector of the $k$-th row of $\tilde{\beta}$. For the $j$-th SNP, $X_j$ is a vector containing the $j$-th row of $X$.

The expanded model is

\[
\tilde{P}_{jk} \sim \begin{cases} 
U[0, 1], & \tilde{\eta}_{jk} = 0, \\
\text{Beta}(\tilde{\alpha}_k, 1), & \tilde{\eta}_{jk} = 1,
\end{cases}
\]
$$\hat{\eta}_{jk} = \begin{cases} 1, & \text{if } \tilde{Z}_{jk} > 0, \\ 0, & \text{if } \tilde{Z}_{jk} \leq 0, \end{cases}$$

$$\tilde{Z}_{jk} = \gamma_k^T X_j + \tilde{\epsilon}_{jk},$$

$$\tilde{\epsilon}_{jk} \sim N(0, \sigma_k^2),$$

where $$\gamma_k = \sigma_k \hat{\beta}_k$$ and $$\sigma_k$$ is the auxiliary parameter whose value is fixed at 1 in the original model.

Let $$\hat{\theta} = \{\hat{\alpha}, \hat{\beta}\}$$ and $$\Theta = \{\hat{\alpha}, \gamma, \sigma\}$$ be the collection of model parameters in the original model and the expanded model, respectively.

For the expanded model, the complete-data log-likelihood is

$$\log \Pr \left( \tilde{P}, \tilde{\eta}, \tilde{Z} | X; \Theta \right) = \log \Pr \left( \tilde{P} | \tilde{\eta}; \hat{\alpha} \right) + \log \Pr \left( \tilde{Z} | X; \gamma, \Sigma \right)$$

$$= \sum_{j=1}^{M} \sum_{k=1}^{2} \tilde{\eta}_{jk} \left( \log \tilde{\alpha}_k + (\tilde{\alpha}_k - 1) \log \tilde{P}_{jk} \right)$$

$$+ \sum_{j=1}^{M} \sum_{k=1}^{2} \left[ -\frac{1}{2} \log (2\pi) - \frac{1}{2} \log \sigma_k^2 - \frac{1}{2\sigma_k^2} \left( \tilde{Z}_{jk} - \gamma_k^T X_j \right)^2 \right].$$

The Q function is evaluated as follows

$$Q = E_{\tilde{\eta}, \tilde{Z}} \log \Pr \left( \tilde{P}, \tilde{\eta}, \tilde{Z} | X; \Theta \right)$$

$$= \sum_{j=1}^{M} \sum_{k=1}^{2} E \left[ \tilde{\eta}_{jk} \right] \left( \log \tilde{\alpha}_k + (\tilde{\alpha}_k - 1) \log \tilde{P}_{jk} \right)$$

$$+ \sum_{j=1}^{M} \sum_{k=1}^{2} \left[ -\frac{1}{2} \log (2\pi) - \frac{1}{2} \log \sigma_k^2 \right]$$

$$+ \sum_{j=1}^{M} \sum_{k=1}^{2} \left[ -\frac{1}{2\sigma_k^2} E \left[ \tilde{Z}_{jk}^2 \right] + \frac{1}{\sigma_k^2} E \left[ \tilde{Z}_{jk} \right] \gamma_k^T X_j - \frac{1}{2\sigma_k^2} \left( \gamma_k^T X_j \right)^2 \right],$$

where the expectation is calculated based on the current $$\hat{\theta}$$ in the original model.

In the PX-E step, we compute the posterior of $$\tilde{\eta}_{jk}$$ as follows

$$E \left[ \tilde{\eta}_{jk} \right] = \Pr \left( \tilde{\eta}_{jk} = 1 | \tilde{P}, X; \hat{\theta} \right) = \frac{\Phi \left( \tilde{\beta}_k^T X_j \right) \tilde{\alpha}_k \tilde{P}_{jk}^{-1}}{\Phi \left( \tilde{\beta}_k^T X_j \right) \tilde{\alpha}_k \tilde{P}_{jk}^{-1} + 1 - \Phi \left( \tilde{\beta}_k^T X_j \right)}.$$
The posterior of $\tilde{Z}_{jk}$ given $\tilde{\eta}_{jk}$ is a truncated bivariate normal distribution. Using the properties of truncated bivariate normal distribution, we can get

$$
E \left( \tilde{Z}_{jk} | \tilde{P}, \mathbf{X}, \tilde{\theta} \right) = \tilde{\beta}_k^T \mathbf{X}_j + E \left[ \tilde{\eta}_{jk} \right] \frac{\phi \left( \tilde{\beta}_k^T \mathbf{X}_j \right)}{\Phi \left( \tilde{\beta}_k^T \mathbf{X}_j \right)} - (1 - E \left[ \tilde{\eta}_{jk} \right]) \frac{\phi \left( \tilde{\beta}_k^T \mathbf{X}_j \right)}{1 - \Phi \left( \tilde{\beta}_k^T \mathbf{X}_j \right)}
$$

$$
= \tilde{\beta}_k^T \mathbf{X}_j + \frac{\phi \left( \tilde{\beta}_k^T \mathbf{X}_j \right)}{\Phi \left( \tilde{\beta}_k^T \mathbf{X}_j \right)} \frac{\tilde{\alpha}_k \tilde{\beta}_k^{\alpha_k - 1} - 1}{\tilde{\alpha}_k \tilde{\beta}_k^{\alpha_k - 1} + 1 - \Phi \left( \tilde{\beta}_k^T \mathbf{X}_j \right)}
$$

$$
= \tilde{\beta}_k^T \mathbf{X}_j + v_{jk},
$$

$$
E \left( \tilde{Z}_{jk}^2 | \tilde{P}, \mathbf{X}, \tilde{\theta} \right) = \tilde{\beta}_k^T \mathbf{X}_j E \left( \tilde{Z}_{jk} | \tilde{P}, \mathbf{X}, \tilde{\theta} \right) + 1.
$$

The incomplete log-likelihood can be evaluated as

$$
L = \sum_{j=1}^{M} \sum_{k=1}^{2} \log \left( \Phi \left( \tilde{\beta}_k^T \mathbf{X}_j \right) \tilde{\alpha}_k \tilde{\beta}_k^{\alpha_k - 1} + 1 - \Phi \left( \tilde{\beta}_k^T \mathbf{X}_j \right) \right).
$$

In the PX-M step, we set the derivative of the Q function with respect to the parameters in $\Theta$ be zero and

$$
\frac{\partial Q}{\partial \tilde{\alpha}_k} = \sum_{j=1}^{M} E \left[ \tilde{\eta}_{jk} \right] \left( \frac{1}{\tilde{\alpha}_k} + \log \tilde{P}_{jk} \right) = 0,
$$

$$
\frac{\partial Q}{\partial \gamma_k} = \sum_{j=1}^{M} \left[ \frac{1}{\sigma_k^2} E \left[ \tilde{Z}_{jk}^T \mathbf{X}_j^T - \frac{1}{\sigma_k^2} \gamma_k^T \mathbf{X}_j \gamma_k \right] \right] = 0,
$$

$$
\frac{\partial Q}{\partial \sigma_k^2} = \sum_{i=1}^{M} \left[ \frac{1}{2 \sigma_k^2} + \frac{1}{2 \sigma_k^4} E \left[ \tilde{Z}_{jk} \right] - \frac{1}{\sigma_k^2} E \left[ \tilde{Z}_{jk} \right] \gamma_k^T \mathbf{X}_j \gamma_k + \frac{1}{2 \sigma_k^2} \left( \gamma_k^T \mathbf{X}_j \right) \right] = 0.
$$

Then we can obtain the updating equations

$$
\tilde{\alpha}_k = - \frac{\sum_{j=1}^{M} E \left[ \tilde{\eta}_{jk} \right]}{\sum_{j=1}^{M} E \left[ \tilde{\eta}_{jk} \right] \log \tilde{P}_{jk}}
$$

$$
\gamma = E \left[ \tilde{Z}^T \right] X \left( X^T X \right)^{-1},
$$

$$
\sigma_k^2 = \frac{1}{M} \sum_{j=1}^{M} \left[ E \left[ \tilde{Z}_{jk}^2 \right] - 2 E \left[ \tilde{Z}_{jk} \right] \gamma_k^T \mathbf{X}_j + \left( \gamma_k^T \mathbf{X}_j \right)^2 \right].
$$
We reduce to the original parameters using the reduction function:

\[ \tilde{\beta}_k = \frac{\gamma_k}{\sigma_k}. \]

Implementation

Input: \( \tilde{P}, X, \tilde{\alpha}, \tilde{\beta}_{k0} = -qnorm(\pi_{1k}) \). Output: \( \tilde{\alpha}, \tilde{\beta} \).

- Initialize \( \tilde{\alpha}, \tilde{\beta} = \begin{pmatrix} \tilde{\beta}_{10} & 0 & \cdots & 0 \\ \tilde{\beta}_{20} & 0 & \cdots & 0 \end{pmatrix} \).

- E-step: For \( j = 1, \ldots, M \) and \( k = 1,2 \), calculate

\[
E[\tilde{\eta}_{jk}] = \frac{\Phi(\tilde{\beta}_k^T X_j) \tilde{\alpha}_k \tilde{P}_{jk}^{\tilde{\alpha}_k^{-1}}}{\Phi(\tilde{\beta}_k^T X_j) \tilde{\alpha}_k \tilde{P}_{jk}^{\tilde{\alpha}_k^{-1}} + 1 - \Phi(\tilde{\beta}_k^T X_j)}.
\]

\[
E(\tilde{Z}_{jk} | \tilde{P}, X; \tilde{\theta}) = \tilde{\beta}_k^T X_j + \frac{\phi(\tilde{\beta}_k^T X_j) (\tilde{\alpha}_k \tilde{P}_{jk}^{\tilde{\alpha}_k^{-1}} - 1)}{\Phi(\tilde{\beta}_k^T X_j) \tilde{\alpha}_k \tilde{P}_{jk}^{\tilde{\alpha}_k^{-1}} + 1 - \Phi(\tilde{\beta}_k^T X_j)}
\]

\[
= \tilde{\beta}_k^T X_j + v_{jk},
\]

\[
E(\tilde{Z}_{jk}^2 | \tilde{P}, X; \tilde{\theta}) = \tilde{\beta}_k^T X_j E(\tilde{Z}_{jk} | \tilde{P}, X; \tilde{\theta}) + 1.
\]

Evaluate \( L \)

\[
L = \sum_{j=1}^{M} \sum_{k=1}^{2} \log \left( \Phi(\tilde{\beta}_k^T X_j) \tilde{\alpha}_k \tilde{P}_{jk}^{\tilde{\alpha}_k^{-1}} + 1 - \Phi(\tilde{\beta}_k^T X_j) \right).
\]

- M-step: For \( k = 1,2 \), update

\[
\tilde{\alpha}_k = -\frac{\sum_{j=1}^{M} E[\tilde{\eta}_{jk}]}{\sum_{j=1}^{M} E[\tilde{\eta}_{jk}] \log \tilde{P}_{jk}},
\]

\[
\gamma = E[\tilde{Z}' \times (X'X)^{-1}],
\]

\[
\sigma_k^2 = \frac{1}{M} \sum_{j=1}^{M} \left[ E[\tilde{Z}_{jk}^2] - 2E[\tilde{Z}_{jk}] \gamma_k^T X_j + (\gamma_k^T X_j)^2 \right].
\]
Then we reduce to the original parameters using the reduction function:

\[ \tilde{\beta}_k = \frac{\gamma_k}{\sigma_k}. \]

- Check the convergence of \( L \).

**Stage 3**

**Implementation**

Input: \( \tilde{P}, X, \tilde{\alpha}, \tilde{\beta} \). Initialize: \( \rho = 0 \). Output: \( \tilde{\alpha}, \tilde{\beta}, \rho \).

- Initialize \( \tilde{\alpha}, \tilde{\beta} \) using warm starts and \( \rho = 0 \).

- E-step: For \( j = 1, \ldots, M \) and \( k = 1, 2 \), calculate

\[
E[\tilde{\eta}_{j1}] = \frac{\Phi_{j11} \hat{\alpha}_1 \hat{\alpha}_1^{-1} \hat{\beta}_1 + \Phi_{j10} \hat{\alpha}_1}{\Phi_j},
\]

\[
E[\tilde{\eta}_{j2}] = \frac{\Phi_{j11} \hat{\alpha}_2 \hat{\alpha}_2^{-1} \hat{\beta}_2 + \Phi_{j01} \hat{\alpha}_2}{\Phi_j},
\]

where

\[
\Phi_{j11} = \Phi_2 \left( \tilde{\beta}_1^T X_j, \tilde{\beta}_2^T X_j, \rho \right),
\]

\[
\Phi_{j10} = \Phi_2 \left( \tilde{\beta}_1^T X_j, -\tilde{\beta}_2^T X_j, -\rho \right) = -\Phi_{j11} + \Phi \left( \tilde{\beta}_1^T X_j \right),
\]

\[
\Phi_{j01} = \Phi_2 \left( -\tilde{\beta}_1^T X_j, \tilde{\beta}_2^T X_j, -\rho \right) = -\Phi_{j11} + \Phi \left( \tilde{\beta}_2^T X_j \right),
\]

\[
\Phi_{j00} = \Phi_2 \left( -\tilde{\beta}_1^T X_j, -\tilde{\beta}_2^T X_j, \rho \right) = 1 + \Phi_{j11} - \Phi \left( \tilde{\beta}_1^T X_j \right) - \Phi \left( \tilde{\beta}_2^T X_j \right),
\]

\[
\Phi_j = \Phi_{j11} \hat{\alpha}_1 \hat{\alpha}_1^{-1} \hat{\beta}_1 + \Phi_{j10} \hat{\alpha}_1 + \Phi_{j01} \hat{\alpha}_2 + \Phi_{j00}.
\]

And

\[
E\left( \tilde{Z}_{j1} \right) = \tilde{\beta}_1^T X_j + \frac{\Phi_{j11}}{\Phi_j} E \left( \tilde{\epsilon}_{j1} | \tilde{\eta}_{j1} = 1, \tilde{\eta}_{j2} = 1, X_j ; \theta \right) + \frac{\Phi_{j10}}{\Phi_j} \hat{\alpha}_1 \hat{\alpha}_1^{-1} - 1 + \frac{\Phi_{j01}}{\Phi_j} \hat{\alpha}_2 \hat{\alpha}_2^{-1} - 1,
\]

[84]
\[ E(\tilde{Z}_{j2}) = \beta_2^T X_j + \frac{\Phi_{j11}}{\Phi_j} E\left(\tilde{\epsilon}_{j2} | \tilde{\eta}_{j1} = 1, \tilde{\eta}_{j2} = 1, X_j; \theta\right) \]
\[ + \frac{\rho \phi(\tilde{\beta}_1^T X_j)}{\Phi_j}\left(\tilde{\alpha}_1 \tilde{P}^{\tilde{\alpha}_1-1} - 1\right) + \frac{\phi(\tilde{\beta}_2^T X_j)}{\Phi_j}\left(\tilde{\alpha}_2 \tilde{P}^{\tilde{\alpha}_2-1} - 1\right), \]

\[ E(\tilde{Z}_{j1}) = \tilde{\beta}_1^T X_j \left(2 E(\tilde{Z}_{j1}) - \tilde{\beta}_1^T X_j\right) + 1 \]
\[ + \frac{\Phi_{j11}}{\Phi_j} E\left(\tilde{\epsilon}_{j1}^2 | \tilde{\eta}_{j1} = 1, \tilde{\eta}_{j2} = 1, X_j; \theta\right) - 1) \]
\[ - \frac{\tilde{\beta}_1^T X_j \phi(\tilde{\beta}_1^T X_j)}{\Phi_j}\left(\tilde{\alpha}_1 \tilde{P}^{\tilde{\alpha}_1-1} - 1\right) - \frac{\rho^2 \tilde{\beta}_2^T X_j \phi(\tilde{\beta}_2^T X_j)}{\Phi_j}\left(\tilde{\alpha}_2 \tilde{P}^{\tilde{\alpha}_2-1} - 1\right), \]

\[ E(\tilde{Z}_{j2}) = \tilde{\beta}_2^T X_j \left(2 E(\tilde{Z}_{j2}) - \tilde{\beta}_2^T X_j\right) + 1 \]
\[ + \frac{\Phi_{j11}}{\Phi_j} E\left(\tilde{\epsilon}_{j2}^2 | \tilde{\eta}_{j1} = 1, \tilde{\eta}_{j2} = 1, X_j; \theta\right) - 1) \]
\[ - \frac{\rho^2 \tilde{\beta}_1^T X_j \phi(\tilde{\beta}_1^T X_j)}{\Phi_j}\left(\tilde{\alpha}_1 \tilde{P}^{\tilde{\alpha}_1-1} - 1\right) - \frac{\rho^2 \tilde{\beta}_2^T X_j \phi(\tilde{\beta}_2^T X_j)}{\Phi_j}\left(\tilde{\alpha}_2 \tilde{P}^{\tilde{\alpha}_2-1} - 1\right), \]

\[ E(\tilde{Z}_{j1} \tilde{Z}_{j2}) = - \tilde{\beta}_1^T X_j \tilde{\beta}_2^T X_j + \tilde{\beta}_1^T X_j E(\tilde{Z}_{j2}) + \tilde{\beta}_2^T X_j E(\tilde{Z}_{j1}) + \rho \]
\[ + \frac{\Phi_{j11}}{\Phi_j} E\left(\tilde{\epsilon}_{j1} \tilde{\epsilon}_{j2} | \tilde{\eta}_{j1} = 1, \tilde{\eta}_{j2} = 1, X_j; \theta\right) - \rho) \]
\[ - \frac{\rho \tilde{\beta}_1^T X_j \phi(\tilde{\beta}_1^T X_j)}{\Phi_j}\left(\tilde{\alpha}_1 \tilde{P}^{\tilde{\alpha}_1-1} - 1\right) - \frac{\rho \tilde{\beta}_2^T X_j \phi(\tilde{\beta}_2^T X_j)}{\Phi_j}\left(\tilde{\alpha}_2 \tilde{P}^{\tilde{\alpha}_2-1} - 1\right). \]

where
\[ \Xi_j = \tilde{\alpha}_1 \tilde{P}^{\tilde{\alpha}_1-1} - \tilde{\alpha}_2 \tilde{P}^{\tilde{\alpha}_2-1} + 1, \]

\[ E\left(\tilde{\epsilon}_{j1} | \tilde{\eta}_{j1} = 1, \tilde{\eta}_{j2} = 1, X_j; \theta\right) = \frac{1}{\Phi_{j11}} \phi\left(\tilde{\beta}_1^T X_j\right) \Phi\left(\left(\tilde{\beta}_1^T X_j - \rho \tilde{\beta}_1^T X_j\right) c\right) \]
\[ + \frac{1}{\Phi_{j11}} \rho \phi\left(\tilde{\beta}_2^T X_j\right) \Phi\left(\left(\tilde{\beta}_2^T X_j - \rho \tilde{\beta}_2^T X_j\right) c\right), \]

85
\[
E \left( \epsilon_{j2} | \tilde{\eta}_{j1} = 1, \tilde{\eta}_{j2} = 1, \mathbf{X}_j; \mathbf{\theta} \right) = \frac{1}{\Phi_{j11}} \phi \left( \tilde{\beta}_2^T \mathbf{X}_j \right) \Phi \left( \left( \tilde{\beta}_1^T \mathbf{X}_j - \rho \tilde{\beta}_2^T \mathbf{X}_j \right) c \right) \\
+ \frac{1}{\Phi_{j11}} \rho \phi \left( \tilde{\beta}_1^T \mathbf{X}_j \right) \Phi \left( \left( \tilde{\beta}_2^T \mathbf{X}_j - \rho \tilde{\beta}_1^T \mathbf{X}_j \right) c \right),
\]

\[
E \left( \epsilon_{j1}^2 | \tilde{\eta}_{j1} = 1, \tilde{\eta}_{j2} = 1, \mathbf{X}_j; \mathbf{\theta} \right) = 1 - \frac{1}{\Phi_{j11}} \tilde{\beta}_1^T \mathbf{X}_j \phi \left( \tilde{\beta}_1^T \mathbf{X}_j \right) \Phi \left( \left( \tilde{\beta}_1^T \mathbf{X}_j - \rho \tilde{\beta}_1^T \mathbf{X}_j \right) c \right) \\
- \frac{1}{\Phi_{j11}} \rho^2 \tilde{\beta}_2^T \mathbf{X}_j \phi \left( \tilde{\beta}_1^T \mathbf{X}_j \right) \Phi \left( \left( \tilde{\beta}_1^T \mathbf{X}_j - \rho \tilde{\beta}_1^T \mathbf{X}_j \right) c \right) \\
+ \frac{1}{\Phi_{j11}} c^{-1} \rho \phi \left( \tilde{\beta}_1^T \mathbf{X}_j \right) \phi \left( \left( \tilde{\beta}_1^T \mathbf{X}_j - \rho \tilde{\beta}_1^T \mathbf{X}_j \right) c \right),
\]

\[
E \left( \epsilon_{j1} \epsilon_{j2} | \tilde{\eta}_{j1} = 1, \tilde{\eta}_{j2} = 1, \mathbf{X}_j; \mathbf{\theta} \right) = \rho - \frac{1}{\Phi_{j11}} \rho \tilde{\beta}_2^T \mathbf{X}_j \phi \left( \tilde{\beta}_1^T \mathbf{X}_j \right) \Phi \left( \left( \tilde{\beta}_1^T \mathbf{X}_j - \rho \tilde{\beta}_2^T \mathbf{X}_j \right) c \right) \\
- \frac{1}{\Phi_{j11}} \rho \tilde{\beta}_2^T \mathbf{X}_j \phi \left( \tilde{\beta}_1^T \mathbf{X}_j \right) \Phi \left( \left( \tilde{\beta}_1^T \mathbf{X}_j - \rho \tilde{\beta}_2^T \mathbf{X}_j \right) c \right) \\
+ \frac{1}{\Phi_{j11}} c^{-1} \phi \left( \tilde{\beta}_1^T \mathbf{X}_j \right) \phi \left( \left( \tilde{\beta}_1^T \mathbf{X}_j - \rho \tilde{\beta}_1^T \mathbf{X}_j \right) c \right),
\]

\[
c = \frac{1}{\sqrt{1 - \rho^2}},
\]

Evaluate \( L \)

\[
L = \sum_{j=1}^{M} \log \Phi_j.
\]
• M-step: For $k = 1, 2$, update

$$\tilde{\alpha}_k = -\frac{\sum_{j=1}^{M} E[\tilde{\eta}_{jk}]}{\sum_{j=1}^{M} E[\tilde{\eta}_{jk}] \log \tilde{P}_j},$$

$$\gamma = E[\tilde{Z}^T X (X^T X)^{-1}],$$

$$\Sigma = \frac{1}{M} \left[ \left( \sum_{j=1}^{M} E[\tilde{Z}_j \tilde{Z}_j^T] \right) - \gamma X^T E[\tilde{Z}] \right].$$

Then we reduce to the original parameters using the reduction function:

$$\tilde{\beta} = D^{-1} \gamma,$$

$$\tilde{R} = D^{-1} \Sigma D^{-1}.$$

• Check the convergence of $L$.

3.3.3 Method to get the parameter estimation in LPM

For $K$ GWASs, we analyze them pairwise using this algorithm and this procedure can be implemented parallelly. For LPM, the estimation $\hat{\alpha}_k$ and $\hat{\beta}_k$ can be obtained using the average estimation of the pairs which containing the $k$-th GWAS. And we can form a matrix $\hat{R}_{\text{pair}}$ using the corresponding estimation $\hat{\rho}$ in pairwise analysis. In real data analysis, the number of SNPs $M$ is often different in each GWAS. To avoid losing much information, we allow different $M$ in each pair analysis. However, since the pairwise analysis is not based on the same data, $\hat{R}_{\text{pair}}$ may not be positive semidefinite which is required for a correlation matrix. Therefore, we solve the following optimization problem to obtain the nearest correlation matrix $\tilde{R}$

$$\min \frac{1}{2} ||R - \hat{R}_{\text{pair}}||^2$$

s.t. $R_{kk} = 1$, $k = 1, ..., K$,

$$R \in S^K_+,$$

where $S^K_+$ is the cone of positive semidefinite matrices in the space of $K \times K$ symmetric matrices, and $|| \cdot ||$ is the Frobenius norm.
3.4 Inference

3.4.1 Identification of risk SNPs

After we obtain the estimation of parameters, we are able to prioritize risk SNPs based on the posterior of $\eta$ which indicate the association of the SNPs with the traits. If we consider the traits separately, the association mapping of the $j$-th SNP on the $k$-th trait can be inferred from

$$\Pr(\eta_{jk} = 1 | P_{jk}, X) = \frac{\Phi \left( \hat{\beta}_k^T X_j \right) \hat{\alpha}_k P_{jk}^{\hat{\alpha}_k - 1}}{\Phi \left( \hat{\beta}_k^T X_j \right) \hat{\alpha}_k P_{jk}^{\hat{\alpha}_k - 1} + 1 - \Phi \left( \hat{\beta}_k^T X_j \right)}.$$ 

In this case, the relationship among traits is ignored and only the current GWAS data is used.

If another GWAS data set is integrated, risk SNPs for both the $k$-th trait and the $k'$-th trait can be inferred from

$$\Pr(\eta_{jk} = 1, \eta_{jk'} = 1 | P_{jk}, P_{jk'}, X) = \frac{\Phi_{j11} \hat{\alpha}_k P_{jk}^{\hat{\alpha}_k - 1} \hat{\alpha}_{k'} P_{jk'}^{\hat{\alpha}_{k'} - 1}}{\Phi_j},$$

where

$$\Phi_{j11} = \Phi_2 \left( \hat{\beta}_k^T X_j, \hat{\beta'}_k^T X_j, \hat{\rho}_{kk'} \right),$$

$$\Phi_{j10} = \Phi_2 \left( \hat{\beta}_k^T X_j, -\hat{\beta'}_k^T X_j, -\hat{\rho}_{kk'} \right) = -\Phi_{j11} + \Phi \left( \hat{\beta}_k^T X_j \right),$$

$$\Phi_{j01} = \Phi_2 \left( -\hat{\beta}_k^T X_j, \hat{\beta'}_k^T X_j, -\hat{\rho}_{kk'} \right) = -\Phi_{j11} + \Phi \left( \hat{\beta'}_k^T X_j \right),$$

$$\Phi_{j10} = \Phi_2 \left( -\hat{\beta}_k^T X_j, -\hat{\beta'}_k^T X_j, \hat{\rho}_{kk'} \right) = 1 + \Phi_{j11} - \Phi \left( \hat{\beta}_k^T X_j \right) - \Phi \left( \hat{\beta'}_k^T X_j \right),$$

$$\Phi_j = \Phi_{j11} \hat{\alpha}_k P_{jk}^{\hat{\alpha}_k - 1} \hat{\alpha}_{k'} P_{jk'}^{\hat{\alpha}_{k'} - 1} + \Phi_{j10} \hat{\alpha}_k P_{jk}^{\hat{\alpha}_k - 1} + \Phi_{j01} \hat{\alpha}_{k'} P_{jk'}^{\hat{\alpha}_{k'} - 1} + \Phi_{j10}.$$

In addition, we can infer the risk SNPs for the $k$-th trait by calculating the marginal posterior

$$\Pr(\eta_{jk} = 1 | P_{jk}, P_{jk'}, X)$$

$$= \Pr(\eta_{jk} = 1, \eta_{jk'} = 1 | P_{jk}, P_{jk'}, X) + \Pr(\eta_{jk} = 1, \eta_{jk'} = 0 | P_{jk}, P_{jk'}, X).$$
where
\[
\Pr(\eta_{jk} = 1, \eta_{jk'} = 0 | P_{jk}, P_{jk'}, X) = \frac{\Phi_{j10} \hat{\alpha}_k P_{jk}^{\hat{\alpha}_k - 1}}{\Phi_j}.
\]

Similarly, we can consider more than two traits, e.g., three traits, and obtain the posterior that the \(j\)-th SNP is associated with the \(k\)-th, \(k'\)-th and \(k''\)-th trait
\[
\Pr(\eta_{jk} = 1, \eta_{jk'} = 1, \eta_{jk''} = 1 | P_{jk}, P_{jk'}, P_{jk''}, X),
\]
the marginal posterior that the \(j\)-th SNP is associated with the \(k\)-th and \(k'\)-th trait
\[
\Pr(\eta_{jk} = 1, \eta_{jk'} = 1 | P_{jk}, P_{jk'}, P_{jk''}, X),
\]
and the marginal posterior that the \(j\)-th SNP is associated with the \(k\)-th trait
\[
\Pr(\eta_{jk} = 1 | P_{jk}, P_{jk'}, P_{jk''}, X).
\]

where
\[
\Phi_{jlmn} = \Phi_3(\mu_{jkk'k''}, R_{kk'k''}), \text{ for } l, m, n \in \{1, 0\},
\]
\[
\Phi_j = \sum_{l=0}^1 \sum_{m=0}^1 \sum_{n=0}^1 \Phi_{jlmn} \left( \hat{\alpha}_k P_{jk}^{\hat{\alpha}_k - 1} \right)^l \left( \hat{\alpha}_{k'} P_{jk'}^{\hat{\alpha}_{k'} - 1} \right)^m \left( \hat{\alpha}_{k''} P_{jk''}^{\hat{\alpha}_{k''} - 1} \right)^n,
\]
\[
\mu_{jkk'k''} = \left( \begin{array}{c}
(2l - 1) \hat{\beta}_k^T X_j \\
(2m - 1) \hat{\beta}_{k'}^T X_j \\
(2n - 1) \hat{\beta}_{k''}^T X_j
\end{array} \right),
\]
\[
R_{kk'k''} = \left( \begin{array}{ccc}
1 & (2l - 1) (2m - 1) \hat{\rho}_{kk'} & (2l - 1) (2n - 1) \hat{\rho}_{kk''} \\
(2l - 1) (2m - 1) \hat{\rho}_{kk'} & 1 & (2m - 1) (2n - 1) \hat{\rho}_{kk''} \\
(2l - 1) (2n - 1) \hat{\rho}_{kk''} & (2m - 1) (2n - 1) \hat{\rho}_{kk''} & 1
\end{array} \right).
\]
and \( \Phi_3 (\mu, \Sigma) = \int_{-\infty}^{\mu_1} \int_{-\infty}^{\mu_2} \int_{-\infty}^{\mu_3} \phi_3 (x, \Sigma) \, dx, \)

\[
\phi_3 (x, \Sigma) = \frac{\exp \left[ -\frac{1}{2} x^T \Sigma^{-1} x \right]}{\sqrt{(2\pi)^3 |\Sigma|}}.
\]

Accordingly, we can calculate the local false discovery rate of the \( j \)-th SNP for the \( k \)-th trait when considering one or more traits simultaneously, i.e.,

\[
fdr_1 (P_{jk}) = 1 - \Pr (\eta_{jk} = 1 | P_{jk}, X),
\]

\[
fdr_1 (P_{jk}, P_{jk'}) = 1 - \Pr (\eta_{jk} = 1 | P_{jk}, P_{jk'}, X),
\]

\[
fdr_1 (P_{jk}, P_{jk'}, P_{jk''}) = 1 - \Pr (\eta_{jk} = 1 | P_{jk}, P_{jk'}, P_{jk''}, X),
\]

and for both the \( k \)-th and \( k' \)-th trait

\[
fdr_2 (P_{jk}, P_{jk'}) = 1 - \Pr (\eta_{jk} = 1, \eta_{jk'} = 1 | P_{jk}, P_{jk'}, X),
\]

\[
fdr_2 (P_{jk}, P_{jk'}, P_{jk''}) = 1 - \Pr (\eta_{jk} = 1, \eta_{jk'} = 1, \eta_{jk''} = 1 | P_{jk}, P_{jk'}, P_{jk''}, X),
\]

and for three traits

\[
fdr_3 (P_{jk}, P_{jk'}, P_{jk''}) = 1 - \Pr (\eta_{jk} = 1, \eta_{jk'} = 1, \eta_{jk''} = 1 | P_{jk}, P_{jk'}, P_{jk''}, X).
\]

We use the following approach to control the global false discovery rate (FDR). We sort SNPs by \( fdr \) from the smallest to the largest and regard the \( j \)-th re-ordered SNP as a risk SNP for each of the three cases if

\[
FDR_{(j)} = \frac{\sum_{i=1}^{j} fdr_{(i)}}{j} \leq \tau,
\]

where \( fdr_{(i)} \) is the \( i \)-th ordered \( fdr \), \( FDR_{(j)} \) is the corresponding global FDR, and \( \tau \) is the threshold of global FDR.
3.4.2 Relationship test among traits

We can test the relationship between two traits in the pairwise analysis by the hypothesis:

\[ H_0 : \rho = 0, \text{ v.s. } H_1 : \rho \neq 0. \]

We use the likelihood ratio test. The test statistic is

\[ \lambda = 2 \left[ \log \Pr \left( \tilde{P} | X; \hat{\theta} \right) - \log \Pr \left( \tilde{P} | X; \hat{\theta}_0 \right) \right], \]

where \( \hat{\theta}_0 \) is the parameter estimates under \( H_0 \), i.e., the estimates we obtain in the second stage of the algorithm. The probability distribution of \( \lambda \) is approximately a \( \chi^2 \) distribution with \( df = 1 \) under the null.

3.4.3 Hypothesis testing of annotation enrichment

When we integrate functional annotation data, we may be interested in the enrichment of annotation for a specific trait. We consider the following test:

\[ H_0 : \beta_{kd} = 0, \text{ v.s. } H_1 : \beta_{kd} \neq 0. \]

To estimate the standard error of \( \hat{\beta}_{kd} \), we consider the single trait case. The log-likelihood for the \( k \)-th trait is

\[ L = \sum_{j=1}^{M} \log \left( \Phi \left( \beta_k^T X_j \right) \alpha_k P_{jk}^{\alpha_k^{-1}} + 1 - \Phi \left( \beta_k^T X_j \right) \right). \]

The information matrix of log-likelihood can be computed by

\[ \mathcal{I} \left( \hat{\theta}_k \right) = -H|_{\hat{\theta}_k} = -\nabla^2 L \left( \theta_k \right) |_{\hat{\theta}_k}, \]

where \( \theta_k = (\alpha_k, \beta_k^T)^T \) and \( \hat{\theta}_k \) is its estimate obtained by LPM. It can be obtained by

\[ \nabla^2 L \left( \theta_k \right) = \begin{pmatrix} \frac{\partial^2 L}{\partial \alpha^2} & \frac{\partial^2 L}{\partial \alpha \partial \beta} \\ \frac{\partial^2 L}{\partial \beta^T \partial \alpha} & \frac{\partial^2 L}{\partial \beta^2} \end{pmatrix}, \]
where

\[
\frac{\partial L}{\partial \alpha} = \sum_{j=1}^{M} \frac{\Phi \left( \beta^T X_j \right) P_j^{\alpha-1} (1 + \alpha \log P_j)}{\Phi (\beta^T X_j) \alpha P_j^{\alpha-1} + 1 - \Phi (\beta^T X_j)},
\]

\[
\frac{\partial L}{\partial \beta} = \sum_{j=1}^{M} \frac{\alpha P_j^{\alpha-1} - 1}{\Phi (\beta^T X_j) \alpha P_j^{\alpha-1} + 1 - \Phi (\beta^T X_j)} \phi \left( \beta^T X_j \right) X_j^T,
\]

\[
\frac{\partial^2 L}{\partial \alpha^2} = -\sum_{j=1}^{M} \left( \frac{\Phi \left( \beta^T X_j \right) P_j^{\alpha-1} (1 + \alpha \log P_j)}{\Phi (\beta^T X_j) \alpha P_j^{\alpha-1} + 1 - \Phi (\beta^T X_j)} \right)^2
+ \sum_{j=1}^{M} \frac{\Phi \left( \beta^T X_j \right) P_j^{\alpha-1} (2 + \alpha \log P_j) \log P_j}{\Phi (\beta^T X_j) \alpha P_j^{\alpha-1} + 1 - \Phi (\beta^T X_j)},
\]

\[
\frac{\partial^2 L}{\partial \alpha \partial \beta} = -\sum_{j=1}^{M} \frac{\Phi \left( \beta^T X_j \right) P_j^{\alpha-1} (1 + \alpha \log P_j) (\alpha P_j^{\alpha-1} - 1) \phi \left( \beta^T X_j \right) X_j^T}{\Phi (\beta^T X_j) \alpha P_j^{\alpha-1} + 1 - \Phi (\beta^T X_j))^2}
+ \sum_{j=1}^{M} \frac{P_j^{\alpha-1} (1 + \alpha \log P_j) \phi \left( \beta^T X_j \right) X_j^T}{\Phi (\beta^T X_j) \alpha P_j^{\alpha-1} + 1 - \Phi (\beta^T X_j)},
\]

\[
\frac{\partial^2 L}{\partial \beta^T \partial \beta} = -\sum_{j=1}^{M} \frac{\left( (\alpha P_j^{\alpha-1} - 1) \phi \left( \beta^T X_j \right) \right)^2 X_j X_j^T}{\Phi (\beta^T X_j) \alpha P_j^{\alpha-1} + 1 - \Phi (\beta^T X_j))^2}
+ \sum_{j=1}^{M} \frac{- (\alpha P_j^{\alpha-1} - 1) \phi \left( \beta^T X_j \right) \beta^T X_j X_j^T}{\Phi (\beta^T X_j) \alpha P_j^{\alpha-1} + 1 - \Phi (\beta^T X_j)}.
\]

Then the inverse of the observed information matrix is an estimator of the asymptotic covariance matrix \( \text{Var} \left( \hat{\theta}_k \right) = \left[ I \left( \hat{\theta}_k \right) \right]^{-1} \). The Wald test statistic is

\[
W = \frac{\hat{\beta}_k^{2 \text{d}}}{\left[ I \left( \hat{\theta}_k \right) \right]^{-1}_{d+1,d+1}}.
\]

The probability distribution of \( W \) is approximately a \( \chi^2 \) distribution with \( df = 1 \) under the null.
3.5 Theorem based on composite likelihood approach

In our study, we analyze the traits pairwise. With increasing number of traits $K$, the parameters in $\mathbf{R}$ only increases quadratically ($K^2$) in LPM. However, for other methods such as GPA, the number of groups will increase exponentially ($2^K$). Theoretically, it is based on the composite likelihood approach [Varin et al., 2011].

For the marginal model for phenotypes $k$ and $k'$, let $\tilde{\theta}_{kk'} = \{\tilde{\alpha}_{kk'}, \tilde{\beta}_{kk'}, \tilde{R}_{kk'}\}$ be the collection of parameters, and $l_{kk'}(\hat{\theta}_{kk'}) = \sum_{j=1}^{M} l_{jkk'}(\hat{\theta}_{kk'})$ the corresponding log-likelihood.

Let $\tilde{\Theta} = \{\tilde{\theta}_{kk'}, 1 \leq k < k' \leq K\}$. Consider the pairwise log-likelihood constructed from all possible pairs of phenotypes

$$ cl(\tilde{\Theta}) = \sum_{1 \leq k < k' \leq K} l_{kk'}(\tilde{\theta}_{kk'}) = \sum_{1 \leq k < k' \leq K} \sum_{j=1}^{M} l_{jkk'}(\tilde{\theta}_{kk'}). $$

This function is also known as a composite or pseudo likelihood [Varin et al., 2011]. Maximizing it with respect to $\tilde{\Theta}$ is equivalent to our proposed method of fitting all pairwise marginal models.

While the classical composite likelihood approach allows one or more parameters to be present in different pieces of the likelihood, we assume that each marginal likelihood has a separate set of parameters, that is, all elements of $\Theta$ are distinct. Let $\tilde{\Theta}^*$ be the set of all parameters in the joint model. One can check that $\alpha_k$ and $\beta_k$ in $\Theta^*$ have $K - 1$ counterparts in $\tilde{\Theta}$, while $\rho_{kk'}$ has a single counterpart. Consequently, there exists a matrix $A$ such that $\Theta^* = A \tilde{\Theta}$. The motivation of this separate parameterization is to ease the computational burden of joint maximization.

Let

$$ s_{jkk'}(\hat{\theta}_{kk'}) = \partial l_{jkk'}(\hat{\theta}_{kk'})/\partial \hat{\theta}_{kk'} $$

and

$$ H_{jkk'}(\hat{\theta}_{kk'}) = \partial s_{jkk'}(\hat{\theta}_{kk'})/\partial \hat{\theta}_{kk'}. $$

Denote by $\hat{\Theta}$ the maximizer of $cl(\tilde{\Theta})$. We have the following theorem.
Theorem 1. Let \( \Omega = -\sum_{j=1}^{M} E\{H_{jkk}(\hat{\theta}_{kk})\}/M \) and \( \Sigma = \sum_{j=1}^{M} \text{cov}\{s_{jkk}(\hat{\theta}_{kk})\}/M \). Then, under mild regularity conditions, \( \sqrt{M} \text{vec}(\hat{\Theta} - \hat{\Theta}) \) converges in distribution to a multivariate normal distribution with mean vector zero and variance-covariance matrix \( \Omega^{-1}\Sigma\Omega^{-1} \).

The proof of this theorem, which we omit, follows directly from the general pseudo likelihood theory. See, for example, Geys et al. [1999] and Steffen and Geert [2006]. To estimate the variance-covariance matrix, we can drop the expectations and covariances, and replace the unknown parameters by their estimates. This empirical estimator is known as the “sandwich” estimator. Let \( \hat{\Theta}^* = A\hat{\Theta} \).

Corollary 2. Under mild regularity conditions, \( \sqrt{M} \text{vec}(\hat{\Theta}^* - \Theta^*) \) converges in distribution to a multivariate normal distribution with mean vector zero and variance-covariance matrix \( A'\Omega^{-1}\Sigma\Omega^{-1}A \).

3.6 Simulation

3.6.1 Simulation of eight traits

Simulation setup

We conducted comprehensive simulations to evaluate the performance of the proposed LPM. We generated the simulation data for eight traits using the generative model. The procedure was as follows. We considered eight traits which were divided into three groups: (i) P1, P2 and P3; (ii) P4, P5, P6; and (iii) P7, P8. Correlation existed only within the groups. Specifically, we set the correlation matrix \( \mathbf{R} \) with the corresponding entries \( \rho_{12} = 0.7, \rho_{13} = 0.4, \rho_{23} = 0.2, \rho_{45} = 0.6, \rho_{46} = 0.3, \rho_{56} = 0.1 \) and \( \rho_{78} = 0.5 \) (all the other entries were set to zeros). The relationship among the traits was depicted in Figure 3.1. The numbers of SNPs and functional annotations were set to be \( M = 100,000 \) and \( D = 5 \) respectively. First, we generated the design matrix \( \mathbf{X} \) and coefficients \( \mathbf{\beta} \) of functional annotations. The entries in \( \mathbf{X} \) excluded the intercept were generated from Bernoulli (0.2). The entries in first column of \( \mathbf{\beta} \) were set to be \(-1\) and the other entries were first generated from \( N(0,1) \) and then transformed to control the signal-noise ratio \( r = \frac{[\text{var}(\mathbf{x}_k\mathbf{\beta}^T)]_{kk}}{[\text{var}(\epsilon)]_{kk}} \) for \( k = 1,\ldots,8 \) to be
Both $X$ and $\beta$ were kept fixed in multiple replications. Then we simulated $\eta_{jk}$ according to the multivariate probit model (3.1)-(3.3). Finally, we generated $P_{jk}$ from $U[0,1]$ if $\eta_{jk} = 0$ and $Beta(\alpha_k,1)$ if $\eta_{jk} = 1$ with $\alpha_1 = 0.2$, $\alpha_2 = 0.35$, $\alpha_3 = 0.5$, $\alpha_4 = 0.3$, $\alpha_5 = 0.45$, $\alpha_6 = 0.55$, $\alpha_7 = 0.25$ and $\alpha_8 = 0.4$.

![Figure 3.1: The true correlation graph among eight traits. The number on the edges and the width of the edges indicate the correlation between the connected traits.](image)

**Performance in characterizing the correlation among the traits**

We first evaluated the performance of LPM in characterizing the correlation among the traits in our simulation. Figure 3.2 shows that the correlation graph is accurately estimated using LPM. We also evaluated the type I error rate and power of LPM for the relationship test among trait. As shown in Figure 3.3a, the type I error rates are almost 0 for all the pairs with no correlation and the powers are almost 1 except for two pairs (P2 and P3, P5 and P6) in which cases the correlations are relatively small and the signal strength are relatively weak, i.e., the corresponding $\rho$ is relatively small and $\alpha$ is relatively large. The comparison results of GPA and graph-GPA are shown in Figures 3.3b and 3.3c. As both GPA and graph-GPA do not adjust the effect of functional annotations, more significant relationships are detected.
Figure 3.2: The average estimated correlation graph using LPM. The number on the edges and the width of the edges indicate the correlation between the connected traits. The results are summarized from 50 replications.

Figure 3.3: Relationship test graph of (a) LPM, (b) GPA, and (c) graph-GPA. The number on the edges and the width of the edges indicate the type I error or power of the relationship test for the connected traits. For LPM and GPA, we controlled family-wise error rate at 0.05. The results are summarized from 50 replications.
Performance in the identification of risk SNPs for one or multiple traits

Then we evaluated the performance of LPM in the identification of risk SNPs for one or multiple traits. To detect risk SNPs for one specific trait, we consider three cases (i) separately analysis of the target trait, (ii) jointly analysis of the target trait with another trait, and (iii) jointly analysis of the target trait with another two traits, using the method described in Section 3.4.1. If the integrated traits are correlated with the target trait, the power to identify risk SNPs is expected to increase in joint analysis. We compared LPM with GPA under these three cases in terms of their empirical FDR and AUC. We controlled global FDR at 0.1 to evaluate empirical FDR. The results based on 50 replications are shown in Figure 3.4-3.6. The empirical FDRs of LPM are indeed controlled at the nominal level. However, the FDRs of GPA are inflated in some cases when the GWAS signal is relatively weak and are conservative when the GWAS signal is relatively strong. This is because the model design of LPM and GPA are different and our method to generate simulation data is in favor of LPM. Moreover, LPM outperformed GPA for all the cases in terms of AUC. As expected, the AUC of LPM increases when correlated traits are integrated. For example, as shown in Figure 3.4b, the power to identify risk SNPs for P1 increases as the correlated traits (P2 and P3) are jointly analyzed. Specifically, integrating trait with high correlation with the target trait could result in a better improvement of AUC.

The comparison performance of LPM and GPA in the identification of SNPs associated with two and three traits based on 50 replications is shown in Figure 3.7 and Figure 3.8 respectively. We controlled global FDR at 0.1 to evaluate empirical FDR. LPM performs better in terms of FDR control and AUC. In the identification of risk SNPs for both P1 and P4, a larger AUC can be achieved by integrating trait which is correlated with either P1 and P4, i.e., integrating P2, P3, P5 or P6.

Estimation of parameters

We also evaluated the accuracy of parameter estimation using LPM based on 50 replications. Figure 3.9 indicates that LPM provided satisfactory estimate of $\alpha$, $\beta$ and $R$. 

97
Figure 3.4: FDR and AUC of LPM and GPA for identification of risk SNPs for P1 (upper panel) and P2 (lower panel). The text in the x-axis indicated the GWAS data sets used in analysis, e.g., P1 indicated separate analysis using only the GWAS data of P1, +P2 in (a) indicated jointly analysis of P1 and P2. The red horizontal line in (b) and was set at the median AUC in separate analysis using LPM as a reference line.
Figure 3.5: FDR and AUC of LPM and GPA for identification of risk SNPs for P3 (upper panel), P4 (middle panel) and P5 (lower panel).
Figure 3.6: FDR and AUC of LPM and GPA for identification of risk SNPs for P6 (upper panel), P7 (middle panel) and P8 (lower panel).
Figure 3.7: FDR and AUC of LPM and GPA for identification of risk SNPs for both P1 and P4. The text in the x-axis indicated the GWAS data sets used in analysis. The red horizontal line in (b) was set at the median AUC in joint analysis of P1 and P4 using LPM as a reference line.

Figure 3.8: FDR and AUC of LPM and GPA for identification of risk SNPs for three specific traits. The text in the x-axis indicated the GWAS data sets used in analysis.
Figure 3.9: Parameter estimation performance of $\alpha$, $\beta$ and $R$ using LPM. We separated the comparison graph of $\beta_{kd}$ for better visualization in (c) and (d). The red dashed lines indicate the diagonal.
Specifically, the estimate of $\tilde{\alpha}$, $\tilde{\beta}$ and $\tilde{R}$ using bLPM for different pairs are stable which shows the consistency and reliability of our algorithm (Figures 3.10-3.19).

Figure 3.10: Comparison of the parameter estimation of $\alpha$ using LPM and $\tilde{\alpha}$ using bLPM. The red dashed lines indicate the true values.

### 3.6.2 Simulation of eight traits without annotation

We simulate the case when there is no functional annotation, i.e., $D = 0$. In this case $X$ only had the intercept term and $\beta$ only had one column which were set to be $-1$. Other settings are the same with those in Section 3.6.1.
Figure 3.11: Comparison of the parameter estimation of $\beta$ using LPM and $\tilde{\beta}$ using bLPM for P1. The red dashed lines indicate the true values.

Figure 3.12: Comparison of the parameter estimation of $\beta$ using LPM and $\tilde{\beta}$ using bLPM for P2. The red dashed lines indicate the true values.

Figure 3.13: Comparison of the parameter estimation of $\beta$ using LPM and $\tilde{\beta}$ using bLPM for P3. The red dashed lines indicate the true values.
Figure 3.14: Comparison of the parameter estimation of $\beta$ using LPM and $\tilde{\beta}$ using bLPM for P4. The red dashed lines indicate the true values.

Figure 3.15: Comparison of the parameter estimation of $\beta$ using LPM and $\tilde{\beta}$ using bLPM for P5. The red dashed lines indicate the true values.

Figure 3.16: Comparison of the parameter estimation of $\beta$ using LPM and $\tilde{\beta}$ using bLPM for P6. The red dashed lines indicate the true values.
Figure 3.17: Comparison of the parameter estimation of $\beta$ using LPM and $\hat{\beta}$ using bLPM for P7. The red dashed lines indicate the true values.

Figure 3.18: Comparison of the parameter estimation of $\beta$ using LPM and $\hat{\beta}$ using bLPM for P8. The red dashed lines indicate the true values.
Figure 3.19: Comparison of the parameter estimation of $R$ using LPM and $\tilde{R}$ using bLPM. The red dashed lines indicate the true values.
Performance in characterizing the correlation among the traits

As shown in Figure 3.20, the relationship test graphs are similar for LPM, GPA and graph-GPA.

![Graphs of LPM, GPA, and graph-GPA](image)

Figure 3.20: Relationship test graph of (a) LPM, (b) GPA, and (c) graph-GPA when there is no functional annotation. The number on the edges and the width of the edges indicate the type I error or power of the relationship test for the connected traits. For LPM and GPA, we controlled family-wise error rate at 0.05. The results are summarized from 50 replications.

Performance in the identification of risk SNPs for one trait

We have shown that when there is no functional annotation, the relationship test graphs are similar for LPM, GPA and graph-GPA. We also compare their performance in the identification of risk SNPs of one specific trait in this case. The results are shown in Figures 3.21-3.23. The performance of LPM and GPA are very close in terms of FDR and AUC. However, for graph-GPA, the empirical FDRs of graph-GPA are conservative and AUCs are relatively lower because the simulations are based on the generative model of LPM.

3.6.3 Simulations of two traits

As we analyze GWASs pairwise in our algorithm, to provide a better illustration for the performance of LPM, we conducted simulations which contain only two traits.
Figure 3.21: FDR and AUC of LPM and GPA for identification of risk SNPs for P1 (upper panel) and P2 (lower panel) when there is no functional annotations. The text in the x-axis indicated the GWAS data sets used in analysis, e.g., P1 indicated separate analysis using only the GWAS data of P1, +P2 in (a) indicated jointly analysis of P1 and P2. The red horizontal line in (b) and was set at the median AUC in separate analysis using LPM as a reference line.
Figure 3.22: FDR and AUC of LPM and GPA for identification of risk SNPs for P3 (upper panel), P4 (middle panel) and P5 (lower panel) when there is no functional annotations.
Figure 3.23: FDR and AUC of LPM and GPA for identification of risk SNPs for P6 (upper panel), P7 (middle panel) and P8 (lower panel) when there is no functional annotations.
Performance in characterizing the correlation between traits

In this simulation, we suppose the signal strength of the traits are the same, i.e., \( \alpha_1 = \alpha_2 = \alpha \). We varied \( \alpha \) at \{0.2, 0.4, 0.6\} and \( r \) at \{0.25, 1, 4\} to obtain the type I error rate and varied \( \rho \) at \{0, 0.05, 0.1, 0.15, 0.2, 0.25\} to obtain the power of LPM for the relationship test between the two traits. The results are shown in Figure 3.24. The type I error rates are appropriately controlled in all cases and the power increases as \( \alpha \) decreases and as \( \rho \) increases. However, we noted that a large signal-noise ratio could lead to a smaller power. This is because the correlation resulting from annotations increases as \( r \) increases and the correlation we aim to estimate becomes relatively smaller.

![Graphs showing type I error rate and power of LPM for different signal-noise ratios and \( \alpha \) values.]

Figure 3.24: Type I error rate and power of LPM for the relationship test between two traits. The bars represent one standard error. We evaluate type I error rate and power at 0.05 significance level. The results are summarized from 500 replications.
Performance in the identification of risk SNPs for one trait

To evaluate the performance of LPM for the identification of risk SNPs of one specific trait when correlated trait is integrated, we suppose the first trait is the target trait with $\alpha_1 = 0.2$. We varied $\alpha_2$ at $\{0.2, 0.4, 0.6\}$ and $\rho$ at $\{0, 0.2, 0.4, 0.6\}$ to obtain the FDR and AUC of LPM for the identification of risk SNPs of the target trait. The results are shown in Figure 3.25.

![Figure 3.25: FDR and AUC of LPM for identification of risk SNPs for the target trait when correlated trait is integrated. We controlled global FDR at 0.1 to evaluate empirical FDR. The results are summarized from 50 replications.](image)

Performance of LPM for the annotation enrichment test

We further conducted simulations to evaluate the type I error rate and power of LPM for the hypothesis testing of annotation enrichment. In this simulation, we suppose the signal strength of the traits are the same, i.e., $\alpha_1 = \alpha_2 = \alpha$, the number of annotations to be $D = 1$ and no correlation $\rho = 0$. We varied $\alpha$ at $\{0.2, 0.4, 0.6\}$ to obtain the type I error rate and varied the coefficient of the annotation $\beta$ at $\{-0.4, -0.3, -0.2, -0.1, 0.1, 0.2, 0.3, 0.4\}$ to obtain the power of LPM for the enrichment test of the annotation. The results are shown in Figure 3.26. We observe that the type I error rate is indeed controlled at the nominal level and the power is almost 1 when the signal strength is relatively strong, i.e., $\alpha = 0.2$ or 0.4, and the coefficient is not very small, i.e., $|\beta| \geq 0.2$. 

113
Figure 3.26: Type I error rate and power of LPM for the annotation enrichment test. The bars represent one standard error. We evaluate type I error rate and power at 0.05 significance level. The results are summarized from 500 replications.

3.6.4 Computational time

Figure 3.27 shows the computational time of LPM with $M = 100,000$ and $D = 5$. For one pair of traits, the computational time depends on the signal strength of GWAS data and their correlation. When the number of trait increases, the time can be largely shortened by using more cores in parallel computation.
3.6.5 More simulations

LPM assumes that the $P$-values of SNPs in each GWAS are from a mixture of uniform and Beta distribution. We have shown that when this assumption is violated LPM is still robust. We consider the situations when $P$-values in non-null group are from other distributions other than the Beta distribution and when $P$-values are obtained from individual-level data.

Simulations based on individual-level data

We consider the two GWASs case where $P$-values are obtained from individual-level data instead of directly simulating from the generative model. We set the number of individuals, SNPs and functional annotations to be $N = 5,000$, $M = 10,000$ and $D = 5$ respectively. The procedure to generate $X$, $\beta$ and $\eta$ is the same with that in Section 3.6.1. To generate the genotype matrix $G$, we first simulate the minor allele frequency for each SNPs from $U[0,1]$ and then encode the entries in $G$ by $\{0, 1, 2\}$ following the Hardy-Weinberg principle. The effect sizes $\beta_{SNP}$ were simulated from $N(0, 1)$ if the corresponding entries in $\eta$ were 1 and were set to be 0 otherwise. The noise level $\sigma_e^2$ was specified to control heritability $h^2 = \frac{\text{var}(G\beta_{SNP})}{\text{var}(G\beta_{SNP}) + \sigma_e^2}$ at given levels. The phenotype data $y$ was generated based on $y = G\beta_{SNP} + e$, where $e_i \sim N(0, \sigma_e^2)$ for $i = 1, ..., N$. Finally, we calculated $P$-values for each SNP using univariate linear regression. We varied heritability $h^2$ at $\{0.3, 0.5, 0.8\}$ and calculated the type I error rate of LPM for the relationship test between the two traits and the empirical FDR for identification of risk SNPs. The results shown in Figures 3.29-3.28, indicating that the type I error rate of LPM for the relationship test and the empirical FDR for identification of risk SNPs are still controlled at the nominal level.

Simulations if $P$-values are not from beta distribution

We simulate the two GWASs case when $P$-values are not from beta distribution. We generated $z$-values for SNPs in the null group from the standard normal distribution and $z$-values for SNPs in the non-null group from the distributions in Table 3.1. Then we converted $z$-scores to obtain $P$-values. In these simulations, $P$-values do not follow the mixture of uniform and beta distribution. We set $M = 100,000$ and $D = 5$. The
Figure 3.28: Type I error rate of LPM for the relationship test between the two traits when \( P \)-values are simulated from individual-level data. The bars represent one standard error. We evaluate type I error rate at 0.05 significance level. The results are summarized from 500 replications.

procedure to generate \( X, \beta \) and \( \eta \) is the same with that in Section 3.6.1. As shown in Figure 3.30-3.31, the type I error rate of LPM for the relationship test between the two traits under these distributions and the empirical FDR for identification of risk SNPs are still controlled at the nominal level.

Table 3.1: Alternative distributions for \( z \)-scores.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>spiky</td>
<td>( 0.4N(0,0.25^2) + 0.2N(0,0.5^2) + 0.2N(0,1^2) + 0.2N(0,2^2) )</td>
</tr>
<tr>
<td>near normal</td>
<td>( \frac{2}{3}N(0,1^2) + \frac{1}{3}N(0,2^2) )</td>
</tr>
<tr>
<td>skew</td>
<td>( \frac{1}{3}N(-2,2^2) + \frac{1}{3}N(-1,1.5^2) + \frac{1}{3}N(0,1^2) + \frac{1}{6}N(1,1^2) )</td>
</tr>
<tr>
<td>big-normal</td>
<td>( N(0,4^2) )</td>
</tr>
</tbody>
</table>

3.7 Real Data Analysis

We applied LPM to analyze 44 GWASs of complex traits integrated with 9 genic category annotations. The source of the summary statistics of GWASs is given in Table 3.2 and the genic category annotations are provided by ANNOVAR [Wang et al., 2010]. The genic category annotations include upstream, downstream, exonic, intergenic, intronic, ncRNA_exonic, ncRNA_intronic, UTR’3 and UTR’5, where ncRNA
Figure 3.29: FDR of LPM for identification of risk SNPs for (a) P1, (b) P2, (c) P1 and P2 when $P$-values are simulated from individual-level data. We controlled global FDR at 0.1 to evaluate empirical FDR. The results are summarized from 50 replications.
Figure 3.30: Type I error rate of LPM for the relationship test between the two traits when \(P\)-values are not from beta distribution. The bars represent one standard error. We evaluate type I error rate at 0.05 significance level. The results are summarized from 500 replications.

refers to RNA without coding annotation. We excluded the SNPs in the MHC region (Chromosome 6, 25-35 Mb) to avoid unusually large GWAS signals.

We show our results in three aspects: the estimated correlation among 44 GWASs (see Figure 3.32), the number of SNPs identified to be associated with each trait (see Figure 3.33), and the enrichment of genic category annotations (see Figures 3.34-3.36).

As shown in Figure 3.32, the correlations among traits are quite dense indicating that pleiotropy are pervasive. In particular, 810 out of 946 trait-pairs have a significant correlation at 0.05 level using Bonferroni correction. According to Figure 3.32, traits can be divided into several groups with relatively high correlations and these partitions are consistent with the categories of the traits. Main groups are group of psychiatric disorders which includes bipolar disorder [Psychiatric GWAS Consortium Bipolar Disorder Working Group, 2011], schizophrenia, neuroticism, depressive symptoms [Okbay et al., 2016a], major depressive disorder [Wray et al., 2018], attention deficit hyperactivity disorder [Demontis et al., 2017], autism spectrum disorder [Grove et al., 2017], and anorexia nervosa [Duncan et al., 2017]; group of red blood cell traits which includes mean cell haemoglobin concentration, mean cell haemoglobin, mean cell volume, red blood cell count, haemoglobin and packed
Figure 3.31: FDR of LPM for identification of risk SNPs for (a) P1, (b) P2, (c) P1 and P2 when P-values are not from beta distribution. We controlled global FDR at 0.1 to evaluate empirical FDR. The results are summarized from 50 replications.
Table 3.2: The source of the 44 GWASs.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at Menopause</td>
<td>Day et al. 2015, Nature Genetics.</td>
</tr>
<tr>
<td>Attention Deficit Hyperactivity Disorder</td>
<td>Demontis et al. 2017, bioRxiv. (ADHD-Pull GWAS)</td>
</tr>
<tr>
<td>BMI</td>
<td>Speliotes et al. 2010, Nature Genetics.</td>
</tr>
<tr>
<td>Celiac Disease</td>
<td>Dubois et al. 2014, Nature Genetics.</td>
</tr>
<tr>
<td>Coronary Artery Disease</td>
<td>Schunkert et al. 2011, Nature Genetics.</td>
</tr>
<tr>
<td>Depressive Symptoms</td>
<td>Okbay et al. 2016a, Nature Genetics.</td>
</tr>
<tr>
<td>Inflammatory Bowel Disease</td>
<td>Jostins et al. 2012, Nature.</td>
</tr>
<tr>
<td>Major Depressive Disorder</td>
<td>Wray et al. 2018, Nature Genetics. (MDD2 2018)</td>
</tr>
<tr>
<td>Multiple Sclerosis</td>
<td>Sawcer et al. 2011, Nature.</td>
</tr>
<tr>
<td>Primary Biliary Cirrhosis</td>
<td>Cordell et al. 2015, Nature Communications.</td>
</tr>
<tr>
<td>Pubertal Growth</td>
<td>Cousminer et al. 2013, Human Molecular Genetics.</td>
</tr>
<tr>
<td>Schizophrenia1</td>
<td>Cross-Disorder Group of the Psychiatric Genomics Consortium 2013b, The Lancet. (SCZ subset)</td>
</tr>
<tr>
<td>Schizophrenia2</td>
<td>Schizophrenia Psychiatric GWAS Consortium 2011, Nature Genetics. (SCZ1)</td>
</tr>
<tr>
<td>Schizophrenia3</td>
<td>Ripke et al. 2013, Nature Genetics. (Sweden+SCZ1)</td>
</tr>
<tr>
<td>Schizophrenia4</td>
<td>Ripke et al. 2014, Nature. (SCZ2)</td>
</tr>
<tr>
<td>Systemic Lupus Erythematosus</td>
<td>Bentham et al. 2015, Nature Genetics.</td>
</tr>
<tr>
<td>Total Cholesterol</td>
<td>Global Lipids Genetics Consortium 2013, Nature Genetics.</td>
</tr>
<tr>
<td>Type 1 Diabetes</td>
<td>Bradfield et al. 2011, PLoS Genetics.</td>
</tr>
<tr>
<td>Type 2 Diabetes</td>
<td>Morris et al. 2012, Nature Genetics.</td>
</tr>
<tr>
<td>Years of Education1</td>
<td>Rietveld et al. 2015, Science.</td>
</tr>
<tr>
<td>Years of Education2</td>
<td>Okbay et al. 2016b, Nature.</td>
</tr>
</tbody>
</table>
Figure 3.32: The estimated $\hat{R}$ for 44 GWASs with 9 genic category annotation integrated.
Figure 3.33: The number of variants identified to be associated with each of the 44 traits using LPM by four different analysis approaches: separately analysis without annotation, separately analysis with annotation, jointly analysis of the top 1 correlated trait with annotation and jointly analysis of the top 2 correlated traits with annotation. We controlled global FDR at 0.1. For visualization purpose, these numbers are normalized by dividing the corresponding number of variants identified by jointly analysis of the top 2 correlated traits with annotation.
**Figure 3.34:** The estimated coefficients of downstream, exonic, intergenic and intronic for 44 GWASs. The bars represent one standard error. The symbol ***,**, and * means the p value for the test of annotation enrichment is ≤ 0.05, ≤ 0.01 and ≤ 0.001 respectively.
Figure 3.35: The estimated coefficients of ncRNA_exonic, ncRNA_intronic, upstream and UTR3 for 44 GWASs. The bars represent one standard error. The symbol "**", "*" and "***" means the p value for the test of annotation enrichment is ≤ 0.05, ≤ 0.01 and ≤ 0.001 respectively.
<table>
<thead>
<tr>
<th>Trait</th>
<th>Estimate of Beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at Menopause</td>
<td>0.5</td>
</tr>
<tr>
<td>Alzheimer</td>
<td></td>
</tr>
<tr>
<td>Anorexia Nervosa</td>
<td>**</td>
</tr>
<tr>
<td>Atopic Dermatitis</td>
<td></td>
</tr>
<tr>
<td>Attention Deficit Hyperactivity Disorder</td>
<td>**</td>
</tr>
<tr>
<td>Autism Spectrum Disorder</td>
<td>**</td>
</tr>
<tr>
<td>Bipolar Disorder</td>
<td>**</td>
</tr>
<tr>
<td>BMI</td>
<td>**</td>
</tr>
<tr>
<td>Celiac Disease</td>
<td>**</td>
</tr>
<tr>
<td>Coronary Artery Disease</td>
<td></td>
</tr>
<tr>
<td>Crohn’s Disease</td>
<td>**</td>
</tr>
<tr>
<td>Depressive Symptoms</td>
<td>**</td>
</tr>
<tr>
<td>Education</td>
<td>**</td>
</tr>
<tr>
<td>Education 1</td>
<td>**</td>
</tr>
<tr>
<td>Ever Smoked</td>
<td>**</td>
</tr>
<tr>
<td>Fasting Glucose</td>
<td>**</td>
</tr>
<tr>
<td>Haemoglobin</td>
<td></td>
</tr>
<tr>
<td>Height</td>
<td>**</td>
</tr>
<tr>
<td>High-density Lipoprotein</td>
<td></td>
</tr>
<tr>
<td>Inflammatory Bowel Disease</td>
<td></td>
</tr>
<tr>
<td>Low-density Lipoprotein</td>
<td></td>
</tr>
<tr>
<td>Major Depressive Disorder</td>
<td>**</td>
</tr>
<tr>
<td>Mean Cell Haemoglobin</td>
<td>**</td>
</tr>
<tr>
<td>Mean Cell Haemoglobin Concentration</td>
<td></td>
</tr>
<tr>
<td>Mean Cell Volume</td>
<td>**</td>
</tr>
<tr>
<td>Multiple Schizophrenia</td>
<td></td>
</tr>
<tr>
<td>Neuroticism</td>
<td></td>
</tr>
<tr>
<td>Packing Cell Volume</td>
<td></td>
</tr>
<tr>
<td>Primary Biliary Cirrhosis</td>
<td>**</td>
</tr>
<tr>
<td>Pubertal Growth</td>
<td>**</td>
</tr>
<tr>
<td>Red Blood Cell Count</td>
<td></td>
</tr>
<tr>
<td>Rheumatoid Arthritis</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>**</td>
</tr>
<tr>
<td>Schizophrenia 1</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 2</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 3</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 4</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 4</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 5</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 6</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 7</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 8</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 9</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 10</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 11</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 12</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 13</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 14</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 15</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 16</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 17</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 18</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 19</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 20</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 21</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 22</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 23</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 24</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 25</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 26</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 27</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 28</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 29</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 30</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 31</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 32</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 33</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 34</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 35</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 36</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 37</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 38</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 39</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 40</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 41</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 42</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 43</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 44</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 45</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 46</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 47</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 48</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 49</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 50</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 51</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 52</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 53</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 54</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 55</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 56</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 57</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 58</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 59</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 60</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 61</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 62</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 63</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 64</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 65</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 66</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 67</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 68</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 69</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 70</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 71</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 72</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 73</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 74</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 75</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 76</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 77</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 78</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 79</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 80</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 81</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 82</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 83</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 84</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 85</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 86</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 87</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 88</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 89</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 90</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 91</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 92</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 93</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 94</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 95</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 96</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 97</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 98</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 99</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 100</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 101</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 102</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 103</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 104</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 105</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 106</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 107</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 108</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 109</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 110</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 111</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 112</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 113</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 114</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 115</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 116</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 117</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 118</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 119</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 120</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 121</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 122</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 123</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 124</td>
<td></td>
</tr>
<tr>
<td>Schizophrenia 125</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.36: The estimated coefficients of UTR’5 for 44 GWASs. The bars represent one standard error. The symbol **, *** and **** means the p value for the test of annotation enrichment is ≤ 0.05, ≤ 0.01 and ≤ 0.001 respectively.

Cell volume [Pickrell 2014], group of autoimmune diseases which includes systemic lupus erythematosus [Bentham et al. 2015], atopic dermatitis [The EArly Genetics and Consortium, Lifecourse Epidemiology (EAGLE) Eczema 2015], primary biliary cirrhosis [Cordell et al. 2015], Crohn’s disease, ulcerative colitis, inflammatory bowel disease [Jostins et al. 2012], celiac disease [Dubois et al. 2010], rheumatoid arthritis [Okada et al. 2014], multiple sclerosis [Sawcer et al. 2011] and type 1 diabetes [Bradfield et al. 2011] and group of lipid-related traits which includes high-density lipoprotein, triglycerides, low-density lipoprotein, and total cholesterol [Global Lipids Genetics Consortium 2013]. We also find some relationship between complex diseases and metabolic traits. For example, relatively high correlations are observed between coronary artery disease [Schunkert et al. 2011] and lipid-related traits, among type 2 diabetes [Morris et al. 2012], fasting glucose and fasting insulin [Manning et al. 2012]. We note that psychiatric disorders are correlated with many other traits, such as BMI [Speliotes et al. 2010], education, HIV [McLaren et al. 2013] and ever smoked [Tobacco, The and Consortium, Genetics 2010]. Similar evidence of relationship has been found by [Hartwig et al. 2016, Breslau et al. 2008, Chandra et al. 20125].
We also discover connections between height and pubertal growth, between age at menopause and fasting insulin, between Alzheimer and lipid-related traits. These findings may provide some new insights of complex traits.

If no annotation is integrated, the estimated correlations are shown in Figure 3.37. It does not change much compared with Figure 3.32. It is because of the presence of linkage disequilibrium (LD) effects.

Figure 3.37: The estimated $\hat{R}$ for 44 GWAS with no annotation integrated.
We also applied graph-GPA to infer the relationship among traits. As graph-GPA is not scalable to a large number of traits, we only analyze a subset of the traits. With the number of traits changes, the relationship estimated by graph-GPA changes a lot (see Figure 3.38). This is because graph-GPA represents the relationship from a conditional independent structure and the results are hard to interpret.

![Diagram showing relationships among traits](image)

Figure 3.38: Relationship inferred by graph-GPA when different numbers of GWASs are considered. The red edges denote the changes in the relationship among traits.

We compared the number of SNPs identified to be associated with each of the 44 traits using LPM by four different analysis approaches: separately analysis without annotation, separately analysis with annotation, jointly analysis of the top 1 correlated trait with annotation and jointly analysis of the top 2 correlated traits with annotation. The details of the top correlated traits are given in Table 3.3. Using the fourth approach as a reference, we calculated the ratio of the number of risk SNPs identified using each approach. Figure 3.33 shows that more risk SNPs can be identified by integrating genic category annotation and correlated traits. The differences between separately analysis with and without annotation are result from genic category annotations. For HIV and anorexia nervosa, a clear improvement is shown between the first two approaches, reflecting a large enrichment of genic category annotations. Consistently, the estimated coefficients of annotations for HIV and anorexia nervosa are relatively large as shown in Figures 3.34-3.36. The differences among approaches with annotations but different number of traits integrated are because of pleiotropy. Low-density lipoprotein (LDL) and total cholesterol (TC) are shown to be highly correlated in Figure 3.32. As a result, jointly analysis of these two diseases
lead to a large increase in the number of risk SNPs identified. The Manhattan plots are provided in Figure 3.39.

![Manhattan plots](image)

Figure 3.39: Manhattan plots of LDL and TC. Top left: separate analysis for LDL. Top right: joint analysis of LDL and TC for LDL. Bottom left: separate analysis for TC. Bottom right: joint analysis of LDL and TC for TC. The blue lines indicate local $\text{fdr} = 0.01$. The green points denote the additional SNPs identified in joint analysis with $\text{fdr} \leq 0.01$. For $\text{fdr}$s smaller than $10^{-30}$, we threshold them at $10^{-30}$.

The estimated coefficients of 9 genic category annotations are shown in Figures 3.34-3.36 and the results for the enrichment test are summarized in Table 3.4. We note that the coefficients for genic category annotations are positive for most traits except for intergenic. Intergenic has negative effects on many traits including puberal growth and ever smoked. By comparing the number of traits with significant coefficients for each genic category annotation, we find that exonic, UTR’3 and UTR’5 have wider
Table 3.3: The top two correlated traits of the 44 GWASs.

<table>
<thead>
<tr>
<th>Trait</th>
<th>top 1</th>
<th>top 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at Menopause</td>
<td>Fasting Insulin</td>
<td>BMI</td>
</tr>
<tr>
<td>Alzheimer</td>
<td>Total Cholesterol</td>
<td>Low-density Lipoprotein</td>
</tr>
<tr>
<td>Anorexia Nervosa</td>
<td>Major Depressive Disorder</td>
<td>Education2</td>
</tr>
<tr>
<td>Atopic Dermatitis</td>
<td>Ulcerative Colitis</td>
<td>Multiple Sclerosis</td>
</tr>
<tr>
<td>Attention Deficit Hyperactivity Disorder</td>
<td>Major Depressive Disorder</td>
<td>Autism Spectrum Disorder</td>
</tr>
<tr>
<td>Autism Spectrum Disorder</td>
<td>Attention Deficit Hyperactivity Disorder</td>
<td>Major Depressive Disorder</td>
</tr>
<tr>
<td>Bipolar Disorder</td>
<td>Schizophrenia2</td>
<td>Schizophrenia3</td>
</tr>
<tr>
<td>BMI</td>
<td>Education2</td>
<td>Major Depressive Disorder</td>
</tr>
<tr>
<td>Celiac Disease</td>
<td>Multiple Sclerosis</td>
<td>Ulcerative Colitis</td>
</tr>
<tr>
<td>Coronary Artery Disease</td>
<td>Triglycerides</td>
<td>Low-density Lipoprotein</td>
</tr>
<tr>
<td>Crohn’s Disease</td>
<td>Inflammatory Bowel Disease</td>
<td>Ulcerative Colitis</td>
</tr>
<tr>
<td>Depressive Symptoms</td>
<td>Neuroticism</td>
<td>Major Depressive Disorder</td>
</tr>
<tr>
<td>Education1</td>
<td>Education2</td>
<td>Attention Deficit Hyperactivity Disorder</td>
</tr>
<tr>
<td>Education2</td>
<td>Education1</td>
<td>Attention Deficit Hyperactivity Disorder</td>
</tr>
<tr>
<td>Ever Smoked</td>
<td>Education2</td>
<td>Major Depressive Disorder</td>
</tr>
<tr>
<td>Fasting Glucose</td>
<td>Type 2 Diabetes</td>
<td>Fasting Insulin</td>
</tr>
<tr>
<td>Fasting Insulin</td>
<td>Type 2 Diabetes</td>
<td>Triglycerides</td>
</tr>
<tr>
<td>Haemoglobin</td>
<td>Packed Cell Volume</td>
<td>Red Blood Cell Count</td>
</tr>
<tr>
<td>Height</td>
<td>Pubertal Growth</td>
<td>High-density Lipoprotein</td>
</tr>
<tr>
<td>High-density Lipoprotein</td>
<td>Triglycerides</td>
<td>Total Cholesterol</td>
</tr>
<tr>
<td>HIV</td>
<td>Crohn’s Disease</td>
<td>Schizophrenia1</td>
</tr>
<tr>
<td>Inflammatory Bowel Disease</td>
<td>Ulcerative Colitis</td>
<td>Crohn’s Disease</td>
</tr>
<tr>
<td>Low-density Lipoprotein</td>
<td>Total Cholesterol</td>
<td>Triglycerides</td>
</tr>
<tr>
<td>Major Depressive Disorder</td>
<td>Depressive Symptoms</td>
<td>Attention Deficit Hyperactivity Disorder</td>
</tr>
<tr>
<td>Mean Cell Haemoglobin</td>
<td>Mean Cell Volume</td>
<td>Red Blood Cell Count</td>
</tr>
<tr>
<td>Mean Cell Haemoglobin Concentration</td>
<td>Red Blood Cell Count</td>
<td>Mean Cell Haemoglobin</td>
</tr>
<tr>
<td>Mean Cell Volume</td>
<td>Mean Cell Haemoglobin</td>
<td>Red Blood Cell Count</td>
</tr>
<tr>
<td>Multiple Sclerosis</td>
<td>Type 1 Diabetes</td>
<td>Primary Biliary Cirrhosis</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>Depressive Symptoms</td>
<td>Major Depressive Disorder</td>
</tr>
<tr>
<td>Packed Cell Volume</td>
<td>Haemoglobin</td>
<td>Red Blood Cell Count</td>
</tr>
<tr>
<td>Primary Biliary Cirrhosis</td>
<td>Multiple Sclerosis</td>
<td>Rheumatoid Arthritis</td>
</tr>
<tr>
<td>Pubertal Growth</td>
<td>Height</td>
<td>BMI</td>
</tr>
<tr>
<td>Red Blood Cell Count</td>
<td>Haemoglobin</td>
<td>Packed Cell Volume</td>
</tr>
<tr>
<td>Rheumatoid Arthritis</td>
<td>Multiple Sclerosis</td>
<td>Primary Biliary Cirrhosis</td>
</tr>
<tr>
<td>Schizophrenia1</td>
<td>Schizophrenia2</td>
<td>Schizophrenia3</td>
</tr>
<tr>
<td>Schizophrenia2</td>
<td>Schizophrenia1</td>
<td>Schizophrenia3</td>
</tr>
<tr>
<td>Schizophrenia3</td>
<td>Schizophrenia2</td>
<td>Schizophrenia4</td>
</tr>
<tr>
<td>Schizophrenia4</td>
<td>Schizophrenia3</td>
<td>Schizophrenia2</td>
</tr>
<tr>
<td>Systemic Lupus Erythematosus</td>
<td>Primary Biliary Cirrhosis</td>
<td>Multiple Sclerosis</td>
</tr>
<tr>
<td>Total Cholesterol</td>
<td>Low-density Lipoprotein</td>
<td>Triglycerides</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>High-density Lipoprotein</td>
<td>Total Cholesterol</td>
</tr>
<tr>
<td>Type 1 Diabetes</td>
<td>Multiple Sclerosis</td>
<td>Primary Biliary Cirrhosis</td>
</tr>
<tr>
<td>Type 2 Diabetes</td>
<td>Fasting Insulin</td>
<td>Fasting Glucose</td>
</tr>
<tr>
<td>Ulcerative Colitis</td>
<td>Inflammatory Bowel Disease</td>
<td>Crohn’s Disease</td>
</tr>
</tbody>
</table>
effects among traits, indicating that SNPs in these regions are enriched for many complex traits. Our results are consistent with the previous study by Schork et al. [2013].

Table 3.4: The number of traits with significant coefficients for each genic category annotation at different significance level.

<table>
<thead>
<tr>
<th>significance level</th>
<th>downstream</th>
<th>exonic</th>
<th>intergenic</th>
<th>intronic</th>
<th>ncRNA exonic</th>
<th>ncRNA intronic</th>
<th>upstream</th>
<th>UTR3</th>
<th>UTR5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>0.01</td>
<td>5</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>0.05</td>
<td>12</td>
<td>18</td>
<td>5</td>
<td>9</td>
<td>7</td>
<td>7</td>
<td>11</td>
<td>14</td>
<td>14</td>
</tr>
</tbody>
</table>

Among the 44 GWASs, we analyzed four different GWASs of schizophrenia, schizophrenia1, schizophrenia2, schizophrenia3 and schizophrenia4. We find that the correlations among them are extremely high (see Table 3.5) and their enrichment of annotations are all similar (see Table 3.6), indicating that our method is replicable. The sample sizes for GWASs from schizophrenia1 to schizophrenia4 become larger, i.e., schizophrenia4 has the largest sample size (see Table 3.7). Correspondingly, the standard error of the estimated coefficients of schizophrenia4 are the smallest comparing with other three GWASs. As shown in Figure 3.33, more risk SNPs are identified to be associated with schizophrenia by jointly analyzing these GWASs. We also analyzed two education GWASs, Years of Education 1 and Years of Education 2 with a larger sample size. Similar results can be found.

Table 3.5: The estimated correlations among four schizophrenia GWASs.

<table>
<thead>
<tr>
<th></th>
<th>Schizophrenia1</th>
<th>Schizophrenia2</th>
<th>Schizophrenia3</th>
<th>Schizophrenia4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schizophrenia1</td>
<td>1</td>
<td>0.9851</td>
<td>0.9485</td>
<td>0.9267</td>
</tr>
<tr>
<td>Schizophrenia2</td>
<td>0.9851</td>
<td>1</td>
<td>0.9700</td>
<td>0.9275</td>
</tr>
<tr>
<td>Schizophrenia3</td>
<td>0.9485</td>
<td>0.9700</td>
<td>1</td>
<td>0.9590</td>
</tr>
<tr>
<td>Schizophrenia4</td>
<td>0.9267</td>
<td>0.9275</td>
<td>0.9590</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 3.6: The estimated coefficients of annotations for four schizophrenia GWASs.

<table>
<thead>
<tr>
<th></th>
<th>Schizophrenia1</th>
<th>Schizophrenia2</th>
<th>Schizophrenia3</th>
<th>Schizophrenia4</th>
</tr>
</thead>
<tbody>
<tr>
<td>downstream</td>
<td>0.4338 (0.3469)</td>
<td>0.1978 (0.2698)</td>
<td>0.2938 (0.2206)</td>
<td>0.2655 (0.1617)</td>
</tr>
<tr>
<td>exonic</td>
<td>0.5406 (0.3553)</td>
<td>0.3241 (0.2772)</td>
<td>0.4612 (0.2257)</td>
<td>0.3436 (0.1655)</td>
</tr>
<tr>
<td>intergenic</td>
<td>0.3564 (0.3523)</td>
<td>0.0870 (0.2745)</td>
<td>0.0997 (0.2236)</td>
<td>0.1303 (0.1640)</td>
</tr>
<tr>
<td>intronic</td>
<td>0.5526 (0.3524)</td>
<td>0.3307 (0.2745)</td>
<td>0.3736 (0.2236)</td>
<td>0.4047 (0.1640)</td>
</tr>
<tr>
<td>ncRNA_exonic</td>
<td>0.7073 (0.3644)</td>
<td>0.4065 (0.2854)</td>
<td>0.4883 (0.2322)</td>
<td>0.2759 (0.1703)</td>
</tr>
<tr>
<td>ncRNA_intronic</td>
<td>0.1963 (0.3535)</td>
<td>-0.0330 (0.2756)</td>
<td>0.1639 (0.2243)</td>
<td>0.1545 (0.1645)</td>
</tr>
<tr>
<td>upstream</td>
<td>0.5568 (0.3474)</td>
<td>0.2426 (0.2700)</td>
<td>0.2986 (0.2207)</td>
<td>0.2702 (0.1618)</td>
</tr>
<tr>
<td>UTR3</td>
<td>0.6340 (0.3552)</td>
<td>0.3957 (0.2773)</td>
<td>0.4404 (0.2258)</td>
<td>0.4779 (0.1655)</td>
</tr>
<tr>
<td>UTR5</td>
<td>0.4916 (0.3740)</td>
<td>0.2863 (0.2950)</td>
<td>0.4978 (0.2392)</td>
<td>0.3919 (0.1746)</td>
</tr>
</tbody>
</table>

The values in the brackets are standard errors of the estimates.

Table 3.7: Sample sizes for four schizophrenia GWASs.

<table>
<thead>
<tr>
<th></th>
<th>Schizophrenia1</th>
<th>Schizophrenia2</th>
<th>Schizophrenia3</th>
<th>Schizophrenia4</th>
</tr>
</thead>
<tbody>
<tr>
<td>case</td>
<td>9,379</td>
<td>9,394</td>
<td>13,833</td>
<td>36,989</td>
</tr>
<tr>
<td>control</td>
<td>7,736</td>
<td>12,462</td>
<td>18,310</td>
<td>113,075</td>
</tr>
</tbody>
</table>
Chapter 4

Conclusion

This thesis has discussed two statistical methods for integrative analysis of genomic data. The first one, which is called LSMM, is designed to integrate functional annotations with GWAS data. It can integrate both genic category annotations and a large amount of cell-type specific functional annotations with GWAS data. LSMM can not only improve the statistical power in the identification of risk SNPs, but also infer relevant cell-type specific functional annotations to the phenotype, offering new insights to explore the genetic architecture of complex traits or diseases. Through comprehensive simulations and real data analysis of 30 GWAS, LSMM is shown to be statistically efficient and computationally scalable. As more annotation data will become publicly available in the future, we believe LSMM is widely useful for integrative analysis of genomic data.

The second one, which is named LPM, is proposed to integrate multiple GWASs with functional annotation. This unified framework can characterize relationship among complex traits, increase the statistical power for association mapping, integrate and investigate the effect of functional annotations simultaneously. With extensive simulations and real data analysis of 44 GWASs, we have shown the statistical efficiency and computational scalability of LPM.

Besides the promising results, our methods can still make some improvements. First, we did not take the linkage disequilibrium (LD) into account and simply assumed the independence among SNPs in both models. However, the presence of LD may strongly influence the identification of risk SNPs for one or more traits. As a
result, it is an important direction of future work to consider the LD effects. Second, other source of omics information can be integrated, such as gene expression data and DNA methylation. It can further help us to understand the aetiology of complex traits.
Bibliography


Demontis, D. et al. (2017). Discovery of the first genome-wide significant risk loci for ADHD. *bioRxiv*.


The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145), 661–678.


CURRICULUM VITAE

Academic qualifications of the thesis author, Ms. MING Jingsi:

- Received the degree of Bachelor of Science (Statistics) from Fudan University, July 2013.

- Received the degree of Master of Finance (Financial Engineering) from Fudan University, July 2015.

August 2018