Ph.D. Dissertation of Science

Pathway-based approach using hierarchical structural component models to analyze multinomial phenotypes

다항 표현형 자료를 이용한 패스웨이 분석 계층적 구조 모형

February 2023

Graduate School
Seoul National University
Statistics Major

Md Kamruzzaman

Pathway-based approach using hierarchical structural component models to analyze multinomial phenotypes

Taesung Park

Submitting a Ph.D. Dissertation of
Statistics

October 2022

Graduate School
Seoul National University
Statistics Major

Md Kamruzzaman

Confirming the Ph.D. Dissertation written by
Md Kamruzzaman
December 2022

Chair	Hee-Seok Oh	(Seal)
Vice Chair	Taesung Park	(Seal)
Examiner	Myunghee Cho Paik	(Seal)
Examiner	Johan Lim	(Seal)
Examiner	Sungkyoung Choi	(Seal)

Abstract

Pathway-based approach using hierarchical structural component models to analyze multinomial phenotypes

Md Kamruzzaman
Department of Statistics
The Graduate School
Seoul National University

To identify novel pathways from markers associated with a particular disease, several statistical methods of pathway analysis have been applied. However, most of the available methods are based on single pathway analyses and do not consider multiple pathways simultaneously. Since pathways are highly correlated, multiple pathway analyses suffer from this correlation. To address this issue, a hierarchical structural component model (HisCoM) was developed, which considered all pathways at the same time, as well as the correlations among them. HisCoM has been successfully applied to the analysis of continuous, counts, and binary phenotypes but it is not readily applicable for analyzing multinomial phenotypes.

In this thesis, we propose novel statistical methods, the hierarchical structural component analysis for multinomial phenotypes (HisCoM-Categ), and hierarchical structural component analysis for longitudinal data with multinomial phenotypes (HisCoM-RCateg). In addition, we also propose a parametric testing approach

rather than a permutation approach for HisCoM to find the association between pathways and phenotypes.

As the existing HisCoM, HisCoM-Categ considers the biomarker and pathway hierarchies while accounting for the correlations of all pathways by using the ridge penalty. For identifying the association between pathways and phenotype, HisCoM-Categ uses the baseline category logit model for nominal phenotypes and the proportional odds model for ordinal phenotypes. HisCoM-RCateg is an extended version of HisCoM-Categ for longitudinal multinomial phenotypes. Like HisCoM-Categ, HisCoM-RCateg can also identify the significant pathways associated with the desired phenotype by analyzing all pathways at a same time. Both HisCoM-Categ and HisCoM-RCateg are flexible enough to be used for various types of omics data. For example, we used our HisCoM-Categ and HisCoM-RCateg methods on a real metabolomic dataset from the Korean Association Resource (KARE) to identify the association between metabolite pathways and type 2 diabetics (T2D). It is noted that T2D is a metabolic disease affected by multiple genetic factors, which is a major public health concern. Application to the KARE metabolite dataset demonstrates that HisCoM-Categ and HisCoM-RCateg are able to identify the pathways that are associated with T2D. Through simulation study, we also show that HisCoM-Categ and HisCoM-RCateg perform better than other methods.

Keywords: Pathway analysis, hierarchical structure, longitudinal data, multinomial phenotype, parametric testing.

Student Number: 2018-34194

Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1. Omics data analysis with biological context

1.1.1 Definition of omics data

The word "omics" refers to a variety biological science fields of research that seeks to characterize and quantify collections of biological molecules that translate into the structure, function, and mapping of an organism or organisms [1]. The suffix "omics" identifies members of the omics group, which includes genomics, transcriptomics, proteomics, and metabolomics. Genomics, transcriptomics, proteomics, and metabolomics stages make up the transmission of every single cell (Figure 1.1) [2].

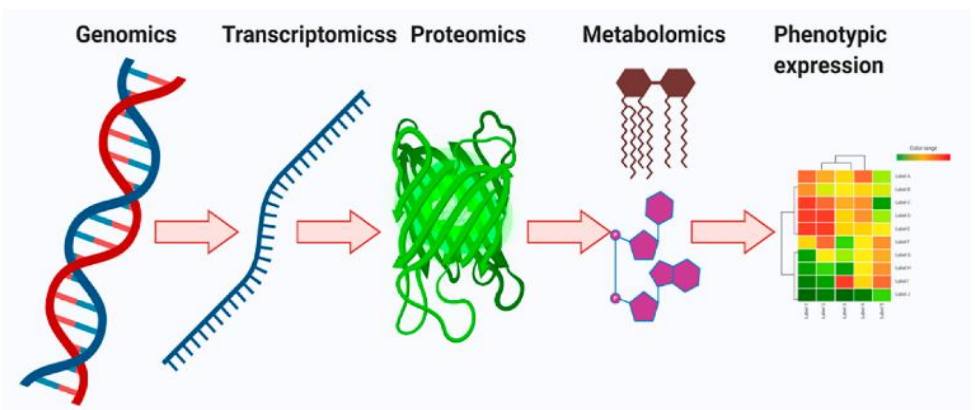
Genomics is central to the science of biology [3] and is the study of the complete set of DNA in an organism, including all of its gene [1]. A human genome has approximately 3 billion DNA base pairs which are distributed across 23 pairs of chromosomes [4]. DNA is structured of two bases that contains nitrogen that couple up to form the molecule. Adenine (A), cytosine (C), guanine (G), and thymine (T) are the four bases of DNA. These bases come in particular pairs (A with T, and G with C). A genomic variant at a single base position in the DNA is known as single nucleotide polymorphism (SNP). Because of the development of high-throughput genomics technology Genome-wide association studies (GWAS) become a widely used stategy to identify associations between SNP and a complex disease of interest such that type 2 diabetics, parkinson's disease, crohn's disease etc [5].

Transcriptomics is one of the popular topics in biology. Transcriptomics is the study of the transcriptome, which is a complete set of all RNA (including mRNA, miRNA) molecules expressed in cell, tissue or organism [6].

The extensive study of proteins, their structure, and their physiological role or function is known as proteomics [7]. Most of the functional information of genes is characterized by the proteome [8]. Identification of the protein or collection of proteins that cause a particular disease is the aim of proteomics [9].

A thorough examination of the metabolites in a biological specimen is called metabolomics [10]. It is used as a complementary approach to genomics, transcriptomics and proteomics [11]. Combination of two or more omics datasets is known as multi-omics data [12]. The major molecules of each omics, such as common and rare variants in genomics, genes in transcriptomics, proteins in proteomics, and metabolites in metabolomics, are collectively referred to as biomarkers in this thesis.

Figure 1.1. Fundamental principle for multi-omics profiling in system biology [2].



1.1.2 Pathways

Since biomarkers interact with one another, they do not work alone. A series of interactions among molecular biomarkers in a cell makes a biological pathway. It can initiate to manufacture of new molecular biomarker like proteins or lipids. Cells are continually receiving chemical cues from both inside and outside the body that are prompted on by injury, infection, stress, etc. Sometimes biological pathways do not work properly. The dysregulation of multiple biomarkers connected in a pathways is caused by complex diseases [13].

Pathway analysis aids the understanding of various omics data collected from high-throughput sequencing methods by using the pre-existing biological knowledge of pathways. Pathway analysis is mostly used to assess the relationship between a disease status and a pathway that consists of a set of biomarkers. Many statistical approaches of pathway analysis have been developed to identify novel pathways connected with a phenotype.

1.1.3 Statistical approach for analyzing omics data

A common approach of association studies in omics data is to search for the relationship between a single biomarker and phenotype. For example, GWAS typically focuses on single SNP biomarker analysis, and it is effective in identifying SNP with large effects. But even with high sample numbers, most biomarkers for complicated diseases have tiny effects, making them challenging to detect [14]. For this reason, instead of analyzing one biomarker at a time, analyze a group of biomarkers that are associated with complex diseases.

Analyze multiple biomarkers together, one such approach is gene set analysis, also known as pathway analysis, which uses prior biological knowledge of gene function. The pathway-based approaches typically examine whether a group of related biomarkers in the same functional pathway are jointly associated with a phenotype of interest [15]. For pathway analysis, several methods were proposed. Gene Set Enrichment Analysis (GSEA), the most widely used analytical technique for pathway analysis using gene expression microarray data [16]. The Kolmogorov-Smirnov statistic is used by GSEA to measure the degree of differential gene expression in a gene-set. Again, the GSEA method was adopted using a minimum p-value approach to analyze the GWAS data motivated by pathway-based methods of microarray data [17]. Similar to the GSEA, metabolite set enrichment analysis (MSEA) was developed to investigate the biological pathway for human and/or mammalian metabolic studies [18].

Similar to the pathway based approach, an adaptive so-called sum of powered score (aSPU) was developed for identifying the association between phenotype and a group of predictors of interest [19]. Later, aSPU for multiple SNPs in a pathway (aSPUpath) to test the association between pathway and phenotype was developed by extending aSPU [20]. Treating an SNP as an ordinal phenotype, POMaSPU was proposed for proportional odds model (POM) to identify the association between SNP and multiple predictors [21].

Those previous pathway methods and association tests consider one pathway at a time. As a result, the correlation among pathways is not considered. Since some of the biomarkers are shared between several pathways simultaneously, which makes high

correlation between pathways. Without considering this correlation results may be wrong. To account for these issues, a pathway based approach using hierarchical structural components of collapsed rare variants (PHARAOH) was developed [22]. PHARAOH used a hierarchical structure of biomarkers and pathways in the model and can analyze the associations between a phenotype and all pathways simultaneously. To account for the correlation among biomarkers and pathways, PHARAOH used ridge penalty on both biomarker and pathway levels. Following PHARAOH, hierarchical structural component analysis of miRNA−mRNA integration (HisCoM−mimi) method have been developed to investigate how miRNA indirectly affect the phenotype accounting for biological relationships between miRNA and mRNA [23]. By taking the advantage of HisCoM model, HisCoM−PAGE was proposed for gene expression data for the survival phenotype [24]. For the survival phenotype, mimi−surv was proposed to identify the significant miRNA−mRNA sets associated with survival phenotype [25]. Recently, DeepHisCoM has been developed that employs deep learning methods to discover the impact of pathway together with complex biomarkers contributions to the phenotype [26]. Later, by expanding the kernel machine regression, the HisCoM−kernel was proposed to identify the non−linear relationships between biomarkers and phenotypes [27]. All the previously developed HisCoM models can handle continuous, binary or survival phenotypes. Following the PHARAOH, PHARAOH−Multi and PHARAOH−GEE was developed for pathway analysis for multiple phenotypes and cluster phenotypes, respectively [28, 29]. However, these approaches are not directly suitable for analyzing multinomial phenotypes. Thus, a pathway approach using

Hierarchical structural Component analysis for multinomial phenotype is needed.

1.2. Objective of the study

The primary purpose of this study is to develop novel statistical methods for pathway analysis of the multinomial phenotypes. Since some of the biomarkers are shared between several pathways simultaneously, which makes high correlation between pathways. Thus, we focus on analyzing multiple pathways simultaneously in a single model using HisCoM. In the first study, we propose HisCoM-Categ by extending HisCoM for pathway analysis of multinomial phenotypes. As the existing HisCoM, the proposed HisCoM-Categ considers the biomarker and pathway hierarchies with accounting the correlations of all pathways using the ridge penalty. For identifying the association between pathways and phenotypes, HisCoM-Categ uses nominal phenotypes as well as ordinal phenotypes.

In the second study, we develop an extended version of HisCoM-Categ for longitudinal multinomial phenotypes using generalized estimating equations approach (HisCoM-RCateg).

To evaluate the significance of association between pathways and phenotype, HisCoM uses the permutation test. Like as HisCoM, HisCoM-Categ and HisCoM-RCateg use the permutation approach for testing the effect of pathways to phenotypes. Finally, in the third study, we develop a parametric test approach of HisCoM to reduce the computational burden and time of the permutation test.

1.3. Layout of the thesis

The structure of thesis is as follows. Chapter 1 is an introduction and goal of this study. Chapter 2 contains the review of the existing pathway-based methods and models for multinomial phenotypes. Chapter 3 and Chapter 4 describe in detail the proposed methods HisCoM-Categ and HisCoM-RCateg, including simulation studies and real data applications. Chapter 5 introduces the parametric testing approach of HisCoM. Lastly, Chapter 6 presents a summary and conclusion of this thesis.

Chapter 2. Review of existing pathway-based methods and models for multinomial phenotypes

2.1. Review of single pathway-based methods

Pathway analysis is a powerful method for analyzing large-scale omics data. Pathway analysis provides a thorough understanding of the molecular processes underlying complex diseases [17]. Several different pathway-based approaches have been developed recently to analyze different kinds of omics data.

2.1.1 Gene set enrichment analysis (GSEA)

For the GSEA method, we consider the total number of biomarkers is N and the predefined pathway set is S. First, we fit the univariate ordinal regression for N biomarkers and compute their regression coefficients ($\hat{\beta}$) and corresponding t-statistic ($=\hat{\beta}/se(\hat{\beta})$). Second, we rank order the N biomarkers according to the t-statistic value ($t_{(1)} < \cdots < t_{(N)}$). Then, we compute the enrichment score (ES) using

$$\mathrm{ES} = \max_{1 \le i \le k} \{|P_{hit} - P_{miss}|\}$$

where

$$P_{hit} = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|t_{(j)}|^p}{N_R}, \text{where } N_R = \sum_{g_j \in S} |t_{(j)}|^p$$

$$P_{miss} = \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{N - N_H}$$

with N_H is the number of biomarkers not in the pathway. When $p = 0$, GSEA reduces to the standard Kolmogorov−Smirnov statistics, that GSEA1. By comparing the observed ES with the permutation distribution values of ES, we evaluate the significance level.

2.1.2 An adaptive sum of power score (aSPU)

Let $y_i^* \in \{1,2,\dots,(J+1) > 2\}$ be the ordinal phenotype for the i^{th} $(i = 1,2,\dots,n)$ subject that can take one of J levels. Let $\boldsymbol{X}_i = \left(x_{i1}, x_{i2}, \dots, x_{ip}\right)^T$ is p multiple markers in a single pathway and $\boldsymbol{Z}_i = (z_{i1}, z_{i2}, \dots, z_{il})^T$ is l adjusting covariates. The POM can be written as

$$logit[\Pr(y_i^* \le j)] = \alpha_j + \boldsymbol{Z}_i\boldsymbol{\delta} + \boldsymbol{X}_i\boldsymbol{\beta}.$$

We want to test the null hypothesis $H_0: \boldsymbol{\beta} = \left(\beta_1, \dots, \beta_p\right)^T = \mathbf{0}$, that is, there is no association between any biomarkers in pathway and phenotype. Suppose U_k is the $k^{th}(k = 1,2,\dots,p)$ component of the score vector $\boldsymbol{U} = \left(U_1, \dots, U_p\right)^T$. For an integer $\gamma \ge 1$, the test statistic of $SPU(\gamma)$ can be defined as

$$SPU(\gamma) = \sum_{k=1}^{p} U_k^{\gamma}.$$

Since we are unsure of which γ value will produce a high power of $SPU(\gamma)$, thus an adaptive SPU test is developed

$$aSPU(\gamma) = \min_{\gamma \in \Gamma} P_{SPU(\gamma)}$$

where the p−value of $SPU(\gamma)$ is $P_{SPU(\gamma)}$ be, and Γ be a set of $\gamma \ge 1$; for $\Gamma = \{1,2,\dots,8,\infty\}$ was used the good performance of the numerical

study. Finally, the permutation approach was used calculate the p-values of all the SPU and aSPU tests.

2.2. Review of multiple pathway-based method: The PHARAOH method

The PHARAOH is a pathway-based approach that uses a hierarchy of rare variant-gene-pathway. A key feature of PHARAOH is the analysis of the entire pathways with a single model:

$$\eta_j = \beta_0 + \sum_{k=1}^{K}\left[\sum_{m=1}^{M_k} x_{jkm} w_{km}\right]\beta_k = \beta_0 + \sum_{k=1}^{K} f_{jk}\beta_k,$$

where η_j is a linear predictor for j^{th} individual, $\boldsymbol{\beta} = [\beta_0 \quad \beta_1 \quad \cdots \quad \beta_K]^T$, K is the number of pathways, x_{jkm} is a value of the m^{th} biomarker in the k^{th} pathway for j^{th} individual, w_{km} is the m^{th} biomarker effect size belonging to the k^{th} pathway, and the number of biomarkers in the k^{th} pathway is M_k.

An alternating least squares (ALS) algorithm was used for solving the following penalized log-likelihood equation to estimate the parameters for PHAROH,

$$\delta = \sum_{i=1}^{n} \log p(y_i; w_{km}, \beta_k) - \frac{1}{2}\lambda_m \sum_{k=1}^{K}\sum_{m=1}^{M_k} w_{km}^2 - \frac{1}{2}\lambda_p \sum_{k=1}^{K} \beta_k^2,$$

where $p(y_i; w_{km}, \beta_k)$ be the probability density function of phenotype for individual i, λ_m and λ_p are the associated tuning parameters corresponding to the biomarkers and pathways, respectively.

The objective function δ was maximized using the iterative reweighted least squares (IRWLS) algorithm. Minimizing the

following penalized least−squares function is equivalent to the maximizing the above objective function δ,

$$\delta = \sum_{i=1}^{n} v_i \left(z_i - \sum_{k}^{K} f_{ik}\beta_k \right)^2 - \frac{1}{2}\lambda_m \sum_{k=1}^{K}\sum_{m=1}^{M_k} w_{km}^2 - \frac{1}{2}\lambda_p \sum_{k=1}^{K} \beta_k^2$$

$$= (\boldsymbol{z} - \boldsymbol{F\beta})^T \boldsymbol{V} (\boldsymbol{z} - \boldsymbol{F\beta}) - \frac{1}{2}\lambda_m \sum_{k=1}^{K}\sum_{m=1}^{M_k} w_{km}^2 - \frac{1}{2}\lambda_p \sum_{k=1}^{K} \beta_k^2,$$

where $\boldsymbol{z}$ is an adjusted response variable with elements $z_i = \eta_i + (y_i - \mu_i)/v_{ji}$, $\boldsymbol{F} = [\boldsymbol{f}_1 \ \cdots \ \boldsymbol{f}_N]^T$ is a latent matrix representing pathways, $\boldsymbol{V}$ is a diagonal matrix with elements $v_j = \left(\partial\mu_j/\partial\eta_j\right)^2/\tau_j$, and τ_j is the variance function. PHARAOH accounts for both correlations between pathways and correlations between biomarkers by imposing ridge penalties (i.e. λ_m, λ_P) on the pathway and gene effects. The ALS algorithm is used to estimate the parameters $\boldsymbol{w}$ and $\boldsymbol{\beta}$. The ALS algorithm iterates the two steps of estimating two parameters by estimating one parameter given the other parameter fixed at a time. The statistical significance is calculated using a permutation test which permutes the phenotype.

2.3. Review of regression for multinomial phenotypes

Let $y_i^* \in \{1,2,\ldots,J > 2\}$ be the multinomial phenotype for subject i $(i = 1,2,\ldots,n)$ that can take one of J levels. Let y_{ij} be the binary variable for $j = 1,2,\ldots,J$, where $y_{ij} = 1$ when subject i is in category j and $y_{ij} = 0$ otherwise. We define the $(J-1)\times 1$ response vector for the i^{th} subject $\boldsymbol{y}_i = \left(y_{i1}, y_{i2}, \ldots, y_{iJ}\right)'$, in which we omitted $y_{i,J+1}$ since

$\sum_{j=1}^{J} y_{ij} = 1$. Let $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})^T$ denote explanatory variable values for subject i. Let $\pi_j(x_i) = P(y_i^* = j|x_i)$.

2.3.1 Nominal phenotypes: Baseline−Category logits models

Consider the response variable y_i^* is nominal. It has no natural ordering. Then the general baseline−category logit model becomes

$$\log\left(\frac{\pi_j(\boldsymbol{x}_i)}{\pi_J(\boldsymbol{x}_i)}\right) = \beta_{0j} + \boldsymbol{\beta}_j^T \boldsymbol{x}_i, \qquad j = 1, \ldots, J-1.$$

where $\beta_j = (\beta_{j1}, \ldots, \beta_{jp})^T$ denote the parameters for the j^{th} baseline−category logit.

2.3.2 Ordinal phenotypes

Cumulative logit model

Consider the phenotype y_i^* is ordinal. Cumulative logit model for ordinal data can be define as

$$\text{logit}(P(y_i^* \leq j|\boldsymbol{x}_i)) = \beta_{0j} + \boldsymbol{\beta}^T \boldsymbol{x}_i, \qquad j = 1, \ldots, J-1,$$

where β_{0j} is the category−specific intercept, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ for the parameters associated with covariates.

Latent variable motivation for cumulative logit models

Let U be the underlaying latent variable and consider

$$U = -\beta^T x + \varepsilon$$

where ε has a standard logistic distribution with cumulative distribution function (cdf) with

$$\mathrm{P}(\varepsilon \le u) = \frac{e^u}{1 + e^u}.$$

Then

$$\begin{aligned} P(Y \le j) &= P\left(U \le \alpha_j\right) \\ &= P\left(-\beta^T x + \varepsilon \le \alpha_j\right) \\ &= P\left(\varepsilon \le \alpha_j + \beta^T x\right) \\ &= \frac{e^{\alpha_j + \beta^T x}}{1 + e^{\alpha_j + \beta^T x}}. \end{aligned}$$

Thus, the ordinal response y^* can be determined by category-specific intercept β_{0j} according to the thresholds

$$Y = j \quad if\ \beta_{0,j-1} < u \le \beta_{0j}$$

where

$$-\infty = \beta_{00} < \beta_{01} < \beta_{02} < \cdots < \beta_{0J} = \infty.$$

Adjacent categories logit models

The adjacent-categories logits for ordinal phenotypes are defined by

$$\mathrm{logit}\left(P(y_i^* \le j | y_i^* \in \{j, j+1\})\right) = \log\left(\frac{\pi_j}{\pi_{j+1}}\right), j = 1, \ldots, J-1.$$

The proportional odds from of the adjacent-categories logit model can be defined as

$$\log\left(\frac{\pi_j(\boldsymbol{x}_i)}{\pi_{j+1}(\boldsymbol{x}_i)}\right) = \beta_{0j} + \boldsymbol{\beta}^T \boldsymbol{x}_i, \qquad j = 1, \ldots, J-1.$$

2.4. Generalized estimating equations for multinomial phenotypes

Let $Y_{it}^{*} \in \{1,2,\dots,(J+1)>2\}$ be the multinomial phenotype for $i^{th}\,(i=1,2,\dots,\mathrm{n})$ subject at $t^{th}\,(t=1,2,\dots,\mathrm{T})$ time point. Define a binary random variable Y_{itj}^{*} for $j=1,2,\dots,(J+1)$ category, where $Y_{itj}^{*}=1$ when i^{th} subject has j^{th} response category at t^{th} time and $Y_{itj}^{*}=0$ otherwise. We convert Y_{it}^{*} into the $J\times 1$ vector $\boldsymbol{Y}_{it}=\left(Y_{it1}^{*},\dots,Y_{itJ}^{*}\right)'$, in which we omitted $Y_{it,J+1}^{*}$ since $\sum_{j}^{J+1}Y_{itj}^{*}=1$. Then, the phenotype vector for the i^{th} subject $\boldsymbol{Y}_i=(\boldsymbol{Y}_{i1}',\boldsymbol{Y}_{i2}',\dots,\boldsymbol{Y}_{iT}')'$: is $TJ\times 1$ vector. Suppose $\boldsymbol{x}_{it}'=\left(x_{it1},x_{it2},\dots,x_{itp}\right)$ is a $p\times 1$ vector of explanatory variables. We also consider $\boldsymbol{Z}_{it}$ is a $J\times(J+k)$ covariate matrix for time t including the intercept, time−stationary, time−varying, and category−specific which is composed from $\boldsymbol{x}_{it}$. Then for i^{th} subject, $\boldsymbol{Z}_i=(\boldsymbol{Z}_{i1}',\dots,\boldsymbol{Z}_{iT}')'$ is the $TJ\times(J+k)$ covariate matrix. The marginal density of $\boldsymbol{Y}_{it}$ is,

$$f(\boldsymbol{y}_{it}|\boldsymbol{Z}_{it};\boldsymbol{\beta})=\prod_{j=1}^{J}\pi_{itj}^{y_{itj}},$$

where $\pi_{itj}=\pi_{itj}(\boldsymbol{\beta})=E\left(Y_{itj}\middle|\boldsymbol{Z}_{it};\boldsymbol{\beta}\right)=\Pr\left(Y_{itj}=1\middle|\boldsymbol{Z}_{it};\boldsymbol{\beta}\right)$ be the probability of the j^{th} phenotype at t^{th} time, and $\boldsymbol{\beta}=(\boldsymbol{\beta_0}'\ \vdots\ \boldsymbol{\beta_x}')$ be a $(J+k)\times 1$ vector of parameters, where the $J\times 1$ vector of category−specific intercepts is $\boldsymbol{\beta_0}=\left(\beta_{01},\beta_{02},\dots,\beta_{0J}\right)'$ and $\boldsymbol{\beta_x}$ is the $k\times 1$ vector of parameters associated with variables. Suppose the marginal probability vector $\boldsymbol{\pi}_i=E(\boldsymbol{Y}_i|\boldsymbol{Z}_i)=(\boldsymbol{\pi}_{i1}',\dots,\boldsymbol{\pi}_{iT}')'$ represents the $TJ\times 1$ mean vector of $\boldsymbol{Y_i}$, where $\boldsymbol{\pi}_{it}=\left(\pi_{it1},\dots,\pi_{itJ}\right)'$.

Let $\boldsymbol{g}:(0,1)^{J}\to\boldsymbol{R}^{J}:(J\times 1)$ be a vector of link functions and we use a multinomial generalized linear model [30] to model the marginal expected vector $\boldsymbol{\pi}_{it}=E(\boldsymbol{Y}_{it}|\boldsymbol{Z}_{it})$ for subject i at time t,

$$\boldsymbol{g}[E(\boldsymbol{Y}_{it}|\boldsymbol{Z}_{it})] = \boldsymbol{g}(\boldsymbol{\pi}_{it}) = \boldsymbol{Z}_{it}\boldsymbol{\beta},$$

where the vector of link functions is chosen such that it consists baseline−category logit functions for nominal responses and cumulative logit link functions or adjacent−categories logit functions for ordinal responses.

The estimate the $\boldsymbol{\beta}$, the generalized estimating equations was solved [31, 32] ,

$$\boldsymbol{S}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{D}_i^T(\boldsymbol{\beta})\boldsymbol{V}_i^{-1}(\boldsymbol{\beta})\big(\boldsymbol{Y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta})\big) = \boldsymbol{0},$$

where $\boldsymbol{D}_i(\boldsymbol{\beta}) = \frac{\partial \boldsymbol{\pi}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$, and $\boldsymbol{V}_i(\boldsymbol{\beta})$ is a $TJ \times TJ$ "working" covariance matrix for $\boldsymbol{Y}_i$ [31, 32]. The covariance matrix $\boldsymbol{V}_i(\boldsymbol{\beta})$ can be decomposed in terms of the working correlation matrix $\boldsymbol{R}_i(\boldsymbol{\alpha})$ and $\boldsymbol{V}_i(\boldsymbol{\beta}) = \boldsymbol{A}_i^{\frac{1}{2}}(\boldsymbol{\beta})\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{A}_i^{\frac{1}{2}}(\boldsymbol{\beta})$, where $\boldsymbol{A}_i$ is the matrix of marginal variances, $\boldsymbol{A}_{it}$, given by $\boldsymbol{A}_{it} = diag\big[\pi_{it1}(1-\pi_{it1}),\ \ldots, \pi_{itJ}\big(1-\pi_{itJ}\big)\big]$ and also $\boldsymbol{A}_i^{\frac{1}{2}} = diag\left[\boldsymbol{A}_{i1}^{\frac{1}{2}}, \ldots, \boldsymbol{A}_{iT}^{\frac{1}{2}}\right]$. Then, $\boldsymbol{R}_i(\boldsymbol{\alpha}) = \boldsymbol{A}_{it}^{-\frac{1}{2}}\boldsymbol{V}_{it}\boldsymbol{A}_{it}^{-\frac{1}{2}}$ is the $J \times J$ diagonal blocks for the correlation matrix $\boldsymbol{Y}_i$, where the $J \times J$ diagonal blocks of $\boldsymbol{V}_i$ is $\boldsymbol{V}_{it} = diag(\boldsymbol{\pi}_{it}) - \boldsymbol{\pi}_{it}\boldsymbol{\pi}_{it}^T$ and the $J \times J$ off−diagonal blocks are $\boldsymbol{\rho}_{itt'} = \boldsymbol{\rho}_{itt'}(\boldsymbol{\alpha}) = Corr(\boldsymbol{Y}_{it}, \boldsymbol{Y}_{it'}), t \neq t'$. Define $\boldsymbol{e}_{it} = \boldsymbol{A}_{it}^{-\frac{1}{2}}(\boldsymbol{Y}_{it} - \boldsymbol{\pi}_{it})$ be the vector of Pearson residual. Then, it follows that

$$\boldsymbol{\rho}_{itt'}(\boldsymbol{\alpha}) = Corr(\boldsymbol{Y}_{it}, \boldsymbol{Y}_{it'}) = E(\boldsymbol{e}_{it}\boldsymbol{e}_{it'}).$$

A various number of working correlation matrices including exchangeable, unstructured etc. were adopted. Finally, the vector of unknown parameters $\boldsymbol{\alpha}$ for the working correlation matrices can be estimated by the method of moments [31].

Chapter 3. Pathway-based Approach using Hierarchical Structural Component Models to Analyze Multinomial Phenotypes

3.1. Introduction

In this chapter, we develop a novel statistical approach, the hierarchical structural component analysis for multinomial phenotypes (HisCoM-Categ). In a summary, the proposed HisCoM-Categ is an extension of the HisCoM for analyzing multinomial phenotypes. As the existing HisCoM, the proposed HisCoM-Categ considers the biomarker and pathway hierarchies while accounting for the correlations of all pathways. For identifying the association between pathways and phenotype, HisCoM-Categ uses the baseline category logit model for nominal phenotypes and the proportional odds model [33] for ordinal phenotypes. HisCoM-Categ is flexible enough to be used for different types of omics data. For example, we used our HisCoM-Categ methods on a real metabolomics dataset from the Korean Association Resource (KARE) to identify the association between metabolite pathways and type 2 diabetics (T2D). It is noted that T2D is a metabolic disease affected by multiple genetic factors [34], which is a major public health concern. Application to the KARE metabolite dataset demonstrates that HisCoM-Categ can well identify the T2D related pathways. Also, through the simulation studies, we evaluate the performance of HisCoM-Categ compared to other pathway analysis methods.

3.2. Methods

3.2.1 Model

Let $y_i^* \in \{1,2,\dots,(J+1) > 2\}$ be the multinomial phenotype for i^{th} $(i = 1,2,\dots,n)$ subject that can take one of $(J+1)$ levels. Let y_{ij} be the binary variable for $j = 1,2,\dots,(J+1)$, where

$$y_{ij} = \begin{cases} 1, \text{when subject } i \text{ is in category } j \\ 0, \text{otherwise} \end{cases}.$$

We define the $J \times 1$ response vector for the i^{th} subject $\boldsymbol{y}_i = \left(y_{i1}, y_{i2}, \dots, y_{iJ}\right)'$, in which we omitted $y_{i,J+1}$ since $\sum_{j=1}^{J+1} y_{ij} = 1$. Let the number of pathways is K and $k^{th}(k = 1,2,\dots,K)$ pathway contains M_k biomarkers. Let x_{ikm} be the m^{th} $(m = 1,2,\dots,M_k)$ biomarker value in the k^{th} pathway for i^{th} subject. Let $\boldsymbol{x}_i = \left(x_{i11}, x_{i12}, \dots, x_{i1M_1}, \dots, x_{iK1}, x_{iK2}, \dots, x_{iKM_K}\right)'$ be a $M \times 1$ vector of consisting all biomarkers for the i^{th} subject across K pathways, where $M = \sum_{k=1}^{K} M_k$. Next, let w_{km} be the weight associated with x_{ikm}, leading to the k^{th} pathway. Let $f_{ik} = \sum_{m=1}^{M_k} w_{km} x_{ikm}$ be the component score for i^{th}subject of the k^{th} pathway. Let $\boldsymbol{f}_i = (f_{i1}, \dots, f_{iK})'$ be a $K \times 1$ vector consisting of all pathways for the i^{th} subject. The probability density function of $\boldsymbol{y}_i$

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i) = \prod_{j=1}^{(J+1)} \left[\pi_j(\boldsymbol{x}_i)\right]^{y_{ij}},$$

where $\pi_j(\boldsymbol{x}_i) = E\left(y_{ij}|\boldsymbol{x}_i\right) = \Pr\left(y_{ij} = 1|\boldsymbol{x}_i\right)$ is the probability of the j^{th} response category i^{th} for subject. Let, the $J \times 1$ mean vector of $\boldsymbol{y}_i$ is $\boldsymbol{\pi}_i = E(\boldsymbol{y}_i|\boldsymbol{x}_i) = \left(\pi_{i1}, \dots, \pi_{iJ}\right)'$. The covariance matrix of multinomial trial is

$$\boldsymbol{\Sigma}_i = Var(\boldsymbol{y}_i) = \boldsymbol{D}_{\boldsymbol{\pi}_i} - \boldsymbol{\pi}_i \boldsymbol{\pi}_i',$$

where $\boldsymbol{D}_{\boldsymbol{\pi}_i}$ is the diagonal matrix of $\boldsymbol{\pi}_i$.

Let $g_j(\cdot)$ and η_{ij} be the link function and linear predictor, respectively for subject i at category j. Then the HisCoM−Categ can be defined as

$$\boldsymbol{\eta}_i = \boldsymbol{g}(\boldsymbol{\pi}_i) = \boldsymbol{F}_i \boldsymbol{\beta} = \boldsymbol{X}_i \boldsymbol{W} \boldsymbol{\beta},$$

where $\boldsymbol{\eta}_i = \left(\eta_{i1}, \eta_{i2}, \ldots, \eta_{iJ}\right)'$ is a $J \times 1$ vector of linear predictors, $\boldsymbol{g}(\boldsymbol{\pi}_i) = \left(g_1(\boldsymbol{\pi}_i), g_2(\boldsymbol{\pi}_i), \ldots, g_J(\boldsymbol{\pi}_i)\right)'$ is a $J \times 1$ vector of link functions, $\boldsymbol{W}$ represents a matrix of weight coefficients linking biomarkers to pathways, $\boldsymbol{\beta}$ is a vector of coefficients linking pathways to phenotype. The vector of link functions is chosen such that it consists baseline−category logit functions for nominal phenotypes and cumulative logit link functions or adjacent−categories logit functions for ordinal phenotypes. The form of $\boldsymbol{F}_i$, $\boldsymbol{X}_i$, $\boldsymbol{W}$ and $\boldsymbol{\beta}$ depend on the link function.

For Baseline−category Logit model, the design matrix $\boldsymbol{X}_i$ is

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & \boldsymbol{x}_i' & 0 & \boldsymbol{0} & \cdots & 0 & \boldsymbol{0} \\ 0 & \boldsymbol{0} & 1 & \boldsymbol{x}_i' & \cdots & 0 & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \boldsymbol{0} & 0 & \boldsymbol{0} & \cdots & 1 & \boldsymbol{x}_i' \end{bmatrix};$$

$$\boldsymbol{\beta} = \left(\beta_{01}, \boldsymbol{\beta}_1^T, \ \ldots, \beta_{0J}, \boldsymbol{\beta}_J^T\right)^T, \text{ and}$$

$$\boldsymbol{\beta}_j = \left(\beta_{1j}, \beta_{2j}, \ \ldots, \beta_{Kj}\right)^T.$$

The weight Matrix $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{W}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{W}_J \end{bmatrix}$ is block diagonal matrix,

where $\boldsymbol{W}_j = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & w_{11j} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & w_{1M_1 j} & 0 & \cdots & 0 \\ 0 & 0 & w_{21j} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & W_{2M_2 j} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & w_{K1j} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & w_{KM_K j} \end{bmatrix}$ is $(M+1) \times (K+1)$

dimensional weight matrix.

For cumulative logit model, the design matrix is

$\boldsymbol{X}_i = \begin{bmatrix} & & \boldsymbol{x}_i' \\ \boldsymbol{I}_{J\times J} & \vdots & \boldsymbol{x}_i' \\ & & \boldsymbol{x}_i' \end{bmatrix}$ and $\boldsymbol{\beta} = (\beta_{01}, \ldots, \beta_{0J}, \beta_1, \beta_2, \ldots, \beta_K)'$

The weight Matrix is $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{I}_{J\times J} & \vdots & \boldsymbol{0}_{J\times K} \\ \cdots & \cdots & \cdots \\ \boldsymbol{0}_{M\times J} & \vdots & \boldsymbol{W_1}: \boldsymbol{M} \times \boldsymbol{K} \end{bmatrix}$ is $(J+M) \times (J+K)$

weight matrix with

$$\boldsymbol{W}_1 = \begin{bmatrix} w_{11} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ w_{1M_1} & 0 & \cdots & 0 \\ 0 & w_{21} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & W_{2M_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & w_{K1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{KM_K} \end{bmatrix}.$$

3.2.2 Parameter estimation

In order to estimate the parameters $\boldsymbol{W}$ and $\boldsymbol{\beta}$, we seek to maximize the following penalized log−likelihood equation

$$Q(\boldsymbol{W},\boldsymbol{\beta}) = l(\boldsymbol{W},\boldsymbol{\beta}) - \frac{1}{2}\lambda_m\|\boldsymbol{W}\|^2 - \frac{1}{2}\lambda_p\|\boldsymbol{\beta}\|^2, \quad (1)$$

with respect to $\boldsymbol{W}$ and $\boldsymbol{\beta}$, subject to $Tr(\boldsymbol{F}\boldsymbol{F}^T) = n\boldsymbol{I}$ [35], where λ_m and λ_p are tuning parameters for the ridge penalty [36] for biomarkers and pathways, respectively. These two penalties are included to control the correlation in both biomarkers and pathways. For a vector or a matrix $\mathbf{B}$, denote $\|\boldsymbol{B}\| = [Tr(\boldsymbol{B}\boldsymbol{B}^T)]^{1/2}$. Also, $l(\boldsymbol{W},\boldsymbol{\beta}) = \sum_{i=1}^{n}\log f(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{W},\boldsymbol{\beta})$. We employed the ALS algorithm to maximize the objective function, that repeats the following two steps until convergence.

Step 1: We update $\boldsymbol{W}$ for fixed $\boldsymbol{\beta}$. Let $\boldsymbol{w} = vec(\boldsymbol{W})$, and by removing all zeros and ones from $\boldsymbol{w}$ vector we constructed a vector $\boldsymbol{w}^*$. To estimate the $\boldsymbol{w}^*$, we solve the following score function

$$\begin{aligned}\frac{\partial Q(\boldsymbol{W},\boldsymbol{\beta})}{\partial \boldsymbol{w}^*} &= \sum_{i=1}^{n}\left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{w}_*}\right)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\pi}_i) - \lambda_m \boldsymbol{w}^* \\ &= \sum_{i=1}^{n}\left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\eta}_i}\boldsymbol{\Phi}_i\right)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\pi}_i) - \lambda_m \boldsymbol{w}^*,\end{aligned}$$

where $\boldsymbol{\Phi}_i$ is a $J \times (J+M)$ matrix constructed by removing the columns of $(\boldsymbol{X}_i \otimes \boldsymbol{\beta}')$ corresponding to the zeros and ones of $\boldsymbol{w}$. Then, using the iterative reweighted least square (IRLS) algorithm, $\boldsymbol{w}^*$ can be estimated by

$$\widehat{\boldsymbol{w}^*} = \left(\sum_{i=1}^{n}\left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\eta}_i}\boldsymbol{\Phi}_i\right)^T \boldsymbol{\Sigma}_i^{-1}\left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\eta}_i}\boldsymbol{\Phi}_i\right) + \lambda_m \boldsymbol{I}\right)^{-1}\left(\sum_{i=1}^{n}\left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\eta}_i}\boldsymbol{\Phi}_i\right)^T \boldsymbol{\Sigma}_i^{-1}\boldsymbol{z}_i\right),$$

where

$$\boldsymbol{z}_i = \boldsymbol{\eta}_i + \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\pi}_i}(\boldsymbol{y}_i - \boldsymbol{\pi}_i).$$

Step 2: We update $\boldsymbol{\beta}$ for fixed $\boldsymbol{W}$, and we solve the following score function

$$\begin{aligned}\frac{\partial Q(\boldsymbol{W}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_i^n \left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\beta}}\right)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\pi}_i) - \lambda_p \boldsymbol{\beta} \\ &= \sum_i^n \left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\eta}_i}\boldsymbol{F}_i\right)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\pi}_i) - \lambda_p \boldsymbol{\beta},\end{aligned}$$

where $\boldsymbol{F}_i = \boldsymbol{X}_i \boldsymbol{W}$. Then, using IRLS algorithm, $\boldsymbol{\beta}$ can be estimated by

$$\widehat{\boldsymbol{\beta}} = \left[\sum_{i=1}^n \left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\eta}_i}\boldsymbol{F}_i\right)^T \boldsymbol{\Sigma}_i^{-1}\left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\eta}_i}\boldsymbol{F}_i\right) + \lambda_p \boldsymbol{I}\right]^{-1}\left(\sum_{i=1}^n \left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\eta}_i}\boldsymbol{F}_i\right)^T \boldsymbol{\Sigma}_i^{-1}\boldsymbol{z}_i\right).$$

Finally, we determine the optimal tuning parameter values of λ_m and λ_p using $k-$fold cross-validation (CV). In CV, we compare the log-likelihood values of a two-dimensional grid of candidate values of λ_m and λ_p.

3.2.3 Penalized HisCoM-Categ estimation

For the penalized HisCoM-Categ, penalized log-likelihood equation in (1) can be written as

$$Q(\boldsymbol{\beta}, \boldsymbol{W}) = l(\boldsymbol{\beta}, \boldsymbol{W}) - n\sum_{k=1}^K \sum_{m=1}^{M_k} p_{\lambda_m}(|w_{km}|) - n\sum_{k=1}^K p_{\lambda_p}(|\beta_k|)$$

where λ_m and λ_p are tuning parameters for biomarkers and pathways, respectively. $p_{\lambda_m}(\cdot)$ and $p_{\lambda_p}(\cdot)$ are the penalty functions associated with biomarkers and pathways. The penalized maximizing likelihood

estimators (PMLEs) are obtained by solving the following estimating equation

$$\frac{\partial Q(\boldsymbol{W},\boldsymbol{\beta})}{\partial \boldsymbol{\Gamma}} = \frac{\partial l(\boldsymbol{W},\boldsymbol{\beta})}{\partial \boldsymbol{\Gamma}} - n\sum_{k=1}^{K}\left[p_{\lambda_m}(|w_{km}|)\right]' - n\sum_{k=1}^{K}\sum_{m}^{M_k}\left[p_{\lambda_p}(|\beta_k|)\right]'$$

where $\boldsymbol{\Gamma} = (\boldsymbol{W},\boldsymbol{\beta})$. By local quadratic approximation (LQA) algorithm [37]

$$\left[p_{\lambda_m}(|w_{km}|)\right]' = p'_{\lambda_m}(|w_{km}|).sgn(w_{km}) \approx \left\{\frac{p'_{\lambda_m}(|w_{km}|)}{|w_{km}|}\right\} w_{km},$$

where p'_{λ_m} is the derivative of penalty function and $sgn(\cdot)$ is the sign function. Similarly, by the LQA algorithm

$$\left[p_{\lambda_p}(|\beta_k|)\right]' = p'_{\lambda_p}(|\beta_k|).sgn(\beta_k) \approx \left\{\frac{p'_{\lambda_p}(|\beta_k|)}{|\beta_k|}\right\} \beta_k.$$

In this study we use three well-known penalty functions. They are the least absolute shrinkage and selection operator (LASSO) [38], the smoothly clipped absolute deviation penalty (SCAD) [37] and the minimum concave penalty (MCP) [39]. The function of LASSO penalty is

$$p_\lambda(|\theta|) = \lambda|\theta|.$$

The function of SCAD penalty is

$$\begin{aligned} p_\lambda(|\theta|) = {} & \lambda|\theta| I(0 \le |\theta| < \lambda) \\ & + \left(\frac{a\lambda(|\theta| - \lambda) - \frac{|\theta|^2 - \lambda^2}{2}}{(a-1)} + \lambda^2 \right) I(\lambda \le |\theta| < a\lambda) \\ & + \left(\frac{(a-1)\lambda^2}{2} + \lambda^2 \right) I(|\theta| > a\lambda), \end{aligned}$$

and the derivative of SCAD penalty is

$$p'_\lambda(|\theta|) = \begin{cases} \lambda & \text{if } |\theta| \le \lambda \\ \dfrac{a\lambda - \theta}{a-1} & \text{if } \lambda < |\theta| \le a\lambda \\ 0 & \text{if } |\theta| > a\lambda \end{cases}$$

for some $a > 2$.

The function of MCP penalty is

$$p_\lambda(|\theta|) = \left(\lambda\theta + \frac{\theta^2}{2a}\right) I(0 \le |\theta| < a\lambda) + \left(\frac{\lambda^2 a}{2}\right) I(|\theta| > a\lambda),$$

and the first derivative of MCP penalty is

$$p'_\lambda(|\theta|) = \begin{cases} \left(\lambda - \dfrac{\theta}{a}\right) \operatorname{sgn}(\theta) & \text{if } |\theta| \le a\lambda \\ 0, & \text{if } |\theta| > a\lambda \end{cases}$$

for some $a > 1$.

3.3. Materials

In this study, we use metabolite data from the Korean Association Resource (KARE) cohort to identify the association between pathways and T2D. This cohort is a community-based cohort established through the Korean Genomic Epidemiologic Study (KoGES) project in the Ansung and Ansan areas of Kyounggi province, South Korea [40]. In 2001-2202, 10,300 individuals aged 40 to 69 were recruited as the baseline, and following surveys were conducted every two years. The dataset was obtained from the from 6^{th}, 7^{th} and 8^{th} follow-ups of the KoGES study and called phase 6, phase 7 and phase 8. The serum metabolites of the subjects were measured using liquid chromatography-mass spectrometry (LC-MS). Among them, 64 metabolites were quantitively analyzed.

Systematical error and batch-effect correction were removed using the systematic error removal using random forest (SERRF) method which may have risen due to instrument and injection time [41]. Then, these 64 metabolites were first mapped to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database. Among 64 metabolites, 52 unique metabolites were mapped to 65 pathways.

Individuals were classified into three different groups such as the normal group, the pre-diabetics group (Pre T2D), and the T2D group. Table 1 displays the total number of samples in each group and follow-up. In phase 6, 691 samples were recruited for metabolomics data collection, with 348 samples being normal, 272 samples having pre-T2D, and the remaining 71 samples having T2D, as shown in Table 3.1. According to Table 3.1, Among the 689 samples in phase 7, there are 330, 226 and 133 samples in the normal, preT2D, and T2D groups, respectively. As shown in Table 3.1, total 666 samples were recruited for metabolomics data collection in phase 8, including 330, 226 and 133 samples in the normal, preT2D, and T2D groups, respectively. In total 664 samples were present in all three phases.

Table 3.1. Frequency of the total number of participants.

T2D category	Phase 6	Phase 7	Phase 8
Normal	348	330	316
Pre T2D	272	226	158
T2D	71	133	192
Total	691	689	666

3.4. Simulation study

3.4.1 Simulation model

To assess the performance of HisCoM−Categ a simulation study is conducted. To evaluate the performance of HisCoM−Categ with other existing methods, we generate the ordinal phenotype. To generate the ordinal phenotype y_i^*, consider the following cumulative logit model

$$\Pr(y_i^* \le j|\boldsymbol{x}_i) = \Pr(\mathrm{U} \le \beta_{0j}|\boldsymbol{x}_i) = G\left(\beta_{0j} - \sum_{k=1}^{K}\left[\sum_{m=1}^{M_k} x_{ikm}w_{km}\right]\beta_k\right),$$

where $i = 1, \dots, n$, $j = 1, \dots, J$ and G denotes the distribution function of the standard logistic distribution. The following latent regression model is considered for generating the y_i^*

$$U_i = \sum_{k=1}^{K}\left[\sum_{m=1}^{M_k} x_{ikm}w_{km}\right]\beta_k + \epsilon_i,$$

where $\epsilon_i \sim G$ and $E[\epsilon_i] = 0$.

Now, we categorize U_i using the corresponding category−specific intercept according to the following threshold to generate y_i^*,

$$y_i^* = j \Leftrightarrow \beta_{0,j-1} < U_i \le \beta_{0j},$$

where $-\infty = \beta_{00} < \beta_{01} < \cdots < \beta_{0J} < \beta_{0,(J+1)} = \infty$.

In simulation study, we use same biomarkers and pathways from real KARE phase 6 metabolite dataset and generate the ordinal

phenotype. Thus, similar to the real dataset, we set the number of categories $(J+1)$ is 3, total number of pathways is $K = 65$. Here we assume that first five pathways are causal pathway and remaining 60 pathways are non-causal pathway. Let $\boldsymbol{\beta}_0 = (\beta_{01}, \beta_{02})' = (-0.3, 0.8)'$. For the causal pathways, we considered two different parameter settings: two biomarker level effects ($w = 0.2$ and 0.3), four pathway-level effects ($\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0.3,\ 0.4,\ 0.5, 0.6$). To evaluate the type I error, we use $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$. Again, for the non-causal pathways we use $\beta_6 = \beta_7 = \cdots = \beta_{65} = 0$. We generate 100 datasets with the sample size for each dataset being the same as the real KARE phase 6 dataset. To calculate the performance, we permute each simulated response 1000 times to calculate the type I error and power. The proportion of cases where at least one true null hypothesis is wrongly rejected is used to calculate the type I error. The proportion of the cases in which all false null hypotheses are correctly rejected is used to compute the statistical power.

3.4.2 Simulation results

In order to demonstrate the statistical performance of the proposed HisCoM-Categ we perform the simulation study. For the purpose of the performance comparison, we compared the type I error and power for HisCoM-Categ with other existing pathway-based methods. We consider GSEA, aSPU and HisCoM as existing pathway-based methods. To use the HisCoM, we use two cases for simulated phenotype (0, 1+2) and (0+1, 2); because HisCoM is for binary phenotype. After generating the phenotype for each simulation,

we obtain the optimal tuning parameter set (λ_m, λ_p) using the 3 folds cross-validation. Then, we evaluate the type I error and power.

Results of the empirical type I error shows in Figure 3.1. Overall, type I errors are shown to well-controlled in various method except GSEA. Especially in HisCoM-Categ method, type I error is well control. Type I error for HisCoM-Categ was well controlled compare to the others methods.

Results of empirical power presents in Figure 3.2, where the x-axis shows the effect sizes of pathways and the y-axis shows the power. The left panel of Figure 3.1 represents the power for biomarkers effect $w = 0.2$ and the right panel is for biomarkers effect $w = 0.3$. HisCoM-Categ and HisCoM (0,1+2) showed similar power for small and large effect sizes for both pathways and biomarkers. HisCoM-Categ outperformed for moderate effect sizes compared to the all other methods. Again, for large effect size HisCoM-Categ, HisCoM (0,1+2) and aSPU provides similar effects where GSEA had the lowest power. Finally, regardless of the effect sizes, HisCoM-Categ and HisCoM (0,1+2) outperformed the conventional methods.

Figure 3.1. Results of the empirical type I error

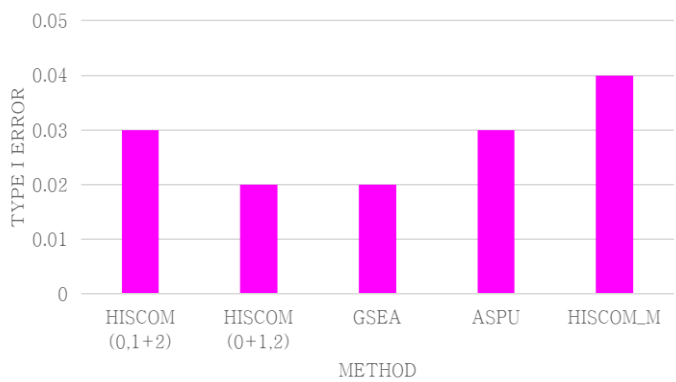
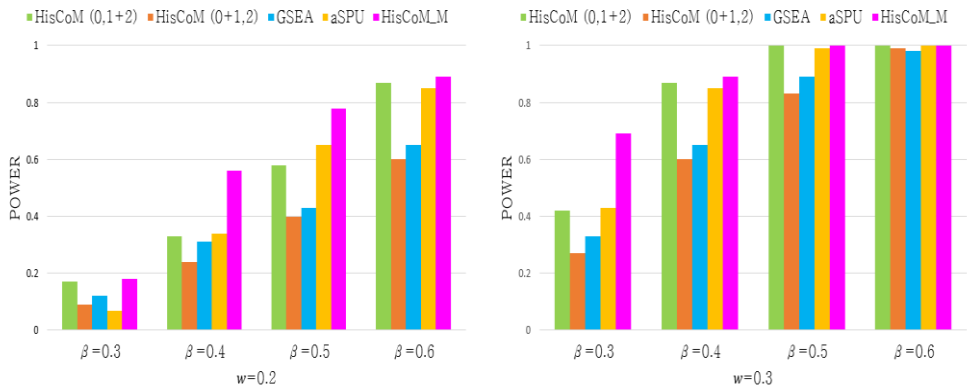


Figure 3.2. Results of the empirical power



3.5. Real data analysis results

3.5.1 Real data analysis results of HisCoM−Categ

In this section, we used the KARE phase 6 dataset to examine the association between pathways using HisCoM−Categ. To identify the pathways that associated with T2D, we performed HisCoM−Categ and aSPU, where age, gender and BMI were included as adjusting covariates. For the KARE phase 6 metabolite dataset, consider the phenotype T2D is an ordinal variable. Let y_i^* be the level of T2D (1 = Normal, 2 = Pre T2D and 3 = T2D) for i^{th} $(i = 1,2,\dots,n)$ subject. Since the phenotype is an ordinal variable, thus we apply the HisCoM−Categ method for the following cumulative logit link,

$$\text{logit}[\Pr(y_i^* \le j)] = \beta_{0j} + \sum_{k=1}^{65} \beta_k \left(\sum_{m=1}^{M_k} x_{ikm} w_{km} \right) + \beta_{66} * \text{age} + \beta_{67} * \text{gender} + \beta_{68} * \text{BMI},$$

for $j = 1, 2$, which represents a proportional odds model.

We use 5−fold cross−validation to choose the best optimal tuning parameters for biomarkers (λ_m) and pathways (λ_p). As a comparison, we also performe an aSPU test and GSEA to find the association between pathways and ordinal phenotypes. Note that, both aSPU and GSEA consider a single pathway at a time. To compare, we also perform HisCoM for a binary phenotype considering normal and T2D for two different cases: (i) normal (0) + pre−T2D (1) vs. T2D (2), and (ii) normal (0) vs. pre−T2D (1) + T2D(2). We use 10000 permutations for calculating the p−values of pathways for all comparative methods. For multiple comparison, FDR adjusted p−values (q−values) were calculated [42]. The q−values of all comparative methods are shown in the Table 3.2. Venn diagrams in Figure 3.3 showes the number of commonly significant pathways from all of the comparative methods. There are 53, 55, 23, 4 and 4 pathways are selected by HisCoM−Categ, aSPU, HisCoM (0, 1+2), HisCoM (0+1, 2) and GSEA methods. Among the selected pathways, 23 pathways are commonly selected by HisCoM−Categ, aSPU and HisCoM (0, 1+2) methods. Table 3.3 summarizes the list 23 commonly significant pathways by HisCoM−Categ, aSPU and HisCoM (0, 1+2) methods. All of these pathways except “pathways of neurodegeneration − multiple diseases” and “propanoate metabolism” have already been identified by HisCoM [43].

Figure 3.3. The number of significantly identified pathways by HisCOM−Categ and other comparative methods

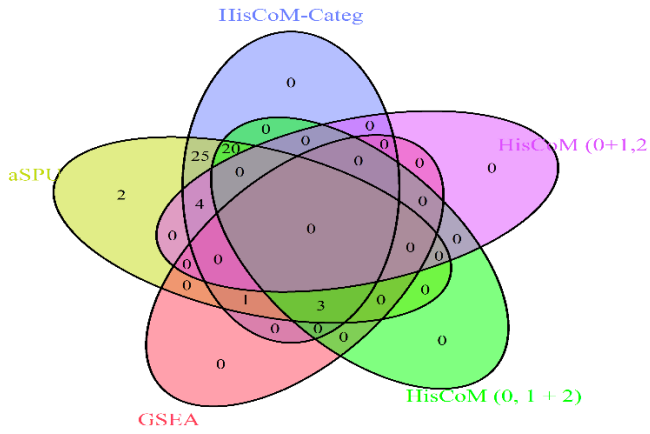


Table 3.2. Detailed results of HisCoM-Categ and other methods

Pathway Names	HisCoM-Categ	q-value			
		aSPU	GSEA	HisCoM 0,1+2	HisCoM 0+1,2
Primary bile acid biosynthesis	0.0007	0.0004	0.3274	0.2133	0.4177
Arginine biosynthesis	0.0002	0.0001	0.6582	0.0626	0.4440
Purine metabolism	0.0002	0.0001	0.7576	0.3362	0.4440
Caffeine metabolism	0.0002	0.0001	0.4559	0.1479	0.4440
Pyrimidine metabolism	0.0115	0.0260	0.3674	0.3367	0.4177
Alanine, aspartate and glutamate metabolism	0.0002	0.0001	0.4192	0.0170	0.4440
Glycine, serine and threonine metabolism	0.0002	0.0001	0.4422	0.3553	0.4440
Cysteine and methionine metabolism	0.0002	0.0001	0.7929	0.3244	0.4440

Valine, leucine and isoleucine degradation	0.0066	0.0038	0.3963	0.3490	0.4440
Valine, leucine and isoleucine biosynthesis	0.0074	0.0096	0.3741	0.8168	0.6829
Lysine degradation	0.0022	0.0016	0.4422	0.1861	0.444
Arginine and proline metabolism	0.0002	0.0001	0.7756	0.1805	0.4177
Histidine metabolism	0.0002	0.0001	0.3741	0.2902	0.444
Tyrosine metabolism	0.0002	0.0001	0.4330	0.0092	0.6365
Phenylalanine metabolism	0.0002	0.0001	0.4192	0.0897	0.6186
Tryptophan metabolism	0.0751	0.0364	0.6398	0.1585	0.7241
Phenylalanine, tyrosine and tryptophan biosynthesis	0.0002	0.0001	0.4422	0.0129	0.4440
beta-Alanine metabolism	0.0006	0.0003	0.7576	0.6604	0.4440
Taurine and hypotaurine metabolism	0.0002	0.0001	0.1114	0.0331	0.4177
Glutathione metabolism	0.0002	0.0001	0.8318	0.2021	0.4177
Glycerophospholipid metabolism	0.4990	0.6713	0.8811	0.7497	0.4440

Pyruvate metabolism	0.0002	0.0001	0.3741	0.1386	0.4440
Glyoxylate and dicarboxylate metabolism	0.0002	0.0001	0.4128	0.0331	0.4440
Propanoate metabolism	0.3512	0.3912	0.9313	0.1861	0.4440
Butanoate metabolism	0.0002	0.0001	0.4422	0.0756	0.4440
Thiamine metabolism	0.0002	0.0001	0.8106	0.0017	0.4440
Nicotinate and nicotinamide metabolism	0.0002	0.0004	0.8811	0.1110	0.4440
Pantothenate and CoA biosynthesis	0.0002	0.0001	0.3741	0.6409	0.4177
Biotin metabolism	0.0543	0.1092	0.7962	0.3667	0.444
Porphyrin metabolism	0.0002	0.0001	0.7929	0.0331	0.6614
Nitrogen metabolism	0.0002	0.0001	0.6398	0.0129	0.444
Sulfur metabolism	0.1583	0.0941	0.3741	0.3362	0.4177
Aminoacyl−tRNA biosynthesis	0.0002	0.0001	0.3741	0.0121	0.4440
Metabolic pathways	0.0007	0.0001	0.7576	0.3255	0.4177
Carbon metabolism	0.0002	0.0001	0.4720	0.0263	0.4177
2−Oxocarboxylic acid metabolism	0.0002	0.0001	0.0407	0.0976	0.4440
Biosynthesis of amino acids	0.0002	0.0001	0.4192	0.0593	0.7103
Biosynthesis of cofactors	0.0002	0.0001	0.6582	0.2152	0.4177
ABC transporters	0.0002	0.0001	0.5307	0.3658	0.4177

cAMP signaling pathway	0.9324	0.9397	0.4422	0.7516	0.9903
Neuroactive ligand–receptor interaction	0.0002	0.0001	0.6555	0.2820	0.4177
Sulfur relay system	0.0002	0.0001	0.4192	0.0129	0.4440
mTOR signaling pathway	0.0032	0.0042	0.7903	0.0259	0.9903
Ferroptosis	0.0002	0.0001	0.2802	0.0129	0.4440
Gap junction	0.0002	0.0001	0.1114	0.1479	0.9903
Thermogenesis	0.2947	0.2842	0.9420	0.4638	0.9903
Synaptic vesicle cycle	0.0002	0.0001	0.7576	0.1585	0.8743
Retrograde endocannabinoid signaling	0.0002	0.0001	0.3865	0.0129	0.9903
Glutamatergic synapse	0.0002	0.0001	0.6398	0.0129	0.4440
Cholinergic synapse	0.4990	0.6713	0.8811	0.7497	0.4440
GABAergic synapse	0.0002	0.0001	0.6398	0.0222	0.6703
Taste transduction	0.0002	0.0001	0.3741	0.2315	0.9903
Proximal tubule bicarbonate reclamation	0.0002	0.0001	0.6398	0.0129	0.4440
Salivary secretion	0.7129	0.7211	0.6362	0.7639	0.9903
Protein digestion and absorption	0.0002	0.0001	0.3741	0.0129	0.4440
Bile secretion	0.1351	0.2391	0.8824	0.6409	0.4177
Vitamin digestion and absorption	0.2244	0.2400	0.8811	0.6046	0.4177

Mineral absorption	0.0002	0.0001	0.7576	0.1106	0.4440
Amyotrophic lateral sclerosis	0.0002	0.0001	0.4330	0.0092	0.9903
Pathways of neurodegeneration – multiple diseases	0.0002	0.0001	0.4422	0.0099	0.9903
Cocaine addiction	0.0002	0.0001	0.0402	0.0017	0.9903
Amphetamine addiction	0.0002	0.0001	0.0402	0.0017	0.9903
Nicotine addiction	0.0002	0.0001	0.4192	0.1106	0.9903
Alcoholism	0.0002	0.0001	0.0402	0.0017	0.9903
African trypanosomiasis	0.0751	0.0309	0.6398	0.1585	0.7241

Table 3.3. List of the 23 commonly significant pathways associated with T2D in all phases by HisCoM–Categ, aSPU and HisCoM (0, 1+2)

Alanine, aspartate and glutamate metabolism	Nitrogen metabolism
Alcoholism	Pathways of neurodegeneration – multiple diseases
Aminoacyl–tRNA biosynthesis	Phenylalanine, tyrosine and tryptophan biosynthesis
Amphetamine addiction	Porphyrin and chlorophyll metabolism

Amyotrophic lateral sclerosis	Protein digestion and absorption
Carbon metabolism	Proximal tubule bicarbonate reclamation
Cocaine addiction	Retrograde endocannabinoid signaling
Ferroptosis	Sulfur relay system
GABAergic synapse	Taurine and hypotaurine metabolism
Glutamatergic synapse	Thiamine metabolism
Glyoxylate and dicarboxylate metabolism	Tyrosine metabolism
mTOR signaling pathway	

3.5.2 Real data analysis results of penalized HisCoM-Categ

In real data analysis using penalized HisCoM-Categ, we use the same metabolomics dataset that we used in HisCoM-categ. Selected pathways using three different penalties are shown in Table 3.3. Among the total 65 pathways SCAD penalty selects 6 pathways, MCP penalty selects 4 pathways and LASSO selects 3 pathways. Commonly selected pathways using three different penalties are shown in Figure 3.4. Venn-diagram in Figure 3.4 shows that 2 pathways such as 'ferroptosis' and 'metabolic pathways'. Additionally, the LASSO penalty selects 'nicotinate and nicotinamide metabolism' pathway which is also selected by the SCAD penalty. Again, SCAD and LASSO penalty commonly selects 4 pathways.

Figure 3.4. Commonly selected pathways using penalized HisCoM-Categ

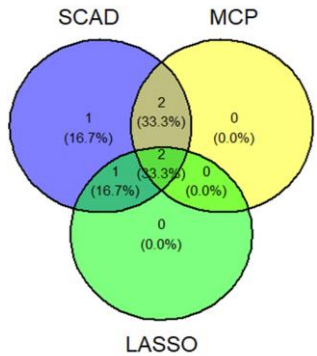


Table 3.4. Results of penalized HisCoM-Categ

Pathway Name	Pathway coefficient (β)		
	SCAD	MCP	LASSO
Cyanoamino acid metabolism	0.2614	-	-
Glutathione metabolism	-1.3157	-2.7944	-
Thiamine metabolism	1.5655	3.0662	-
Nicotinate and nicotinamide metabolism	0.1892	-	0.0609
Ferroptosis	-0.4190	-0.5276	-0.0001
Metabolic Pathways	0.4306	0.4388	0.8013

3.6. Discussion

In summary, Hierarchical Structural Component Models of Pathway Analysis for Multinomial Phenotypes (HisCoM-Categ) is propose for identifying pathways that have been associated with multinomial a phenotype. HisCoM-Categ considers the hierarchies among pathways and biomarkers. HisCoM-Categ evaluates the relationship between pathways and a multinomial phenotype in a

single model. HisCoM-Categ also enables us to control the correlations among pathways and among biomarkers. HisCoM-Categ is flexible enough to be used for both nominal and ordinal phenotypes. Using the simulation data, we also show the comparison of propose HisCoM-Categ with other comparative methods. Based on the simulation results, performance of HisCoM-Categ is higher than all other methods and control type I error well. We also apply three different penalties in HisCoM-Categ method. The real metabolite data analysis shows that HisCoM-Categ is able to identify the well-known pathways that have been associated with multinomial phenotypes. Therefore, we hope that HisCoM-Categ may be able to help the researchers identify the pathways that are associated with multinomial phenotypes. We also think that HisCoM-Categ is robust for use with any other types of omics data, such as microbiome data.

Chapter 4. Pathway-based Approach using Hierarchical Structural Component Models to Analyze longitudinal Multinomial Phenotypes

4.1. Introduction

In this chapter, we propose a novel statistical approach, the Hierarchical Structural Component Models to Analysis longitudinal Multinomial phenotypes using Generalized Estimating Approach (HisCoM-Rcateg). In a summary, the proposed HisCoM-Rcateg is an extension of the HisCoM-Categ method for analyzing longitudinal multinomial phenotypes. As an extension of the existing HisCoM-Categ, the proposed HisCoM-Rcateg considers the biomarker and pathway hierarchies while accounting for the correlations of all pathways by using the ridge penalty. Like HisCoM-Categ, for identifying the association between pathways and phenotype, HisCoM-Rcateg uses the baseline category logit model for nominal phenotypes and the proportional odds model [33] for ordinal phenotypes. HisCoM-Rcateg is also flexible enough to be used for different types of omics data. For example, we used our HisCoM-Rcateg methods on a real metabolomic dataset from the Korean Association Resource (KARE) to identify the association between metabolite pathways and type 2 diabetics (T2D). Application to the KARE metabolite dataset demonstrates that HisCoM-Rcateg can well identify the T2D related pathways.

4.2. Methods

4.2.1 Model

Let $y_{it}^{*} \in \{1,2,\dots,(J+1)>2\}$ be the multinomial phenotype for i^{th} $(i=1,2,\dots,n)$ subject at $t^{th}(t=1,2,\dots,T)$ time point that can take one of $(J+1)$ levels. Let y_{itj}^{*} be the binary variable for $j=1,2,\dots,(J+1)$, where $y_{itj}=1$ when i^{th} subject is in j^{th} category at t^{th} time and $y_{itj}^{*}=0$ otherwise. We define the $J\times 1$ response vector for the i^{th} subject at t^{th} time $\boldsymbol{y}_{it}=\left(y_{it1}^{*},\dots,y_{itJ}^{*}\right)'$, in which we omitted $y_{it,J+1}^{*}$ since $\sum_{j=1}^{J+1}y_{itj}^{*}=1$. Then, the response vector for the i^{th} subject $\boldsymbol{y}_i=(\boldsymbol{y}_{i1}',\ \boldsymbol{y}_{i2}',\ \dots,\ \boldsymbol{y}_{iT}')'$: is $TJ\times 1$ vector. Let the total number of pathways are K and $k^{th}(k=1,2,\dots,K)$ pathwa contains M_k biomarkers. Let x_{itkm} be the m^{th} $(m=1,2,\dots,M_k)$ biomarker value in the k^{th} pathway for i^{th} subject at time t. Let $\boldsymbol{x}_{it}=\left(x_{it11},\ x_{it12},\dots,x_{it1M_1},\dots,x_{itK1},\ x_{itK2},\dots,x_{itKM_K}\right)'$ is a $M\times 1$ vector of consisting all biomarkers for the i^{th} subject across K pathways, where $M=\sum_{k=1}^{K}M_k$. Next, let w_{km} be the weight associated with x_{itkm}, leading to the k^{th} pathway. Let $f_{itk}=\sum_{m=1}^{M_k}w_{km}x_{itkm}$ be the component score for i^{th}subject of the k^{th} pathway at time t. Let $\boldsymbol{f}_{it}=(f_{it1},\dots,f_{itK})'$ be a $K\times 1$ vector consisting of all pathways for the i^{th} subject at time t. The marginal density function of $\boldsymbol{y}_{it}$ is consider to the multinomial distribution, that is

$$f(\boldsymbol{y}_{it}|\boldsymbol{x}_{it})=\prod_{j=1}^{(J+1)}\left[\pi_{tij}\right]^{y_{itj}},$$

where $\pi_{itj}=E\left(y_{itj}\middle|\boldsymbol{x}_{it}\right)=\Pr\left(y_{itj}=1\middle|\boldsymbol{x}_i\right)$ be the probability of the j^{th} phenotype category for i^{th} subject at t^{th} time. Let $\boldsymbol{\pi}_{it}=$

$E(\boldsymbol{y}_{it}|\boldsymbol{x}_{it}) = \left(\pi_{it1}, \ldots, \pi_{itJ}\right)'$ is the $J \times 1$ mean vector of $\boldsymbol{y}_{it}$, and the $TJ \times 1$ mean vector of $\boldsymbol{y}_i$ is $\boldsymbol{\pi}_i = E(\boldsymbol{y}_i|\boldsymbol{X}_i) = (\boldsymbol{\pi}'_{i1}, \ldots, \boldsymbol{\pi}'_{iT})'$.

Let $g_j(\cdot)$ and η_{itj} be the link function and linear predictor, respectively for subject i in category j at time t. Then the HisCoM-RCateg can be defined as

$$\boldsymbol{\eta}_{it} = \boldsymbol{g}_j(\boldsymbol{\pi}_{it}) = \boldsymbol{F}_{it}\boldsymbol{\beta}_t = \boldsymbol{X}_i\boldsymbol{W}\boldsymbol{\beta}_t,$$

where $\boldsymbol{\eta}_{it} = \left(\eta_{it1}, \eta_{it2}, \ldots, \eta_{itJ}\right)'$ is a $J \times 1$ vector of linear predictors, $\boldsymbol{g}(\boldsymbol{\pi}_{it}) = \left(g_1(\boldsymbol{\pi}_{it}), g_2(\boldsymbol{\pi}_{it}), \ldots, g_J(\boldsymbol{\pi}_{it})\right)'$ is a $J \times 1$ vector of link functions, $\boldsymbol{W}$ represents a matrix of weight coefficients that make the link between biomarkers and pathways, and $\boldsymbol{\beta}$ denotes a vector of coefficients of pathways to phenotype. The choice of the vector of link functions $\boldsymbol{g}$ is the baseline-category logit function for nominal response and cumulative link function for ordinal response. The form of $\boldsymbol{F}_{it}$, $\boldsymbol{X}_{it}$, $\boldsymbol{W}$ and $\boldsymbol{\beta}_t$ depend on the link function.

4.2.2 Parameter estimation

To estimate the parameters $\boldsymbol{W}$ and $\boldsymbol{\beta}$, we maximize the following penalized generalized estimating equation with respect to $\boldsymbol{W}$ and $\boldsymbol{\beta}$

$$\boldsymbol{U}(\boldsymbol{W}, \boldsymbol{\beta}) = \boldsymbol{S}(\boldsymbol{W}, \boldsymbol{\beta}) - p'_{\lambda_m}(\boldsymbol{w}) \; - p'_{\lambda_b}(\boldsymbol{b}),$$

subject to $Tr(\boldsymbol{F}\boldsymbol{F}^T) = n\boldsymbol{I}$ [35], where $\boldsymbol{S}(\boldsymbol{W}, \boldsymbol{\beta})$ is the generalized estimating equation for parameters, λ_m and λ_p are tuning parameters for the ridge penalty [36] for biomarkers and pathways, respectively and $p'_\lambda(\cdot)$ is the first derivation of ridge penalty with $p_\lambda(\boldsymbol{B}) = \frac{1}{2}\|\boldsymbol{B}\|^2$.

For a vector or a matrix $\boldsymbol{B}$, denote $\|\boldsymbol{B}\| = [Tr(\boldsymbol{B}\boldsymbol{B}^T)]^{1/2}$. We employ these two ridge penalties to regulate the correlation in both the biomarkers and the pathways.

Now the generalized estimating equation (GEE) for the parameters is

$$\boldsymbol{S}(\boldsymbol{w},\boldsymbol{b}) = \sum_{i}^{n} \boldsymbol{D}_i'(\boldsymbol{w},\boldsymbol{b})\Sigma_i^{-1}(\boldsymbol{w},\boldsymbol{b})\big(\boldsymbol{y}_i - \boldsymbol{\pi}_i(\boldsymbol{w},\boldsymbol{b})\big)$$

where $\boldsymbol{D}_i(\boldsymbol{w},\boldsymbol{b}) = \frac{\partial \boldsymbol{\pi}_i(\boldsymbol{w},\boldsymbol{b})}{\partial(\boldsymbol{w},\boldsymbol{b})}$ and $\boldsymbol{\Sigma}_i(\boldsymbol{w},\boldsymbol{b})$ is a $TJ \times TJ$ "working" covariance matrix of $\boldsymbol{y}_i$. Let $\boldsymbol{R}_i(\boldsymbol{\alpha})$ be the working correlation matrix for i^{th} the subject. Then, the working covariance matrix $\boldsymbol{\Sigma}_i(\boldsymbol{w},\boldsymbol{b})$ can be written as

$$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_i(\boldsymbol{w},\boldsymbol{b}) = A_i^{\frac{1}{2}}(\boldsymbol{w},\boldsymbol{b})\boldsymbol{R}_i(\boldsymbol{\alpha})A_i^{\frac{1}{2}}(\boldsymbol{w},\boldsymbol{b})$$

where $\boldsymbol{A}_i$ is the matrix of marginal variances, $\boldsymbol{A}_{it}$, detailed in section 2.3.

To maximize the penalized generalized estimating equation function $\boldsymbol{U}(\boldsymbol{w},\boldsymbol{b})$, we used the alternating iterative algorithm that repeats the following steps until convergence.

To maximize the objective function, we used the alternating least squares (ALS) algorithm, which iterates the following two steps until convergence.

Step 1: We update $\boldsymbol{w}$ for fixed $\boldsymbol{b}$. Let $\boldsymbol{w} = vec(\boldsymbol{W})$, and $\boldsymbol{w}_*$ is the vector formed by eliminating all zero and one elements of $\boldsymbol{w}$. To estimate the $\boldsymbol{w}_*$, we solve the following score function

$$\boldsymbol{U}(\boldsymbol{W},\boldsymbol{\beta}) = \boldsymbol{S}(\boldsymbol{W},\boldsymbol{\beta}) - p'_{\lambda_m}(\boldsymbol{w})$$

$$= \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{w}_*}\right)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\pi}_i) - \lambda_m \boldsymbol{w}$$

$$= \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\eta}_i}\boldsymbol{\Phi}_i\right)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\pi}_i) - \lambda_m \boldsymbol{w}$$

where $\boldsymbol{\Phi}_i$ is a $J \times (J + M)$ matrix constructed by removing the columns of $(\boldsymbol{X}_i \otimes \boldsymbol{\beta}')$ corresponding to the zero and one elements of $\boldsymbol{w}$. Then, using the IWRLS algorithm, $\boldsymbol{w}_*$ can be estimated by

$$\widehat{\boldsymbol{w}}_* = \left(\sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\eta}_i}\boldsymbol{\Phi}_i\right)^T \boldsymbol{\Sigma}_i^{-1} \left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\eta}_i}\boldsymbol{\Phi}_i\right) + \lambda_m \boldsymbol{I}\right)^{-1} \left(\sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\eta}_i}\boldsymbol{\Phi}_i\right)^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{z}_i\right),$$

where

$$\boldsymbol{z}_i = \boldsymbol{\eta}_i + \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\pi}_i}(\boldsymbol{y}_i - \boldsymbol{\pi}_i).$$

Step 2: We update $\boldsymbol{b}$ for fixed $\boldsymbol{w}$, and we solve the following score function

$$\boldsymbol{U}(\boldsymbol{W}, \boldsymbol{\beta}) = \boldsymbol{S}(\boldsymbol{W}, \boldsymbol{\beta}) - p'_{\lambda_p}(\boldsymbol{b})$$

$$= \sum_{i}^{n} \left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\beta}}\right)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\pi}_i) - \lambda_p \boldsymbol{b}$$

$$= \sum_{i}^{n} \left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\eta}_i}\boldsymbol{F}_i\right)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\pi}_i) - \lambda_p \boldsymbol{\beta},$$

where $\boldsymbol{F}_i = \boldsymbol{X}_i \boldsymbol{W}$. Then, using IWRLS algorithm, $\boldsymbol{b}$ can be estimated by

$$\widehat{\boldsymbol{b}} = \left[\sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\eta}_i}\boldsymbol{F}_i\right)^T \boldsymbol{\Sigma}_i^{-1} \left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\eta}_i}\boldsymbol{F}_i\right) + \lambda_p \boldsymbol{I}\right]^{-1} \left(\sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\eta}_i}\boldsymbol{F}_i\right)^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{z}_i\right).$$

Finally, we apply $k-$fold cross−validation (CV) to determine the values of λ_m and λ_p which compares the multiclass AUC [44] values of a two−dimensional grid of candidate values of λ_m and λ_p.

4.3. Simulation study

4.3.1 Simulation model

We conduct a simulation study to demonstrate the performance of HisCoM−RCateg. To demonstrate the performance of HisCoM−RCateg, we generate correlated ordinal phenotype.

In order to generate the correlated ordinal phenotype y_{it}^*, consider the following marginal cumulative logit model

$$P(y_{it}^* \leq j|\boldsymbol{x}_{it}) = G\left(\beta_{0j} + \boldsymbol{x}_{it}^T\boldsymbol{W}\boldsymbol{\beta}\right),$$

where $i = 1,2,\dots,n;\ t = 1,2,\dots,T;\ j = 1,2,\dots,J$, and G is the cdf of the standard logistic distribution. The following multivariate latent regression model is considered for generating the y_{it}^*

$$\boldsymbol{u}_i = \begin{bmatrix} u_{i1} \\ \vdots \\ u_{iT} \end{bmatrix} = \begin{bmatrix} \mu_{i1} \\ \vdots \\ \mu_{iT} \end{bmatrix} + \begin{bmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{iT} \end{bmatrix} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i,$$

where $\mu_{it} = -\boldsymbol{x}_{it}^T\boldsymbol{W}\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}_i$, $i = 1,\dots,n$ denote n independent random vectors and marginally, $\epsilon_{it} \sim G\ \forall i,t$. Let $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ be a $T \times T$ latent correlation matrix for $\boldsymbol{\epsilon}_i$. Then, NORTA (NORmal To Anything) method was used [45] to generate ϵ_{it} for the marginal distribution function G with $\boldsymbol{R}_{\varepsilon}$. The NORTA method was originally introduced to generate data for any kind of marginal distribution [46]. In NORTA method, first a vector $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, \dots, Z_{iT})$ with correlation matrix $\boldsymbol{R}_Z$ was generated from standard multivariate normal (MVN) distribution. Then, the transformation $\epsilon_{it} = F^{-1}[\Phi(Z_{it})]\ \forall t$ was used, where Φ represents the

cdf of the standard normal distribution. Then $\boldsymbol{R}_Z$ can be approximated by $\boldsymbol{R}_\varepsilon$ using some mild regularity conditions, i.e. $\boldsymbol{R}_Z \approx \boldsymbol{R}_\varepsilon$ [46]. The NORTA approach, then, guarantees that marginally $\epsilon_{it} \sim G$. Then, we categorize U_{it} by the corresponding category−specific intercepts according to the following threshold to generate y_{it}^*

$$y_{it} = j \leftrightarrow \beta_{0,j-1} < U_{it} < \beta_{j0},$$

where $-\infty = \beta_{00} < \beta_{01} < \beta_{02} < \cdots < \beta_{0J} < \beta_{0,(J+1)} = \infty$.

In this simulation study, we use same biomarkers and pathways from real longitudinal metabolite data set that describe in Section 3.3 and generate the ordinal phenotype. Thus, we set the number of categories $(J+1)$ is 3, total number of pathways is $K = 65$. Here we assume that first five pathways are causal pathway and remaining 60 pathways are non−causal pathway. Let $\boldsymbol{\beta}_0 = (\beta_{01}, \beta_{02})' = (-0.3,\ 0.8)'$. For the causal pathways, we considered two different parameter settings: two biomarker level effects ($w = 0.2$ and 0.3), four pathway−level effect ($\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0.3,\ 0.4,\ 0.5, 0.6$). For non−causal pathways $\beta_6 = \beta_7 = \cdots = \beta_{65} = 0$. To generate the correlated $\boldsymbol{u}_i$ from latent regression we consider the following latent correlation matrix,

$$\boldsymbol{R}_\varepsilon = \begin{bmatrix} 1.00 & 0.85 & 0.85 \\ 0.85 & 1.00 & 0.85 \\ 0.80 & 0.85 & 1.00 \end{bmatrix}.$$

We generate 100 datasets with the sample size for each dataset being the same as the real dataset. We permute each simulated response 1000 times to calculate the p−value for pathways.

4.3.2 Simulation Results

To demonstrate the performance of the propose HisCoM_RCateg, we perform the simulation study. For the purpose of the performance comparison, we compare the type I error and power for HisCoM−RCateg with other existing pathway−based methods. We consider GSEA, and HisCoM−GEE method as existing pathway−based method. To use the HisCoM−GEE method, we use two case for simulated phenotype 0, 1+2 and 0+1, 2; because HisCoM−GEE is for binary phenotype. After generating the phenotype for each simulation, we obtained the optimal tuning parameter set (λ_m, λ_p) using the 4 folds cross−validation. Then, we evaluate the type I error and power.

Results of the empirical type I error shows in Figure 4.1, Overall, type I errors were shown to well−controlled in various method except GSEA. Especially, type I error for HisCoM−RCateg method and HisCoM−GEE (0+1,2) are more conservation compare to HisCoM−GEE (0, 1+2) and GSEA.

Results of empirical power presents in Figure 4.2, where the x−axis shows the effect sizes of pathways and y−axis shows the power. The top panel of Figure 4.2 represents the power for biomarkers effect $w = 0.2$ and the bottom panel is for biomarkers effect $w = 0.3$. HisCoM−RCateg has higher power compare to all other methods for small and large effect sizes both pathways and biomarkers. For small effect size HisCoM−GEE (0, 1+2) has higher power than HisCoM−GEE (0+1, 2). For moderate to large effect sizes HisCoM−GEE (0+1, 2) always higher power than HisCoM−GEE (0, 1+2). Again, the power of GSEA always smaller than all methods for small to large effect size GSEA. Finally, the HisCoM−

RCateg outperforms the other approaches regardless of the effect sizes.

Figure 4.1. Results of empirical type I error

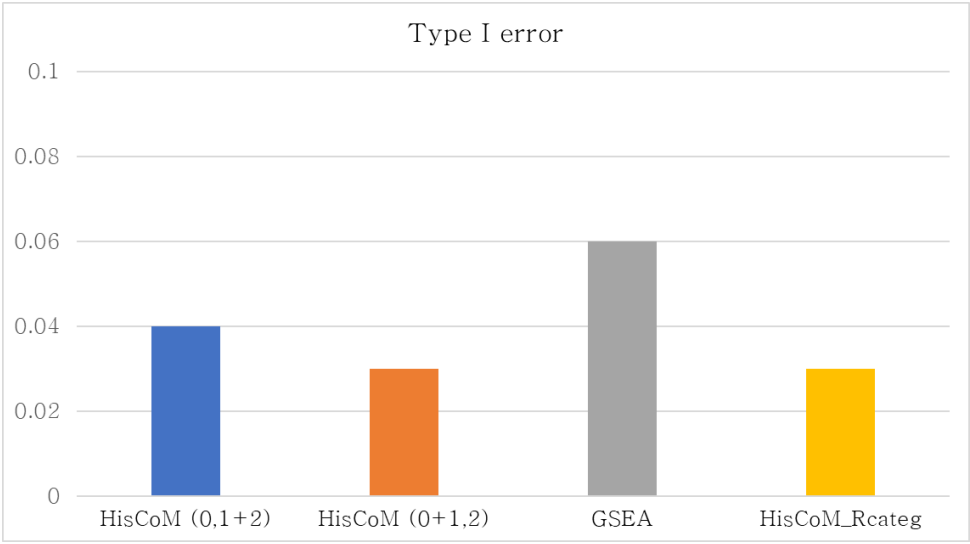
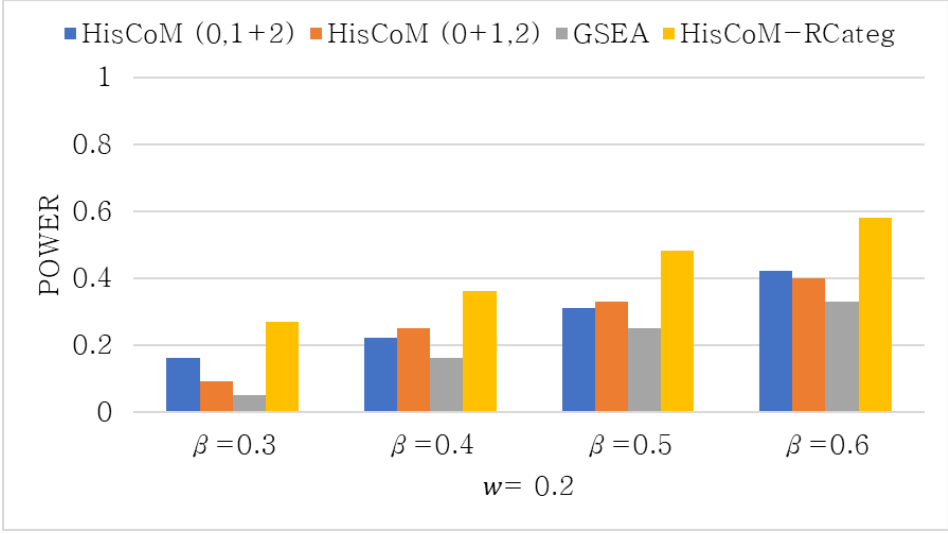
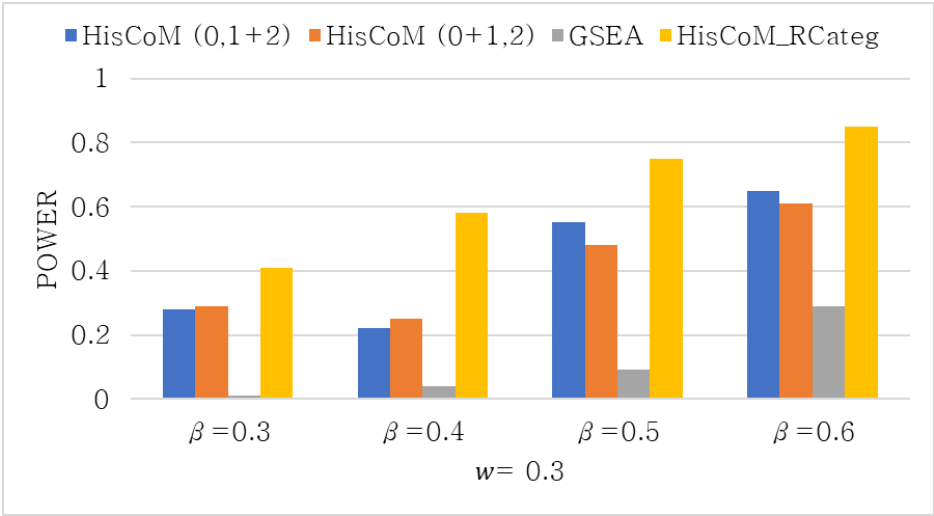


Figure 4.2. Results of empirical power





4.4. Real data analysis results

In this section, we use the KARE dataset to examine the association between pathways using HisCoM−RCateg. Description of the KARE dataset is in Section 3.3. To apply the HisCoM−RCateg, we used KARE phase 6, phase 7 and phase 8 datasets. Thus, we use 664 samples that are common in 3 different phases. To identify the pathways that associated with T2D, we performed HisCoM−RCateg, where age, gender and BMI were included as adjusting covariates. To apply the HisCoM−RCateg approach to KARE data, we considered the T2D as an ordinal phenotype. Then, ordinal phenotype y_{it}^{*} is the T2D level ($1 =$ Normal, $2 =$ Pre T2D and $3 =$ T2D) for i^{th} $(i = 1, 2, \dots, 664)$ subject at time $t(t = 1, 2, 3)$. Since the phenotype is ordinal, we applied the HisCoM−RCateg method using the following proportional odds model to the cumulative logits,

$$\text{logit}[\Pr(y_{it}^{*} \le j)] = \beta_{0j} + \sum_{k=1}^{65} \beta_{tk} \left(\sum_{m=1}^{M_k} x_{ikm} w_{km} \right) + \beta_{66} * \text{age}$$

$$+\beta_{67} * \text{gender} + \beta_{68} * \text{BMI},$$

for $j = 1, 2$.

We used 4−fold cross−validation to select the optimal tuning parameters for biomarkers (λ_m) and pathways (λ_p). We used 10000 permutations for calculating the p−values of pathways for the HisCoM−RCateg. For multiple comparison, FDR adjusted p−values (q−values) were calculated. The null hypothesis $H_0: \beta_{k1} = \beta_{k2} = \beta_{k3} = 0$ can be tested to get the global effect of a pathway. To combine the p−values we use the Fisher's method [47]. Accorfing to the Fisher's method, the test statistic for is $\mathrm{T} = -2\sum_{i=1}^{Q} log(p_i) \sim \chi^2_{(2Q)}$, where p_i is the individual p−value for each phase and $Q = 3$ is the total number of phases. The q−values for each pathway from HisCoM−RCateg are presented in the Table 4.1.

Table 4.1. Results of the q−values from HisCoM−RCateg

Pathway	Phase 6	Phase7	Phase8	Global
2−Oxocarboxylic acid metabolism	0.0072	0.0992	0.0128	1.73E−04
ABC transporters	0.0162	0.0858	0.0189	4.39E−04
African trypanosomiasis	0.0506	0.1579	0.4431	0.0497
Alanine, aspartate and glutamate metabolism	0.0022	0.0235	0.0036	1.72E−06
Alcoholism	0.0022	0.0136	0.0036	1.01E−06
Aminoacyl−tRNA biosynthesis	0.0055	0.0520	0.0105	4.00E−05
Amphetamine addiction	0.0022	0.0136	0.0036	1.01E−06
Amyotrophic lateral sclerosis	0.0022	0.0097	0.0105	8.29E−07
Arginine and proline metabolism	0.0022	0.0252	0.0251	1.98E−05
Arginine biosynthesis	0.0022	0.0340	0.0340	3.80E−05

beta-Alanine metabolism	0.2314	0.4837	0.2065	0.2039
Bile secretion	0.0237	0.0097	0.0459	5.35E-05
Biosynthesis of amino acids	0.0055	0.0858	0.0140	1.21E-04
Biosynthesis of cofactors	0.0022	0.0827	0.0036	7.25E-06
Biotin metabolism	0.2863	0.4225	0.1013	0.1257
Butanoate metabolism	0.0022	0.0385	0.0036	3.04E-06
Caffeine metabolism	0.0133	0.2579	0.3479	0.0199
cAMP signaling pathway	0.4197	0.1320	0.1488	0.0950
Carbon metabolism	0.0022	0.0113	0.0036	4.43E-07
Cholinergic synapse	0.2363	0.0984	0.1255	0.0391
Cocaine addiction	0.0022	0.0136	0.0036	1.01E-06
Cysteine and methionine metabolism	0.0510	0.1579	0.0113	0.0016
Ferroptosis	0.0022	0.0136	0.0036	1.01E-06
GABAergic synapse	0.0022	0.0136	0.0089	2.72E-06
Gap junction	0.0022	0.0136	0.0036	1.01E-06
Glutamatergic synapse	0.0022	0.0136	0.0113	3.52E-06
Glutathione metabolism	0.0022	0.0097	0.0065	4.43E-07
Glycerophospholipid metabolism	0.2363	0.0984	0.1255	0.0391
Glycine, serine and threonine metabolism	0.0133	0.0375	0.0464	0.0004
Glyoxylate and dicarboxylate metabolism	0.0022	0.0097	0.0202	2.06E-06
Histidine metabolism	0.0022	0.0505	0.0128	1.95E-05
Lysine degradation	0.0640	0.0525	0.1063	0.0051
Metabolic pathways	0.0043	0.0984	0.0262	2.04E-04
Mineral absorption	0.0517	0.0858	0.0189	0.0014
mTOR signaling pathway	0.0506	0.0841	0.2904	0.0162
Neuroactive ligand-receptor interaction	0.0022	0.0113	0.0036	4.43E-07

Nicotinate and nicotinamide metabolism	0.1232	0.5597	0.2904	0.1910
Nicotine addiction	0.0022	0.0113	0.0036	4.43E−07
Nitrogen metabolism	0.0022	0.0136	0.0113	3.52E−06
Pantothenate and CoA biosynthesis	0.0994	0.2788	0.1013	0.0388
Pathways of neurodegeneration − multiple diseases	0.0022	0.0113	0.0036	4.43E−07
Phenylalanine metabolism	0.0022	0.0510	0.0140	2.32E−05
Phenylalanine, tyrosine and tryptophan biosynthesis	0.0022	0.0858	0.0089	2.18E−05
Porphyrin metabolism	0.0022	0.0097	0.0065	4.43E−07
Primary bile acid biosynthesis	0.1650	0.0113	0.0740	0.0010
Propanoate metabolism	0.9550	0.7710	0.5584	0.9340
Protein digestion and absorption	0.0055	0.0520	0.0105	4.00E−05
Proximal tubule bicarbonate reclamation	0.0022	0.0136	0.0113	3.52E−06
Purine metabolism	0.0055	0.0097	0.1156	3.35E−05
Pyrimidine metabolism	0.2539	0.5597	0.5704	0.4937
Pyruvate metabolism	0.0087	0.7150	0.1864	0.0191
Retrograde endocannabinoid signaling	0.0022	0.0136	0.0036	1.01E−06
Salivary secretion	0.2742	0.1061	0.1124	0.0441
Sulfur metabolism	0.8921	0.1438	0.6540	0.4788
Sulfur relay system	0.0055	0.0623	0.0036	1.46E−05
Synaptic vesicle cycle	0.0022	0.0097	0.0036	2.36E−07
Taste transduction	0.0022	0.0136	0.0036	1.01E−06
Taurine and hypotaurine metabolism	0.0022	0.0136	0.0036	8.29E−07
Thermogenesis	0.3163	0.1055	0.2565	0.0987

Thiamine metabolism	0.0022	0.0136	0.0036	8.29E−07
Tryptophan metabolism	0.0506	0.1579	0.4431	0.0500
Tyrosine metabolism	0.0022	0.0984	0.0036	9.90E−06
Valine, leucine and isoleucine biosynthesis	0.1664	0.1579	0.0289	0.0125
Valine, leucine and isoleucine degradation	0.1664	0.2662	0.0299	0.0206
Vitamin digestion and absorption	0.4446	0.7450	0.2283	0.4609

4.5. Discussion

In this chapter we proposed a new method, HisCoM−RCateg, to find the association between pathway and longitudinal multinomial phenotype. While our previous HisCoM−Categ method can handle only multinomial phenotype from cross−sectional data, whereas HisCoM−RCateg uses longitudinal multinomial phenotypes. HisCoM−RCateg also able to handle both time dependent and time independent biomarkers. HisCoM−RCateg evaluates the relationship between pathways and a multinomial phenotype in a single model. To develop the HisCoM−RCateg we use the basic framework of the GEE for categorical response. Like as HisCoM−Categ, HisCoM−RCateg is flexible enough to be used for both nominal and ordinal phenotypes. Through the simulation study we show that HisCoM−RCateg performs better than other methods. The analysis of a real dataset with T2D phenotypes, HisCoM−RCateg, can identify pathways that have an associated with multinomial phenotype. Therefore, we fully

expect that HisCoM-Categ will help the researchers identify the pathways that are associated with multinomial phenotypes. In conclusion, we hope that HisCoM-RCateg can serve as a main tool for pathway analysis of longitudinal multinomial phenotypes for omics data.

Chapter 5. Parametric testing for hierarchical structural component models

5.1. Introduction

In chapter 5, a parametric testing approach HisCoM is propose. For testing the significance of the pathway effect, the original HisCoM uses the permutation approach. When the asymptotic distribution is unknown, the permutation test is useful for generating the exact distribution under the null hypothesis. Sometimes, the permutation test is problematic for high-dimensional data because of its computational burden and time. HisCoM was originally developed for high-dimensional data, so it requires a long time to get the significant value of pathways. To account for this issue, in this chapter we introduce a parametric test to get the significant value of pathways. To do this, first, we estimated the asymptotic variance of pathways and then use the Wald type test. After that, we performe the non-centrality test to find the significant value of pathways.

5.2. Methods

5.2.1 HisCoM

From chapter 2, the HisCoM model can be written as,

$$\eta_i = g(\pi_i) = \beta_0 + \sum_{k=1}^{K} f_{ik}\beta_k = \beta_0 + \sum_{k=1}^{K}\left[\sum_{m=1}^{M_k} x_{ikm}w_{km}\right]\beta_k = \boldsymbol{F\beta} = \boldsymbol{XW\beta}$$

We aim to maximize the following penalized log-likelihood function to estimate the parameter

$$Q(\boldsymbol{W},\boldsymbol{\beta}) = l(\boldsymbol{W},\boldsymbol{\beta}) - \frac{1}{2}\lambda_G\|\boldsymbol{W}\|^2 - \frac{1}{2}\lambda_P\|\boldsymbol{\beta}\|^2$$

where $Q(\boldsymbol{W},\boldsymbol{\beta})$ is the penalized loglikelihood function and $l(\boldsymbol{W},\boldsymbol{\beta})$ is the log likelihood function.

Theorem 5.1: Assume $y_1, \ldots, y_n$ are independent with pdf $f(y_i|\,\boldsymbol{w}_0, \boldsymbol{\beta}_0)$ for $\boldsymbol{w}_0 \in \boldsymbol{\Omega}_{\boldsymbol{w}_o}$ and $\boldsymbol{\beta}_0 \in \boldsymbol{\Omega}_{\boldsymbol{\beta}_o}$, where $\boldsymbol{w}_0$ and $\boldsymbol{\beta}_0$ are the true values of $\boldsymbol{w}$ and $\boldsymbol{\beta}$. If $n \to \infty$, $\lambda_m = O(\sqrt{n})$ and $\lambda_p = O(\sqrt{n})$ then

$$\|\widehat{\boldsymbol{w}}_0 - \boldsymbol{w}_0\| = O_p\left(\frac{1}{\sqrt{n}}\right),\ \|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0\| = O_p\left(\frac{1}{\sqrt{n}}\right)$$

Proof:

We want to show that for any given $\varepsilon > 0$, there exist a large constant C such that

$$\mathrm{P}\left\{\sup_{\boldsymbol{u}=(\boldsymbol{u}_1^T,\boldsymbol{u}_2^T)^T:\,\|\boldsymbol{u}\|=C} Q\left(\boldsymbol{w}_0 + \frac{1}{\sqrt{n}}\boldsymbol{u}_1, \boldsymbol{\beta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{u}_2\right) < Q(\boldsymbol{w}_0,\boldsymbol{\beta}_0)\right\} \geq 1 - \varepsilon$$

This implies with probability $1-\varepsilon$ that there exists a local maximizer of $Q(\boldsymbol{w}_0,\boldsymbol{\beta}_0)$ in the ball $\left\{\boldsymbol{w}_0 + \frac{1}{\sqrt{n}}\boldsymbol{u}_1, \boldsymbol{\beta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{u}_2 \colon \left\|(\boldsymbol{u}_1^T, \boldsymbol{u}_2^T)^T\right\| \leq C\right\}$. Hence, there exists a local maximizer such that

$$\|\widehat{\boldsymbol{w}}_0 - \boldsymbol{w}_0\| = O_p\left(\frac{1}{\sqrt{n}}\right),\ \|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0\| = O_p\left(\frac{1}{\sqrt{n}}\right)$$

We have,

$$\begin{aligned} D(\boldsymbol{w}_0,\boldsymbol{\beta}_0) &= Q\left(\boldsymbol{w}_0 + \frac{1}{\sqrt{n}}\boldsymbol{u}_1, \boldsymbol{\beta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{u}_2\right) - Q(\boldsymbol{w}_0,\boldsymbol{\beta}_0) \\ &= \left[L\left(\boldsymbol{w}_0 + \frac{1}{\sqrt{n}}\boldsymbol{u}_1, \boldsymbol{\beta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{u}_2\right) - L(\boldsymbol{w}_0,\boldsymbol{\beta}_0)\right] \\ &\quad - \lambda_m \sum_{j=1}^{m}\left\{\left(w_{0j} + \frac{1}{\sqrt{n}}u_{1j}\right)^2 - w_{0j}\right\} \end{aligned}$$

$$-\lambda_p \sum_{k=1}^{K} \left\{ \left(\beta_{0k} + \frac{1}{\sqrt{n}} u_{2k} \right)^2 - \beta_{0k} \right\}$$
$$= (G_1) - (G_2).$$

Now, using the Taylor series approximation

$$(G_1) = \left[L\left(w_0 + \frac{1}{\sqrt{n}} \boldsymbol{u}_1, \beta_0 + \frac{1}{\sqrt{n}} \boldsymbol{u}_2 \right) - L(w_0, \beta_0) \right]$$
$$= L(w_0, \beta_0) + \frac{1}{\sqrt{n}} \Delta_{w_o} L(w_0, \beta_0)\, \boldsymbol{u}_1 + \frac{1}{\sqrt{n}} \Delta_{\beta_o} L(w_0, \beta_0)\, \boldsymbol{u}_2$$
$$+ \frac{1}{2} \boldsymbol{u}_1^T \left(\frac{1}{n} \right) \Delta_{w_o}^2 L(w_0, \beta_0)\, \boldsymbol{u}_1 + \frac{1}{2} \boldsymbol{u}_2^T \left(\frac{1}{n} \right) \Delta_{\beta_o}^2 L(w_0, \beta_0)\, \boldsymbol{u}_2$$
$$+ \boldsymbol{u}_1^T \left(\frac{1}{n} \right) \Delta_{w_o \beta_o}^2 L(w_0, \beta_0)\, \boldsymbol{u}_2 - L(w_0, \beta_0)$$
$$= \frac{1}{\sqrt{n}} \Delta_{w_o} L(w_0, \beta_0)\, \boldsymbol{u}_1 + \frac{1}{\sqrt{n}} \Delta_{\beta_o} L(w_0, \beta_0)\, \boldsymbol{u}_2$$
$$+ \frac{1}{2} \boldsymbol{u}_1^T \left(\frac{1}{n} \right) \Delta_{w_o}^2 L(w_0, \beta_0)\, \boldsymbol{u}_1 + \frac{1}{2} \boldsymbol{u}_2^T \left(\frac{1}{n} \right) \Delta_{\beta_o}^2 L(w_0, \beta_0)\, \boldsymbol{u}_2$$
$$+ \boldsymbol{u}_1^T \left(\frac{1}{n} \right) \Delta_{w_o \beta_o}^2 L(w_0, \beta_0)\, \boldsymbol{u}_2.$$

Since,

$$\frac{1}{\sqrt{n}} \Delta_{w_o} L(w_0, \beta_0) = O_p(1)$$
$$\frac{1}{\sqrt{n}} \Delta_{\beta_o} L(w_0, \beta_0) = O_p(1)$$
$$-\frac{1}{n} \Delta_{w_o}^2 L(w_0, \beta_0) \to^p \boldsymbol{I}_{w_0}(w_0, \beta_0)$$
$$-\frac{1}{n} \Delta_{w_o \beta_o}^2 L(w_0, \beta_0) \to^p \boldsymbol{I}_{w_o \beta_o}(w_0, \beta_0)$$
$$-\frac{1}{n} \Delta_{\beta_o}^2 L(w_0, \beta_0) \to^p \boldsymbol{I}_{\beta_o}(w_0, \beta_0).$$

Therefore,

$$G_1 = O_p(1)\boldsymbol{u}_1 + O_p(1)\boldsymbol{u}_2 - \frac{1}{2} \boldsymbol{u}_1^T \boldsymbol{I}_{w_0}(w_0, \beta_0) \boldsymbol{u}_1$$

$$-\frac{1}{2}\boldsymbol{u}_1^T\boldsymbol{I}_{w_o\beta_o}(w_0,\beta_0)\boldsymbol{u}_2 \; -\frac{1}{2}\boldsymbol{u}_2^T\boldsymbol{I}_{\beta_o}(w_0,\beta_0)\boldsymbol{u}_2$$

$$= G_{11} + G_{12} + G_{13} + G_{14} + G_{15}.$$

Here, $G_{11} + G_{12}$ is dominated by $G_{13} + G_{14} + G_{15}$ for sufficiently large C.

Theorem 5.2: Let $\boldsymbol{\gamma} = (\boldsymbol{w}^T, \boldsymbol{\beta}^T)^T$ and $p_\lambda(\boldsymbol{\gamma}) = \frac{1}{2}\lambda_G\|\boldsymbol{W}\|^2 + \frac{1}{2}\lambda_P\|\boldsymbol{\beta}\|^2$. Assume $y_1, \ldots, y_n$ are independent with pdf $f(y_i|\,\boldsymbol{\gamma})$ for $\boldsymbol{\gamma_0} \in \boldsymbol{\Omega_\gamma}$, where $\boldsymbol{\gamma_0}$ are the true values of $\boldsymbol{\gamma}$. Let $\widehat{\boldsymbol{\gamma}}$ are the estimates of $\boldsymbol{\gamma}$. Then

$$\sqrt{n}\left(\boldsymbol{I}(\boldsymbol{\gamma}_0) + \frac{1}{n}\boldsymbol{p}''_\lambda(\boldsymbol{\gamma_0})\right)\left\{(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma_0}) + \frac{1}{n}\left(\boldsymbol{I}(\boldsymbol{\gamma}_0) + \frac{1}{n}\boldsymbol{p}''_\lambda(\boldsymbol{\gamma_0})\right)^{-1}\boldsymbol{p}'_\lambda(\boldsymbol{\gamma_0})\right\} \rightarrow N\big(0, \boldsymbol{I}(\boldsymbol{\gamma}_0)\big)$$

where $\boldsymbol{p}'_\lambda$ and $\boldsymbol{p}''_\lambda$ are the first and second derivative of $p_\lambda(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$.

Proof: Expanding the function $\frac{\partial Q(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}$ into a Taylor series about $\boldsymbol{\gamma}_0$ and evaluation it at $\widehat{\boldsymbol{\gamma}}$, we get

$$\frac{\partial Q(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}|_{\boldsymbol{\gamma}=\widehat{\boldsymbol{\gamma}}} = \frac{\partial l(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}|_{\boldsymbol{\gamma}=\widehat{\boldsymbol{\gamma}}} - \boldsymbol{p}'_\lambda(\widehat{\boldsymbol{\gamma}})$$

$$\Rightarrow \boldsymbol{0} = \frac{\partial l(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}} + \frac{\partial l^2(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}^T}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma_0}) - \boldsymbol{p}'_\lambda(\boldsymbol{\gamma_0}) - \boldsymbol{p}''_\lambda(\boldsymbol{\gamma_0})(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma_0})$$

$$\Rightarrow \frac{\partial l(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}} = \left[-\frac{\partial l^2(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}^T} + \boldsymbol{p}''_\lambda(\boldsymbol{\gamma_0})(\boldsymbol{\gamma_0})\right](\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma_0}) + \boldsymbol{p}'_\lambda(\boldsymbol{\gamma_0})$$

$$\Rightarrow \frac{1}{\sqrt{n}}\frac{\partial l(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}} = \sqrt{n}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma_0})\left[-\frac{1}{n}\frac{\partial l^2(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}^T} + \frac{1}{n}\boldsymbol{p}''_\lambda(\boldsymbol{\gamma_0})(\boldsymbol{\gamma_0})\right] + \frac{1}{\sqrt{n}}\boldsymbol{p}'_\lambda(\boldsymbol{\gamma_0}),$$

Since $E\left(\frac{\partial l(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}}\right) = \boldsymbol{0}$, by central limit theorem (CLT)

$$\frac{1}{\sqrt{n}}\frac{\partial l(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}} \rightarrow^{D} N\big(0, I(\boldsymbol{\gamma}_0)\big),$$

By the law of large numbers,

$$-\frac{1}{n}\frac{\partial l^2(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}^T} \rightarrow^{P} E\left(-\frac{1}{n}\frac{\partial l^2(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}^T}\right).$$

That is

$$-\frac{1}{n}\frac{\partial l^2(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}^T} \rightarrow^{P} I(\boldsymbol{\gamma}_0).$$

Then,

$$\left\{\sqrt{n}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma_0})\left(I(\boldsymbol{\gamma}_0) + \frac{1}{n}\boldsymbol{p}''_{\boldsymbol{\lambda}}(\boldsymbol{\gamma_0})\right) + \frac{1}{\sqrt{n}}\boldsymbol{p}'_{\boldsymbol{\lambda}}(\boldsymbol{\gamma_0})\right\} \rightarrow N\big(0, I(\boldsymbol{\gamma}_0)\big)$$

$$\Rightarrow \sqrt{n}\left(I(\boldsymbol{\gamma}_0) + \frac{1}{n}\boldsymbol{p}''_{\boldsymbol{\lambda}}(\boldsymbol{\gamma_0})\right)\left\{(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma_0}) + \frac{1}{n}\left(I(\boldsymbol{\gamma}_0) + \frac{1}{n}\boldsymbol{p}''_{\boldsymbol{\lambda}}(\boldsymbol{\gamma_0})\right)^{-1}\boldsymbol{p}'_{\boldsymbol{\lambda}}(\boldsymbol{\gamma_0})\right\} \rightarrow N\big(0, I(\boldsymbol{\gamma}_0)\big)$$

Thus, the asymptotic covariance matrix of $\widehat{\boldsymbol{\gamma}}$ is,

$$\{I_n(\boldsymbol{\gamma}_0) + \boldsymbol{p}''_{\boldsymbol{\lambda}}(\boldsymbol{\gamma_0})\}^{-1} I_n(\boldsymbol{\gamma}_0) \{I_n(\boldsymbol{\gamma}_0) + \boldsymbol{p}''_{\boldsymbol{\lambda}}(\boldsymbol{\gamma_0})\}^{-1}$$

where $I_n(\boldsymbol{\gamma}_0) = nI(\boldsymbol{\gamma}_0)$.

5.3. Hypothesis test

To check the effect of an individual pathway on the phenotype, we consider the following null hypothesis

$$H_0: \beta_k = 0 \; vs. \; H_a: \beta_k \neq 0.$$

To perform the test of hypothesis we use the following Wald type test

$$z = \frac{\hat{\beta}_k}{se(\hat{\beta}_k)} \sim N(0,1).$$

Under the H_0, the Wald statistic $W_\beta = z^2 \sim \chi^2_{(1)}$. Under the H_a, the Wald statistic $W_\beta \sim \chi^2_{(1,\delta)}$, where δ is the non-centrality parameter. The mean of this non-central $\chi^2_{(df,\delta)}$ random variable is $\delta + df$. Thus, the mean of W_β is $\delta + 1$, and we estimate the non-centrality parameter as $\hat{\delta} = \max(0, \hat{\mu} - 1)$, where $\hat{\mu}$ is the mean of W_β under the H_0. To estimate $\hat{\mu}$, we permute the phenotype a few times and calculate W_β, then take a sample mean for W_β as $\hat{\mu}$. Empirically we determine the number of permutations is 100 for calculating the sample mean $\hat{\mu}$. Then we calculate the pathway significance value using both the central and non-central approach and compare them with our gold standard permutation p-value.

Again, to adjust the asymptotic test, we do the saddle point approximation and df adjustment. Further, we also perform the modified asymptotic test using the modification for the objective function of parameter estimation of the original HisCoM. To do this we revisit the HisCoM and consider the single ridge penalty for the product of biomarker effect and pathway effect rather than consider the double ridge penalty.

5.4. Modified asymptotic test

Recall the penalized log-likelihood function

$$Q(\boldsymbol{W}, \boldsymbol{\beta}) = l(\boldsymbol{W}, \boldsymbol{\beta}) - \frac{1}{2}\lambda_G \|\boldsymbol{W}\|^2 - \frac{1}{2}\lambda_P \|\boldsymbol{\beta}\|^2$$

Instead of using separate penalty function, consider the single penalty function for the product of $\boldsymbol{W}$ and $\boldsymbol{\beta}$. Thus, the objective function is

$$Q(\boldsymbol{W},\boldsymbol{\beta}) = l(\boldsymbol{W},\boldsymbol{\beta}) - \frac{1}{2}\lambda \|\boldsymbol{W\beta}\|^2$$

To estimate the $\boldsymbol{W}$ and $\boldsymbol{\beta}$, separately, derivative of $p_\lambda(\boldsymbol{W\beta})$ with respect to $\boldsymbol{W}$ and $\boldsymbol{\beta}$. The first and second derivative of $p_\lambda(\boldsymbol{W\beta})$ with respect to $\boldsymbol{W}$ is

$$\frac{\partial[p_\lambda(\boldsymbol{W\beta})]}{\partial \boldsymbol{W}} = diag(\beta_1, \dots \beta_1, \ \dots, \beta_K, \dots, \beta_K)\boldsymbol{w}_*\boldsymbol{\beta}_*,$$

$$\Rightarrow \frac{\partial^2[p_\lambda(\boldsymbol{W\beta})]}{\partial \boldsymbol{W}\partial \boldsymbol{W}^T} = \lambda diag(\beta_1^2, \dots, \beta_1^2, \dots, \beta_K^2, \dots, \beta_K^2)$$

$$= \lambda bdiag\left(diag(\beta_1^2), \ \dots, diag(\beta_K^2)\right),$$

where $\boldsymbol{\beta}_*$ is the vector of $\boldsymbol{\beta}$ without intercept term and $\boldsymbol{w}_*$ is the matrix of $\boldsymbol{W}$ without first column and first row and $diag(\beta_k^2)$, $k = 1, \dots, K$ is $M_k \times M_k$ diagonal matrix, where $\mathrm{M_k}$ is the number of biomarkers in the $\mathrm{k^{th}}$ pathway. Again, the first and second derivative of $p_\lambda(\boldsymbol{W\beta})$ with respect to $\boldsymbol{\beta}$

$$\frac{\partial[p_\lambda(\boldsymbol{W\beta})]}{\partial \boldsymbol{\beta}} = \lambda \boldsymbol{W}^T\boldsymbol{W\beta},$$

$$\Rightarrow \frac{\partial^2[p_\lambda(\boldsymbol{W\beta})]}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} = \lambda \boldsymbol{W^2}.$$

Theorem 5.3: Let $\boldsymbol{\gamma_{wb}} = \boldsymbol{W\beta}$ and $p_\lambda(\boldsymbol{\gamma_{wb}})$ $\frac{1}{2}\lambda \|\boldsymbol{W\beta}\|^2$ Assume $\mathrm{y}_1, \dots, y_n$ are independent with pdf $f(y_i|\,\boldsymbol{\gamma_{wb}})$ for $\boldsymbol{\gamma_{wb}^0} \in \boldsymbol{\Omega_{\gamma_{wb}}}$, where $\boldsymbol{\gamma_{wb}^0}$ are the true values of $\boldsymbol{\gamma_{wb}}$. Let $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{wb}}$ are the estimates of $\boldsymbol{\gamma_{wb}}$. Then

$$\sqrt{n}\left(\boldsymbol{I}(\boldsymbol{\gamma}_{\boldsymbol{wb}}^{0})+\frac{1}{n}\boldsymbol{p}_{\boldsymbol{\lambda}}''(\boldsymbol{\gamma}_{\boldsymbol{wb}}^{0})\right)\left\{(\widehat{\boldsymbol{\gamma}}_{\boldsymbol{wb}}-\boldsymbol{\gamma}_{\boldsymbol{wb}}^{0})\right.$$
$$\left.+\frac{1}{n}\left(\boldsymbol{I}(\boldsymbol{\gamma}_{\boldsymbol{wb}}^{0})+\frac{1}{n}\boldsymbol{p}_{\boldsymbol{\lambda}}''(\boldsymbol{\gamma}_{\boldsymbol{wb}}^{0})\right)^{-1}\boldsymbol{p}_{\boldsymbol{\lambda}}'(\boldsymbol{\gamma}_{\boldsymbol{wb}}^{0})\right\}\rightarrow \boldsymbol{N}\left(0,\boldsymbol{I}(\boldsymbol{\gamma}_{\boldsymbol{wb}}^{0})\right).$$

Theorem 5.3 can be proved in the similar way to the Theorem 5.2.

Then, the asymptotic covariance matrix of $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{wb}}$ is,

$$\left\{\boldsymbol{I}_{n}(\boldsymbol{\gamma}_{\boldsymbol{wb}}^{0})+\boldsymbol{p}_{\boldsymbol{\lambda}}''(\boldsymbol{\gamma}_{\boldsymbol{wb}}^{0})\right\}^{-1}\boldsymbol{I}_{n}(\boldsymbol{\gamma}_{\boldsymbol{wb}}^{0})\left\{\boldsymbol{I}_{n}(\boldsymbol{\gamma}_{\boldsymbol{wb}}^{0})+\boldsymbol{p}_{\boldsymbol{\lambda}}''(\boldsymbol{\gamma}_{\boldsymbol{wb}}^{0})\right\}^{-1},$$

where $\boldsymbol{I}_{n}(\boldsymbol{\gamma}_{\boldsymbol{wb}}^{0})=n\boldsymbol{I}(\boldsymbol{\gamma}_{\boldsymbol{wb}}^{0})$.

Using the result of this asymptotic theorem, we then perform hypothesis test of each biomarker effect to the phenotype via pathways. That is the null hypothesis for m^{th} biomarker in k^{th} pathway is

$$H_{0}: w_{km}\beta_{k}=0 \text{ vs } H_{1}: w_{km}\beta_{k}\neq 0.$$

We use the Wald type of test and permutation test to perform the testing of the above null hypothesis. Compare their results discuss in result section.

Moreover, using this asymptotic result we further test the effect of pathway effect using the following two hypotheses.

Hypothesis 1: Consider full degrees of freedom

$$H_{0}: \boldsymbol{C}_{k}\boldsymbol{W}\boldsymbol{\beta}=0$$

Hypothesis 2: Consider one degrees of freedom

$$H_{0}: \sum_{m=1}^{M_{k}}\beta_{k}w_{km}=0$$

In summary, to assess the pathway effect to the phenotype we perform 7 different test including permutation test.

1. Permutation test for $H_0: \beta_k = 0$
2. Asymptotic test for $H_0: \beta_k = 0$
3. Non-centrality test for $H_0: \beta_k = 0$
4. DF adjustment for $H_0: \beta_k = 0$
5. Saddle point Approximation for $H_0: \beta_k = 0$
6. Modified Asymptotic test for with full df, $H_0: \boldsymbol{C}_k \boldsymbol{W} \boldsymbol{\beta} = 0$
7. Modified Asymptotic test for with one df, $H_0: \sum_{m=1}^{M_k} \beta_k w_{km} = 0$

5.5. Results

To compare the results of the parametric test with the permutation test we perform a simulation study and real data analysis. In real data analysis, we consider four examples. In example 1, we choose 5 non-overlapping pathways; in example 2, 10 overlapping pathways; in example 3, 20 overlapping pathways from the KARE phase 6 dataset and finally for example 4, we consider the KARE phase 6 dataset.

5.5.1 Number of permutations of non-central test

To determine the number of permutations for the non-central parameter we use the first three examples. First, we permute the phenotype 10, 20, 30, 40, 50, and 100 times; and calculate the noncentral parameters. Then repeat this process 10 times and

calculate the mean and confidence interval. Figure 5.1 shows the results for the first example with five pathways. In Figure 5.1, the x−axis is for pathways; and the y−axis is for mean and CI for the noncentral parameter. Figure 5.1 shows that when the number of permutations is small noncentral parameter is varied but for a large number of permutations noncentral parameter is not varied. Results of the second and third examples are shown in Figure 5.2 and Figure 5.3. In both Figure 5.2 and Figure 5.3, the x−axis shows the number of repetitions and the y−axis shows the mean and CI, and each panel is for each pathway. Figure 5.2 shows the mean and CI of 6 pathways from 10 pathways in example 2. Figure 5.3 shows the mean and CI of 6 pathways from 20 pathways in example 2. Both Figure 5.2 and Figure 5.3 shows that when the number of permutation small noncentral parameter vary a lot but for a large number of permutations noncentral parameter changes slightly. Thus, in our study, we fix the number of permutations as 100 for calculating the non−central parameter for hypothesis testing.

Figure 5.1. Mean and CI for noncentral parameter with repetition for example with 5 pathways

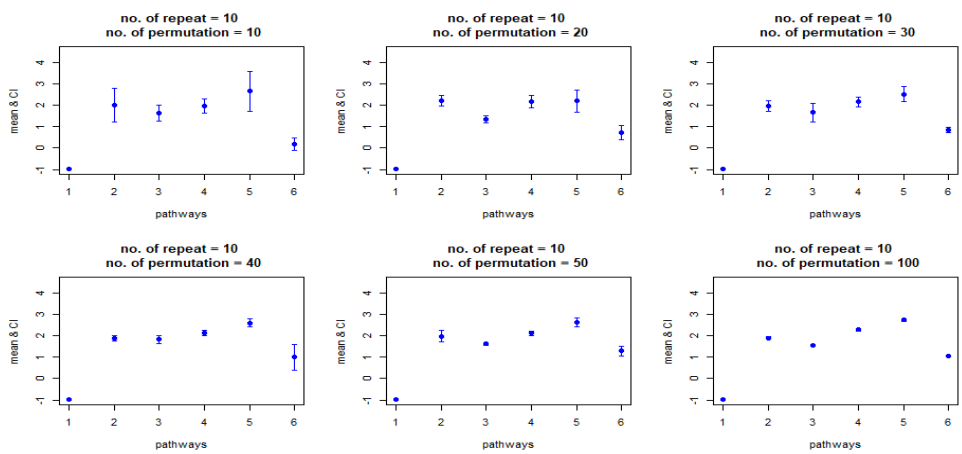


Figure 5.2. Mean and CI for noncentral parameter with repetition for example with 10 pathways

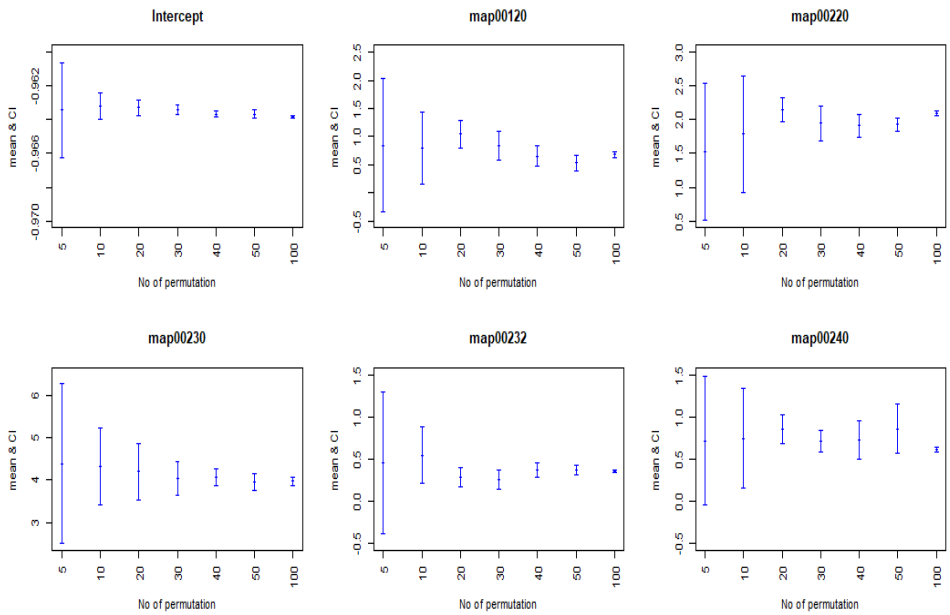
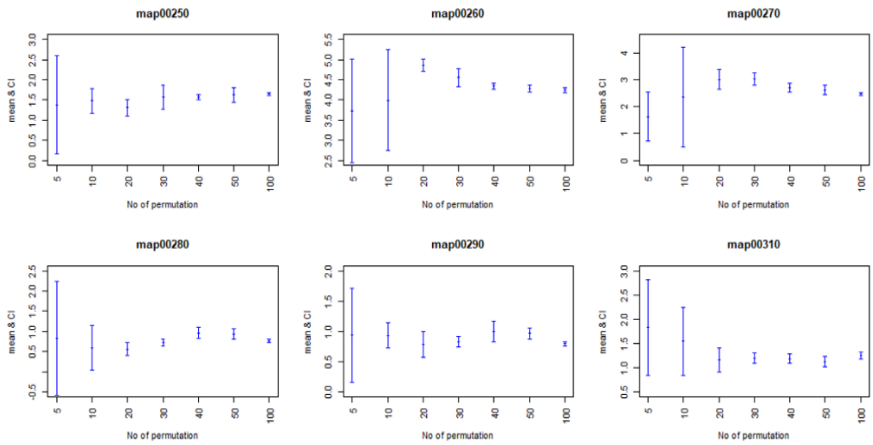


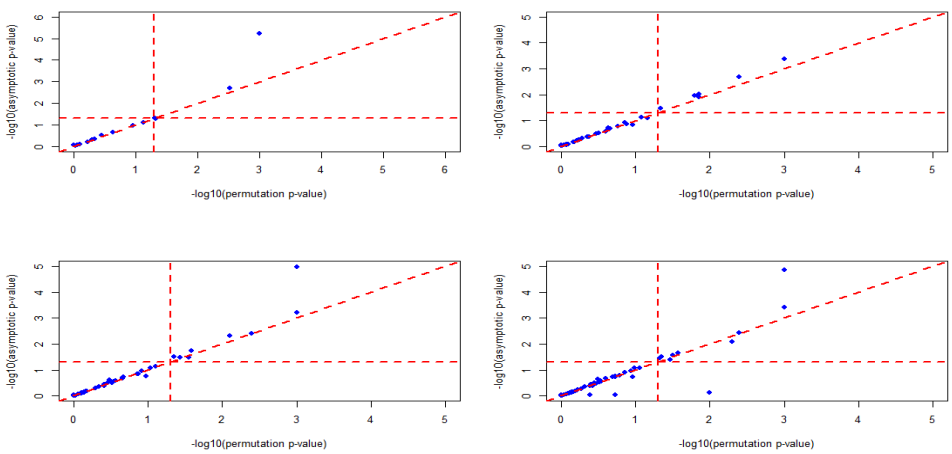
Figure 5.3. Mean and CI for noncentral parameter with repetition for example with 20 pathways



5.5.2 Comparison of the results of testing the modified asymptotic test

Figure 5.4 shows the comparison of $-\log_{10}(p\text{-value})$ between permutation and asymptotic test for $H_0: w_{km}\beta_k = 0$ test. In Figure 5.4, the left top plot is for example 1, the right top plot is for example 2, the left bottom plot is for example 3, and the right bottom plot is for example 4. Figure 5.4 shows that the asymptotic test is similar to the permutation for most of the cases.

Figure 5.4. Comparison of $-log_{10}(p-\text{value})$ for permutation vs asymptotic test $H_0: w_{km}\beta_k = 0$



5.5.3 Comparison of the results of pathway effect test

Figure 5.5 to Figure 5.8 shows the comparison of $-\log_{10}(p\text{-value})$ of pathway effect test using different hypotheses. For comparison, we considered the permutation approach as a gold standard. We use the correlation to measure the degree of the

relation of p−values from two different approaches. Also, to test whether two sets of p−values from different testing approaches come from the same distribution or not, we perform the Kolmogorov−Smirnov test. For all examples, p−values from the non−centrality test are close to the permutation test and their correlation is higher than the other methods. SAP and df adjustment contribute nothing, it's similar to the asymptotic result. The modified asymptotic test with full df performs better than the modified asymptotic test with one df. For the first three examples, the modified asymptotic test with full df is almost similar to the asymptotic test. Figure 5.8 shows that, the p−values of the modified asymptotic test with full df inflated more than that of the asymptotic test. The modified asymptotic test with one df always provides less power compared to the other methods. In summary, the non−central test provides consistent results for all examples and it can be used as the alternative method for the permutation approach in HisCoM.

Figure 5.5. Comparison of $-\boldsymbol{log_{10}}(\boldsymbol{p}-\text{value})$ of pathway effect test for example 1

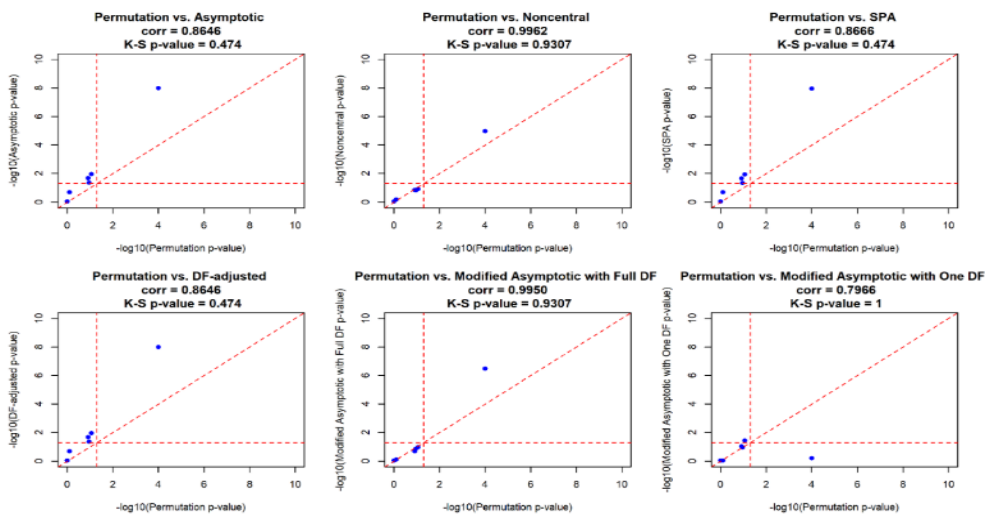


Figure 5.6. Comparison of $-\boldsymbol{log_{10}}(\boldsymbol{p}-\text{value})$ of pathway effect test for example 2

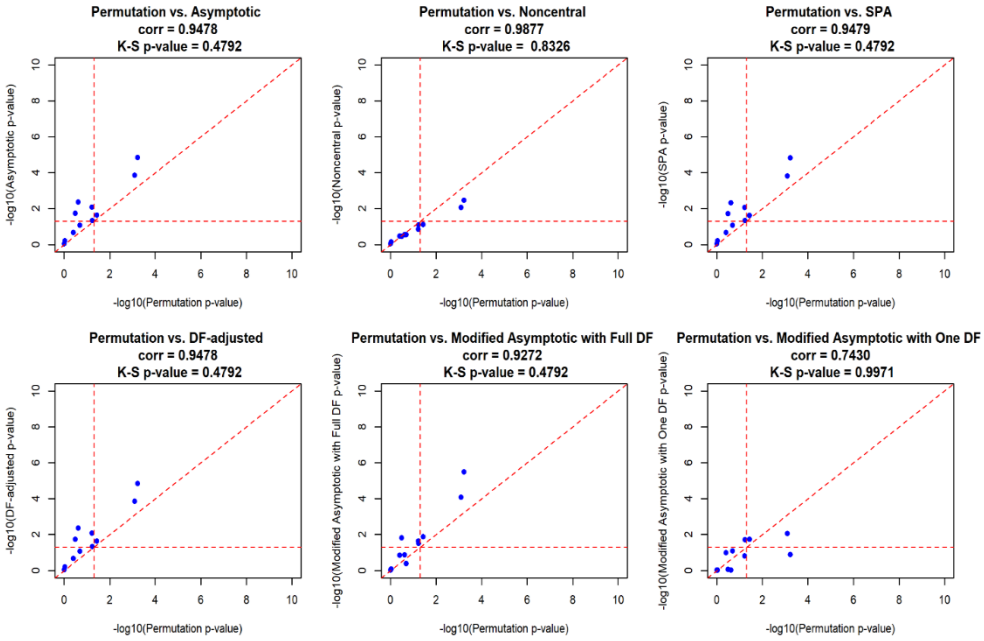


Figure 5.7. Comparison of $-\boldsymbol{log_{10}}(\boldsymbol{p}-\text{value})$ of pathway effect test for example 3

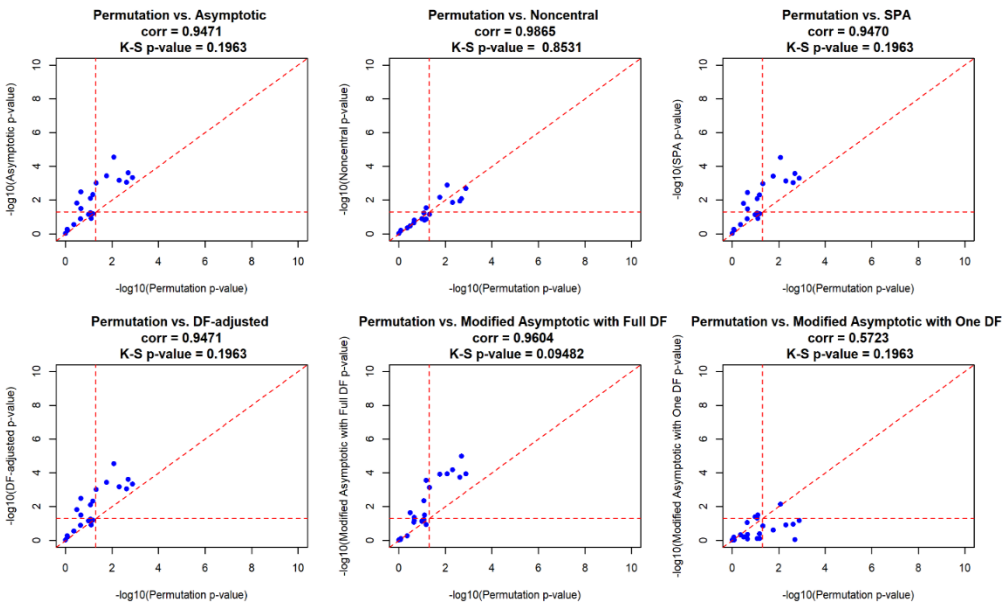
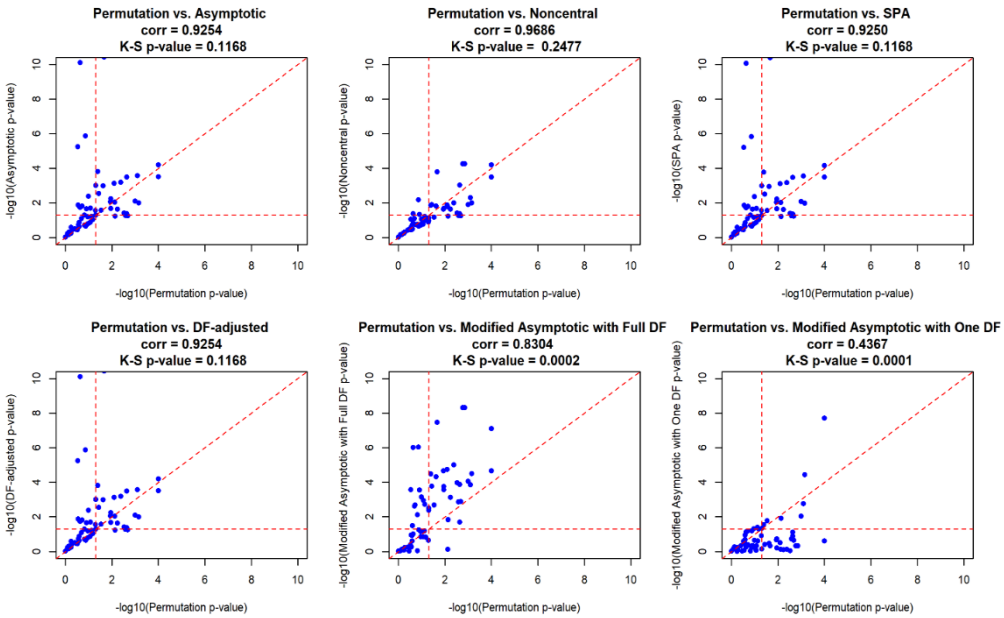


Figure 5.8. Comparison of $-\boldsymbol{log_{10}}(\boldsymbol{p}-$value) of pathway effect test for example 4



5.6. Simulation study

5.6.1 Simulation model

We perform a simulation study to compare the performance of propose different types of parametric tests in HisCoM. To evaluate the performance, we generate a binary phenotype. Also, in our simulation study, we use real metabolite data from KARE phase 6. We conduct the simulation from metabolite data. In generate the binary phenotype, consider the following logit model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{k=1}^{K}\left[\sum_{m=1}^{M_k} x_{ikm}\right]\beta_k,$$

where $i = 1, \ldots, n$. The total number of pathways $K = 65$ is in the metabolite dataset in the metagenome dataset. From metabolite data, we randomly select 5 pathways are causal pathways and the remaining 60 pathways as non-causal pathways. For the causal pathways, we considered two different parameter settings: two biomarker level effects ($w = 0.2$ and 0.3), and four pathway-level effects ($\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0.3,\ 0.4,\ 0.5,\ 0.6$), and the effect of non-causal pathways is zero. We generated 1000 datasets with the sample size for each dataset being the same as the real dataset for Type I error and 100 datasets for power calculation. To calculate the p-value for the permutation test, we permute each simulated response 1000 times to calculate the type-I error and power. Again, to calculate the p-value for the non-central χ^2 test, we permuted each simulated response 1000 times.

5.6.2 Simulation results

To demonstrate the statistical performance of the proposed parametric testing approach for HisCoM we perform the simulation study. For the performance comparison, we compare the type I error and power for seven different tests of HisCoM. To do so, first, we generate the binary phenotype from KARA phase 6 metabolite data. After generating the phenotype for each simulation, we obtained the optimal tuning parameter set (λ_m, λ_p) using the 3 folds cross-validation. Then, we evaluate the type I error and power.

Results of the empirical type I error are shown in Figure 5.9. The permutation test, the modified asymptotic test with full df, and the modified asymptotic test with one df successfully control the

type I error in simulation for the metabolite data. Type I error for the non−central χ^2 test is seeming good after 100 data generation but it is higher after 1000 data generation. Again, the type I error cannot be controlled by an asymptotic χ^2 test, saddle point approximation, and df adjustment. The modified asymptotic test with one df provides a lower type I error compare to all other methods.

Results of empirical power are present in Figure 5.11, where the x−axis shows the effect sizes of pathways and the y−axis shows the power. The left panel of Figure 5.11 represents the power for biomarkers effect $w = 0.2$ and the right panel is for biomarkers effect $w = 0.3$. Power for the non−central χ^2 test is comparable with the permutation test. Power for the asymptotic test, the SPA, and the df adjustment test is always higher than permutation and noncentral test; but they cannot control type I error. Again, the modified asymptotic test with full df and the modified asymptotic test with one df always provide higher power compared to the permutation test and non−central test, also they control type I error well. The modified asymptotic test with one df provides slightly higher power compared to the modified asymptotic test with full df.

Based on the simulation study, we can use the non−central χ^2 test for the original HisCoM rather than the permutation test to reduce the computational burden. Otherwise, we can use the modified asymptotic test with full df and the modified asymptotic test with full df rather than the permutation test in the original HisCoM. In the permutation test, we test the individual pathway effect (i.e., $\mathrm{H}_0: \beta_k = 0$) but in the modified asymptotic test we check the effect of the biomarker to phenotype via pathway ($H_0: \boldsymbol{C}_k \boldsymbol{W} \boldsymbol{\beta} = 0$ and $H_0: \sum_{m=1}^{M_k} \beta_k w_{km} = 0$). Since

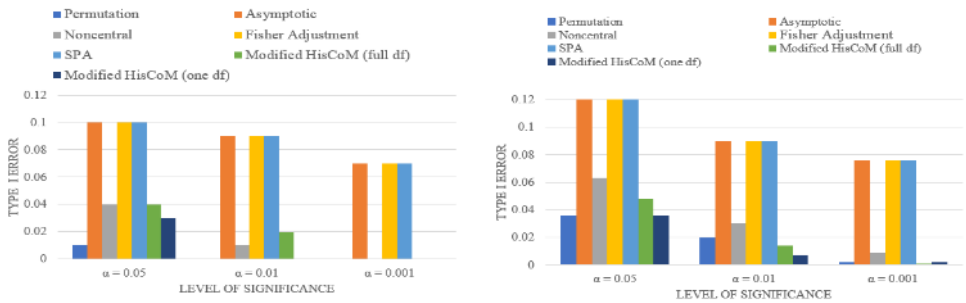
$$\mathrm{H}_0: \beta_k = 0 \Rightarrow\ H_0: \boldsymbol{C}_k \boldsymbol{W} \boldsymbol{\beta} = 0$$
$$H_0: \boldsymbol{C}_k \boldsymbol{W} \boldsymbol{\beta} = 0 \Rightarrow\ \mathrm{H}_0: \beta_k = 0$$

and

$$\mathrm{H}_0: \beta_k = 0 \Rightarrow\ H_0: \sum_{m=1}^{M_k} \beta_k w_{km} = 0$$
$$H_0: \sum_{m=1}^{M_k} \beta_k w_{km} = 0 \Rightarrow\ \mathrm{H}_0: \beta_k = 0.$$

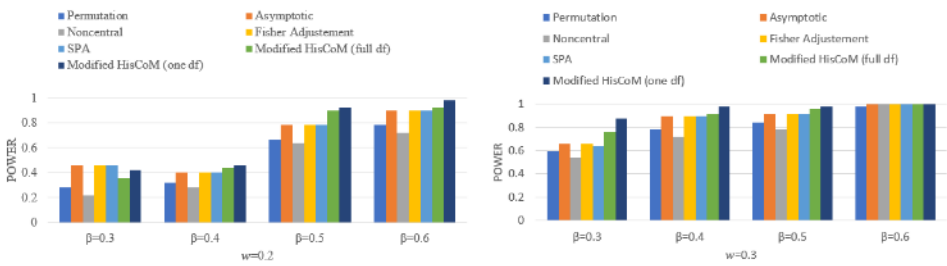
Thus, the hypothesis of the individual pathway effect test and the hypothesis for the modified asymptotic test are equivalent.

Figure 5.9. Empirical type I errors computed from metabolite data



(a) number of data generate − 100 (b) number of data generate − 1000

Figure 5.10. Empirical power from metabolite data



5.7. Conclusion

In summary, we proposed a parametric testing approach to identify the association between pathways and phenotype. The main contribution of this study is to provide a p-value for testing the association between pathway and phenotype with a simple and effective parametric procedure instead permutation approach. This parametric testing approach reduces the computational burden and computational time compared to the permutation test. We use different types of adjustment for the asymptotic test such that non-central test, modified asymptotic test, etc., and then compare their results with the permutation test results. Real data analysis shows that results from the non-centrality test are close to the permutation test results. However, the number of permutations to calculate the non-central parameter was determined empirically. We also perform a simulation study to compare the performance of tests. In simulation study shows that the power of the non-centrality test is comparable with the permutation test. Modified asymptotic test with one df has higher power compared to the all other methods and control type I error well.

Chapter 6. Summary and Conclusion

In this study, we propose a novel pathway-based approach for multinomial phenotypes. To handle this issue, we propose a hierarchical structural component analysis for the multinomial phenotype (HisCoM-Categ) and its extension HisCoM-RCateg for the longitudinal multinomial phenotype. Furthermore, we proposed penalized version of both HisCoM-Categ. All approaches use the biological context of hierarchies among pathways and biomarkers. For the penalized version, we consider three penalty functions i.e., LASSO, SCAD and MCP.

In chapter 3, we propose a novel method, HisCoM-Categ and penalized HisCoM-Categ for the multinomial phenotype to identify the significant pathway. HisCoM-Categ is flexible to use a variety of omics data with both nominal categorical phenotype and ordinal categorical phenotype. In simulation studies, we compare the performances of HisCoM-Categ with the original HisCoM, GSEA, and aSPU. From that comparison, HisCoM-Categ shows better performance than the other three methods. Also, in real data analysis, HisCoM-Categ successfully identified the pathways which are associated with T2D.

In chapter 4, we propose HisCoM-RCateg and its penalized version, an extension of HisCoM-Categ for the longitudinal multinomial phenotype. In application to the real data analysis, HisCoM-RCateg successfully identified the pathways. In simulation studies, we compare the performance of HisCoM-RCateg with HisCoM-GEE. The simulation studies showed that HisCoM-RCateg

has than the HisCoM-GEE approach for multinomial responses and controlled the type I error very well.

In chapter 5, we propose a parametric test for HisCoM to identify the significant pathways. To reduce the computational burden for the original HisCoM, we propose an asymptotic test and then compare their results with permutation test results. We also use many different adjustments of the proposed parametric test. Real data analysis shows that the results of non-centrality adjustment are close to the permutation test. Thus, we believe that the proposed parametric test helps to identifying pathways when fitting any large and high-dimensional data.

Bibliography

1. Vailati-Riboni, M., V. Palombo, and J.J. Loor, *What are omics sciences?*, in *Periparturient diseases of dairy cows*. 2017, Springer. p. 1-7.
2. Raja, G., et al., *Recent advances of microbiome-associated metabolomics profiling in liver disease: Principles, mechanisms, and applications.* International Journal of Molecular Sciences, 2021. **22**(3): p. 1160.
3. Lesk, A.M., *Introduction to genomics*. 2017: Oxford University Press.
4. Oyelade, J., et al., *Overview of the human genome*, in *Genome Plasticity in Health and Disease*. 2020, Elsevier. p. 9-26.
5. Yang, Y., S. Basu, and L. Zhang, *A Bayesian hierarchical variable selection prior for pathway-based GWAS using summary statistics.* Statistics in medicine, 2020. **39**(6): p. 724-739.
6. Morozova, O., M. Hirst, and M.A. Marra, *Applications of new sequencing technologies for transcriptome analysis.* Annual review of genomics and human genetics, 2009. **10**: p. 135-151.
7. Anderson, N.L. and N.G. Anderson, *Proteome and proteomics: new technologies, new concepts, and new words.* Electrophoresis, 1998. **19**(11): p. 1853-1861.
8. Aslam, B., et al., *Proteomics: technologies and their applications.* Journal of chromatographic science, 2017. **55**(2): p. 182-196.
9. Verrills, N.M., *Clinical proteomics: present and future prospects.* Clinical Biochemist Reviews, 2006. **27**(2): p. 99.
10. Clish, C.B., *Metabolomics: an emerging but powerful tool for precision medicine.* Molecular Case Studies, 2015. **1**(1): p. a000588.
11. Oliver, S.G., et al., *Systematic functional analysis of the yeast genome.* Trends in biotechnology, 1998. **16**(9): p. 373-378.
12. Krassowski, M., et al., *State of the field in multi-omics research: From computational needs to data mining and sharing*. Frontiers in Genetics, 2020. **11**: p. 610798.
13. Liu, Y. and M.R. Chance, *Pathway analyses and understanding disease associations.* Current genetic medicine reports, 2013. **1**(4): p. 230-238.

14. Sham, P.C. and S.M. Purcell, *Statistical power and significance testing in large-scale genetic studies.* Nature Reviews Genetics, 2014. **15**(5): p. 335-346.
15. Wang, K., M. Li, and H. Hakonarson, *Analysing biological pathways in genome-wide association studies.* Nature Reviews Genetics, 2010. **11**(12): p. 843-854.
16. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences, 2005. **102**(43): p. 15545-15550.
17. Wang, K., M. Li, and M. Bucan, *Pathway-based approaches for analysis of genomewide association studies.* The American Journal of Human Genetics, 2007. **81**(6): p. 1278-1283.
18. Xia, J. and D.S. Wishart, *MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data.* Nucleic acids research, 2010. **38**(suppl_2): p. W71-W77.
19. Pan, W., et al., *A powerful and adaptive association test for rare variants.* Genetics, 2014. **197**(4): p. 1081-1095.
20. Pan, W., I.-Y. Kwak, and P. Wei, *A powerful pathway-based adaptive test for genetic association with common or rare variants.* The American Journal of Human Genetics, 2015. **97**(1): p. 86-98.
21. Kim, J., W. Pan, and A.s.D.N. Initiative, *Adaptive testing for multiple traits in a proportional odds model with applications to detect SNP-brain network associations.* Genetic epidemiology, 2017. **41**(3): p. 259-277.
22. Lee, S., et al., *Pathway-based approach using hierarchical components of collapsed rare variants.* Bioinformatics, 2016. **32**(17): p. i586-i594.
23. Kim, Y., et al., *Hierarchical structural component modeling of microRNA-mRNA integration analysis.* BMC bioinformatics, 2018. **19**(4): p. 25-34.
24. Mok, L., et al., *HisCoM-PAGE: hierarchical structural component models for pathway analysis of gene expression data.* Genes, 2019. **10**(11): p. 931.
25. Kim, Y., et al., *Identifying miRNA-mRNA Integration Set Associated With Survival Time.* Frontiers in Genetics, 2021. **12**: p. 634922.
26. Park, C., B. Kim, and T. Park, *DeepHisCoM: deep learning pathway analysis using hierarchical structural component models.* Briefings in Bioinformatics, 2022.

27. Hwangbo, S., et al., *Kernel-based hierarchical structural component models for pathway analysis.* Bioinformatics, 2022. **38**(11): p. 3078-3086.
28. Lee, S., et al., *Pathway-based approach using hierarchical components of rare variants to analyze multiple phenotypes.* BMC bioinformatics, 2018. **19**: p. 85-97.
29. Lee, S., et al., *Pathway analysis of rare variants for the clustered phenotypes by using hierarchical structured components analysis.* BMC medical genomics, 2019. **12**(5): p. 1-9.
30. Fahrmeir, L., et al., *Multivariate statistical modelling based on generalized linear models*. Vol. 425. 1994: Springer.
31. Liang, K.-Y. and S.L. Zeger, *Longitudinal data analysis using generalized linear models.* Biometrika, 1986. **73**(1): p. 13-22.
32. Lipsitz, S.R., K. Kim, and L. Zhao, *Analysis of repeated categorical data using generalized estimating equations.* Statistics in medicine, 1994. **13**(11): p. 1149-1163.
33. McCullagh, P., *Regression models for ordinal data.* Journal of the Royal Statistical Society: Series B (Methodological), 1980. **42**(2): p. 109-127.
34. Ouyang, Y., et al., *Metabolome-Genome-Wide Association Study (mGWAS) Reveals Novel Metabolites Associated with Future Type 2 Diabetes Risk and Susceptibility Loci in a Case-Control Study in a Chinese Prospective Cohort.* Global Challenges, 2021. **5**(4): p. 2000088.
35. Takane, Y. and H. Hwang, *An extended redundancy analysis and its applications to two practical examples.* Computational statistics & data analysis, 2005. **49**(3): p. 785-808.
36. Hoerl, A.E. and R.W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems.* Technometrics, 1970. **12**(1): p. 55-67.
37. Fan, J. and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties.* Journal of the American statistical Association, 2001. **96**(456): p. 1348-1360.
38. Tibshirani, R., *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society: Series B (Methodological), 1996. **58**(1): p. 267-288.
39. Zhang, C.-H., *Nearly unbiased variable selection under minimax concave penalty.* 2010.
40. Kim, Y., B.-G. Han, and K. Group, *Cohort profile: the Korean genome and epidemiology study (KoGES) consortium.*

International journal of epidemiology, 2017. **46**(2): p. e20-e20.
41. Fan, S., et al., *Systematic error removal using random forest for normalizing large-scale untargeted lipidomics data.* Analytical chemistry, 2019. **91**(5): p. 3590-3596.
42. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal statistical society: series B (Methodological), 1995. **57**(1): p. 289-300.
43. Jung, T., et al., *Integrative Pathway Analysis of SNP and Metabolite Data Using a Hierarchical Structural Component Model.* Frontiers in genetics, 2022. **13**.
44. Hand, D.J. and R.J. Till, *A simple generalisation of the area under the ROC curve for multiple class classification problems.* Machine learning, 2001. **45**: p. 171-186.
45. Touloumis, A., *Simulating Correlated Binary and Multinomial Responses under Marginal Model Specification: The SimCorMultRes Package.* R J., 2016. **8**(2): p. 79.
46. Cario, M.C. and B.L. Nelson, *Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix*. 1997, Technical Report, Department of Industrial Engineering and Management ….
47. Fisher, R., *Statistical Methods for Research Workers. 4th edn Oliver and Boyd: London.* 1932.

초록

그동안 특정 질병과 관련된 마커로부터 새로운 경로를 식별하기 위해 경로 분석의 여러 통계적 방법이 적용되어 왔습니다. 하지만 사용 가능한 대부분의 방법은 단일 경로 분석을 기반으로 하며 여러 경로를 동시에 고려하지 않습니다. 경로는 높은 상관 관계가 있기 때문에 다중 경로 분석은 이러한 상관 관계의 문제를 겪습니다. 이 문제를 해결하기 위해 HisCoM(계층적 구조 구성 요소 모델)이 개발되었습니다. 이 모델은 모든 경로와 경로 간의 상관 관계를 동시에 고려했습니다. HisCoM 은 연속형, 이산형 및 이진형 데이터 분석에 성공적으로 적용되었지만 다항 표현형 분석에는 쉽게 적용할 수 없습니다.

본 논문에서는 다항 표현형에 대한 계층적 구조 성분 분석(HisCoM-Categ)과 다항 표현형 종단 데이터에 대한 계층적 구조 성분 분석(HisCoM-RCateg)이라는 새로운 통계 방법을 제안한다. 또한 HisCoM 이 경로와 표현형 간의 연관성을 찾기 위해 순열 접근 방식이 아닌 모수적 가설검정 접근 방식을 제안한다.

HisCoM-Categ 는 기존 HisCoM 과 마찬가지로 바이오마커와 경로 계층을 고려하면서 릿지 페널티를 사용하여 모든 경로의 상관관계를 고려합니다. 경로와 표현형 사이의 연관성을 확인하기 위해 HisCoM-Categ 는 명목상 표현형에 대한 기본 범주 로짓 모델과 서수 표현형에 대한 비례 확률 모델을 사용합니다. HisCoM-RCateg 는 세로 다항 표현형을 위한 HisCoM-Categ 의 확장 버전입니다. HisCoM-Categ 와 마찬가지로 HisCoM-RCateg 도 동시에 모든 경로를 분석하여 원하는 표현형과 관련된 중요한 경로를 식별할 수 있습니다. HisCoM-Categ 및 HisCoM-RCateg 는 모두 다양한 유형의 오믹스 데이터에 사용할 수 있을 만큼 유연합니다. 예를 들어, 우리는 Korean Association Resource (KARE)의 실제 대사체 데이터 세트에서 HisCoM-Categ 및 HisCoM-RCateg 방법을 사용하여 대사 경로와 제 2 형 당뇨병(T2D) 사이의 연관성을 확인했습니다. T2D 는 여러

유전적 요인에 의해 영향을 받는 대사성 질환이며, 이는 주요 공중 보건 문제입니다. KARE 대사산물 데이터 세트에 대한 적용은 HisCoM-Categ 및 HisCoM-RCateg 가 T2D 와 관련된 경로를 식별할 수 있음을 보여줍니다. 또한 시뮬레이션 연구를 통해 HisCoM-Categ 및 HisCoM-RCateg 가 다른 방법보다 더 나은 성능을 보인다는 것을 보여줍니다.

주요어 : 패스웨이 분석, 계층적 구조, 종적 데이터, 다항 표현형, 모수적 가설검정

학번 : 2018-34194