이 학 석 사 학 위 논 문

Parametric Testing for Hierarchical Structural Component Model

계층적 구조 모형의 모수적 검정

2020년 2월

서울대학교 대학원
통계학과
정 석 호

Parametric Testing for Hierarchical Structural Component Model

by

Seokho Jeong

A thesis
submitted in fulfillment of the requirement
for the degree of Master in
Statistics

Department of Statistics
College of Natural Sciences
Seoul National University
February 2020

<제목>

지도교수 박 태 성

이 논문을 이학석사 학위논문으로 제출함

2019년 12월

서울대학교 대학원
통계학과

정 석 호

정석호의 이학석사 학위논문을 인준함
2019년 12월

위원장 _____(인)
부위원장 _____(인)
위 원 _____(인)

Abstract

Parametric Testing for Hierarchical Structural Component Model

Seokho Jeong

Department of statistics

The Graduate School

Seoul National University

Hierarchical Structural Component models (HisCoM) has added a hierarchical structure to the generalized linear models to represent hierarchical structure of biological data. Extension of this method has been successfully extended to fit the variety of biological data types, from clinical data to microbiome data. HisCoM and its variations utilized permutation test to discover association in individual feature and higher structures. Along with the advent of high-throughput technologies in genomics, genomic

datasets from massive nationwide cohort such as UK biobank have posed computational challenges. In this study, parametric test for feature individual effect and latent level effect which does not require permutation for faster analysis. Simulation study presented parametric test showed high power and fast speed under right distribution assumptions. Application of the test on real data analysis using RNA expression and 16S rRNA microbiome data from Cancer patients were also presented.

Keyword: Hierarchical components, Parametric testing, Pathway analysis, Human microbiome

Student Number: 2018-24959

Contents

List of tables

List of figures

Chapter 1.

Introduction

With the advent of high-throughput technologies such as next-generation sequencing(NGS), new types of biological data have emerged and have been widely used for biological and clinical analysis. Not only these data have data-type oriented properties, but also the analysis of such data depends on the pre-existing biological information. For example, RNA expression data and microbiome data share characteristics such as non-negativity and sparsity because both use NGS when producing. On the other hand, because the sources of RNA expression and the microbiome is different, prior biological knowledge differs. Nevertheless, such biological information is often given in a hierarchical manner. For example, phylogenic trees of microbiome species express the hierarchical relationship of species. Pathway in the human mRNA gene and protein

expression also can be interpreted as a hierarchical structure of the biological process. Therefore, models assuming a prior hierarchical structure can be informative and interpretable.

Models to illustrate and associate hierarchical structure of biological data has been proposed often motivated from a specific biological data. For example, SNP-set kernel Association Test (SKAT) has been developed to collectively test SNP-set into a sequence [1]. This method made expansion to the Microbiome Regression-Based Kernel Association Test (MiRKAT), which performs association tests on using the phylogenic relation [2]. The Gene Set Enrichment Analysis (GSEA) is also widely used to explain the hierarchical structure of biological pathways [3].

As one of such efforts, Pathway-based approach using HierArchical structure of collapsed Rare variant Of High-throughput sequencing data (PHARAOH) has been developed for the pathway analysis of the rare variants in human gene sequence [4]. It is one of the first statistical approaches that take multiple pathways into account. The method is based on the generalized structured component analysis (GSCA) and is originally designed to a single and

uncorrelated phenotype with a general exponential family distribution (response variable); extensions to clustered phenotypes (PHARAOH-GEE) and multiple phenotypes (PHARAOH-multi) have been also developed [5] [6] [7].

The hierarchical structured component model (HisCoM) has been proposed to suit broader types of biological data. Such attempts include the integration of the miRNA and mRNA expressions and connection to the phenotype (HisCoM-Mimi), an extension of HisCoM structure using Cox Proportional Regression for the prognosis of patients (HisCoM-PAGE), and an application to drug response prediction model from MRM-MS (multiple reaction monitoring mass spectrometry) data [8] [9] [10]. HisCoM-GGI takes SNP (Single Nucleotide Polymorphism) interaction term into gene analysis [11].

For the testing significance of individual pathway or gene effect, the permutation test is presented in the original HisCoM model. The permutation test is useful because it generates an empirical null distribution of both pathway and gene effect. However, compared with the permutation test, parametric tests for HisCoM has not been well developed. The parametric test of HisCoM can be useful because the computation burden and calculation time might pose a

problem when fitting large and high dimensional data when performing permutation tests.

In this study, revisiting and interpretation of HisCoM are given. Under HisCoM without regularization, two parametric tests for significance are introduced. Wald type test using the sandwich form of individual effect covariance is first introduced. Also, we will show that the distribution of each pathway effect is expressed as a quadratic form of multivariate Gaussian variables. Using this property, the distribution is given as a weighted mixture of independent chi-square distributions. Detailed simulation study and real data analysis are also performed and presented.

Chapter 2.

Methodology

2.1 The PHARAOH Method

For the discussion of further study, the revisit of the methodology in PHARAOH is made here. One of the distinctive features of PHARAOH is that a genetic pathway, a collection of molecular regulators that interact with each other and with other substances in the cell to govern the gene expression levels of mRNA and proteins, is considered for association analysis. Input of the PHARAOH consist of rare genetic variants on each sample. One or more variants are included in each gene, and each gene is assigned to one or more pathways. To handle the rare variants, the PHARAOH first uses collapsing technique. Let g_{ij} be the genotype of t^{th}genetic variant of j^{th} gene by the

number of minor alleles. Collapsed genotype of j^{th} gene can be generated using weighted sum of these genotypes included in j^{th} gene as $x_j = \sum_t \omega_t g_{tj}$, where ω_t is a pre-defined weight such as Minor Allele Frequency(MAF).

After integrating rare variants into collapsed genotypes, genotypes consist pathway scores as their weighted sums. Since pathway scores is fitted to simple GLM, different types of phenotypes such as count, binary, categorical phenotypes can be considered. In other words, coefficient connecting the pathway score linearly predicts the parameter for an exponential family, via link function. To illustrate this, first define linear predictor of i^{th} sample η_i, then the presumed distribution becomes as equation (1),

$$p(y_i; \mu_i, \phi) = \exp(y_i \mu_i - \frac{a(\mu_i)}{b(\phi)} + c(\mu_i, \phi)),\ \eta_i = \gamma(\mu_i) \quad (1)$$

where γ is a link function. To explain η_i, suppose k^{th} pathway and its respective gene set G_k of an i^{th} sample, then k^{th} pathway score becomes weighted sum of x_j's in k^{th} pathway, i.e. $f_{ik} = \sum_{j \in G_k} w_{jk} x_{ijk}, x_{ijk} = \sum_t \omega_t g_{itj}$. Lastly, linear predictor is expressed as $\eta_i = \sum_k \beta_k f_{ik} = \sum_k \sum_{j \in G_k} \beta_k w_{jk} x_{ijk}$. Thus, the objective parameters for estimation are weights on gene and pathway coefficients. By adopting double regularization by the ridge penalty,

objective function using penalized likelihood becomes as equation (2),

$$l(W,\beta|y) = \sum_{i=1}^{N}(y_i\mu_i - \frac{a(\mu_i)}{b(\phi)} + c(\mu_i,\phi)) - \frac{1}{2}\lambda_G \sum_{k=1}^{K}\sum_{j\in G_k} w_{jk}^2 - \frac{1}{2}\lambda_p \sum_{k=1}^{K}\beta_k^2 \quad (2)$$

where W is the weights on each gene, and β is all the pathway coefficients. For fair comparison of effect size among pathways and assuring of identifiable model, constraint $\sum_{i=1}^{N} f_{ik}^2 = N$ is given as scaling constraint [2].

The PHARAOH method adopts ALS method to estimate the model parameters. On each iteration, linear predictor has two formulations (3) and (4)

$$\eta_i = \sum_k \beta_k(\sum_{j\in G_k} w_{jk}x_{ijk}) \quad (3)$$

$$\eta_i = \sum_k \sum_{j\in G_k} w_{jk}(\sum_{j\in G_k} \beta_k x_{ijk}) \quad (4)$$

Thus, weights w_j's are estimated using (4), on fixed β_k's and β_k's are estimated on fixed w_{jk}'s. On each step, Iteratively Reweighted Least Square (IRLS) with penalty is utilized to estimate the target parameters.

2.2 Sandwich Variance Estimate of $W\beta$

Focusing on the estimation of weights on collapsed genotypes and pathway effect in PHARAOH, more

generalized biological pathways can be explained as presented in other studies in HisCoM [8] [9] [10]. Let us suppose we have already known biological pathway structures between covariates when building regression models. Properly estimated weights on individual gene can represent latent variables as quantity of each pathway. If covariates are gene quantifications, the latent variable becomes pathway quantity. If covariates are microbiome counts of species level, latent variable can represent higher phylogenic structure such as family and genus. As shown in Figure 1, the general structure of HisCoM consists of individual covariates and latent variable indicating pathway information. In this study, simple formulation of HisCoM, where no penalties are taken into account and no multiple phenotypes. If independent variables of N samples have K latent structures in the data matrix, the total number of covariates can be expressed as the summation of number of independent variables in each latent structure (T_k). Let us state corresponding weights and pathway effects as w_{kt}, β_k.

X can be expressed as collection of pathway

$\mathrm{X} = X_{N\times\sum_{k=1}^{K} T_k} = \left(X_{N\times T_1}, X_{N\times T_1}, \dots, X_{N\times T_K}\right)$, let $\mathrm{P} = \sum_{k=1}^{K} T_k$

Let the weight matrix and pathway effect as

$$W_{P\times K} = (w_{jk}) = \begin{cases} w_{ij} & \text{if i th variable is included in the k th latent variable} \\ 0 & \text{o.w.} \end{cases},$$

$$\beta_{K\times 1} = (\beta_1, \beta_2, \dots, \beta_K)^T$$

The log likelihood of HisCoM model for exponential family is as shown

$$l(W,\beta) = \sum_{i=1}^{N} \log p(y_i;\mu_i,\phi)\ \exp(y_i\mu_i - \frac{a(\mu_i)}{b(\phi)} + c(\mu_i,\phi)) \quad (5)$$

where link function $\gamma(\mu_i) = \eta_i = X_i W\beta$. Objective function becomes $l(W,\beta)$, or equivalently $l(W\beta)$. By maximizing objective function, optimal weight and pathway effect is estimated. To make the pathway effect β consistent with the magnitude of estimated latent variable, and $\sum_{i=1}^{N} f_{ik}^2 = \sum_{i=1}^{N}(\sum_{j=1}^{T_k} w_{jk}X_{ijk})^2 f_{ik}^2 = N \ for\ \forall k$ is still maintained by scaling of weights.

For example, if y is normally distributed, the objective function becomes

$$\underset{W,\beta}{\operatorname{argmin}}\ (y - XW\beta)^t(y - XW\beta)\ ,\ w.r.t.\ diag((XW)^tXW) = NI_{(K,K)} \quad (6)$$

Using the ALS algorithm, we can get the estimates of W and β.

Let $l^1 := \frac{\partial l}{\partial(W\beta)}, l^2 := \frac{\partial^2 l}{\partial(W\beta)^T\partial(W\beta)}$, then by property of maximum likelihood estimator [12]

$$\frac{1}{\sqrt{n}}(\widehat{W\beta} - W\beta) \to N(0, (-l_n^2)^{-1}\ \sum(l_n^1(W\beta))^2\ (-l_n^2)^{-1} \text{ as } n->\infty \quad (7)$$

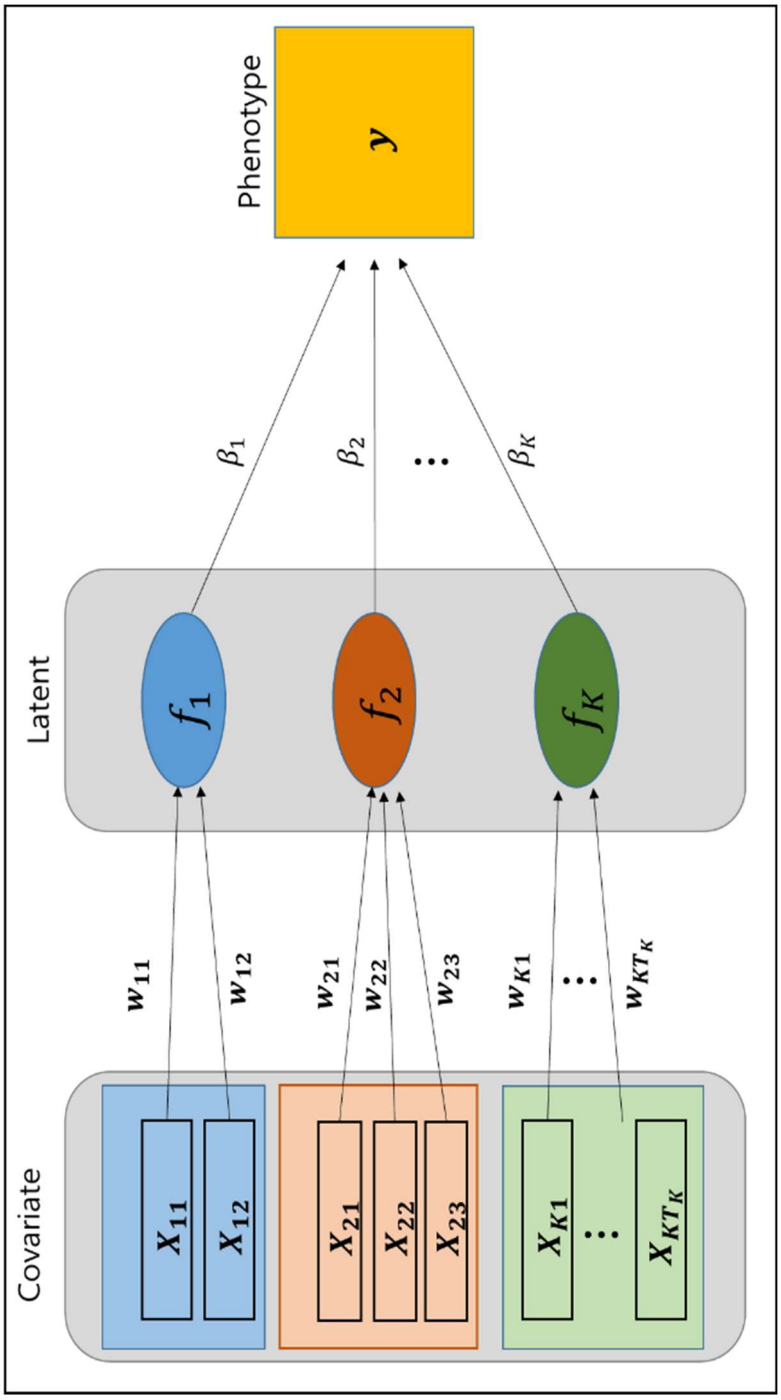


Figure 1 General HisCoM model structure diagram

2.3 Test statistics for the pathway effect

Consider testing the kth pathway effect with the covariates $X_k \coloneqq X_{N\times T_k}$ under the null hypothesis. From (2.2.1), the weights and effect on each independent variable included in kth pathway becomes as follows.

$$w_k = (w_{1k}, \dots, w_{T_k k})^T, w_k\beta_k = (w_{1k}\beta_k, \dots, w_{T_k k}\beta_k)^T \quad (8)$$

$$\widehat{W\beta_k} \sim N\left(0, \Sigma_{T_k,T_k}\right), for\ \ (X_k w_k)^t X_k w_k = N \quad (9)$$

$$\Sigma_{T_k,T_k} = U_{T_k,r} D_{r,r} U_{T_k}^T = B_k {B_k}^T \quad (10)$$

The covariance matrix of (2.3.2) is a partial matrix of (2.2.1) which can be decomposed by Cholesky-like form using singular value decomposition. Using the scaling constraint after samplewise standardization of latent variables, $N\widehat{\beta_k}^2$ can be expressed as a quadratic form of Gaussian random variables.

$$N\widehat{\beta_k}^2 = (\widehat{W\beta_k})^T {X_k}^T X_k \left(\widehat{W\beta_k}\right) = Z^T {B_k}^T {X_k}^T X_k B_k Z,$$

$$where B_k Z = \widehat{W\beta_k}$$

$$B_k{}^T X_k{}^T X_k B_k = \ P^T \Lambda P \ where\ P\ is\ orthogonal\ and\ \Lambda\ eigenvalues$$

$$\Lambda = diag(\lambda_1, \lambda_2, \dots, \lambda_r)\ \ with\ \ rank\ \ r$$

Then distribution of $N\widehat{\beta_k}^2$ can be formulated as the squares of normal variables, or equivalently mixture of chi-squared variables [13].

$$N\widehat{\beta_k}^2 = \sum_{j=1}^{r} \lambda_j (U_j)^2 \quad (11)$$

$$U = (U_1, U_2, \ldots, U_r)^T \sim N(0, I_{r,r}) \quad (12)$$

The approximated p-value under the null hypothesis is then calculated via the survival function of chi-square mixtures using eigenvalues via R package "CompQuadForm" [14].

Chapter 3.

Results

3.1 Simulation Study

To evaluate the performance of the parametric test in HisCoM a multiple linear regression with the independent normal error was assumed. 200 samples with 100 replicates were generated to compute the power and type I

error. Causal pathways and individual covariates were selected. Different parameters such as effect size, number of pathway variables, number of variables in a pathway were applied to check the tendency of the result. Table 1 summarizes the hyperparameters chosen for the simulation study. The accuracy of estimated variance for individual effects was checked by comparing variance estimates and theoretical variances. Then, the null distribution of the latent estimates was generated using estimated eigenvalues and compared with the theoretical distribution. The comparison was made using a density plot of effects and QQ plots. Finally, Type-I error and Power comparison was made between the proposed test and permutation test results.

As shown in figure 2, the variance estimates and theoretical variance were compared on each effect. It showed high accuracy depending on different effect sizes. Figure 3 shows the quantile-quantile plot of theoretical null distribution and the estimated value of the noncausal pathway and causal(significant) pathway. It shows the null distribution is well kept under the noncausal pathway.

Figure 4 shows the overall power of the parametric test was higher than the power of the permutation test with

controlled type-1 error all over the different effect sizes. The q-q plots showed that the effect sizes of latent variables are well approximated by the chi-square mixture method. However, there was a tendency; as the effect increases, the p-value of the permutation test also increases compared to that of the parametric test. The number of pathway and number of genes in a pathway did not show any tendency as seen in supplementary figures.

Table 1 Hyperparameter setting for Simulation Study (Variance is set to 1)

Hyperparamter Type	Value	Other Specification
Effect size	0.01,0.032,0.1,0.2 0.3,0.4,0.5,12	# of Pathway = 10 Genes in a pathway = 20
Number of Pathway	3, 5, 10	Effect size = 0.4 Genes in a pathway = 10
# of genes in each Pathway	5, 10, 20	Effect size = 0.4 # of Pathway = 4

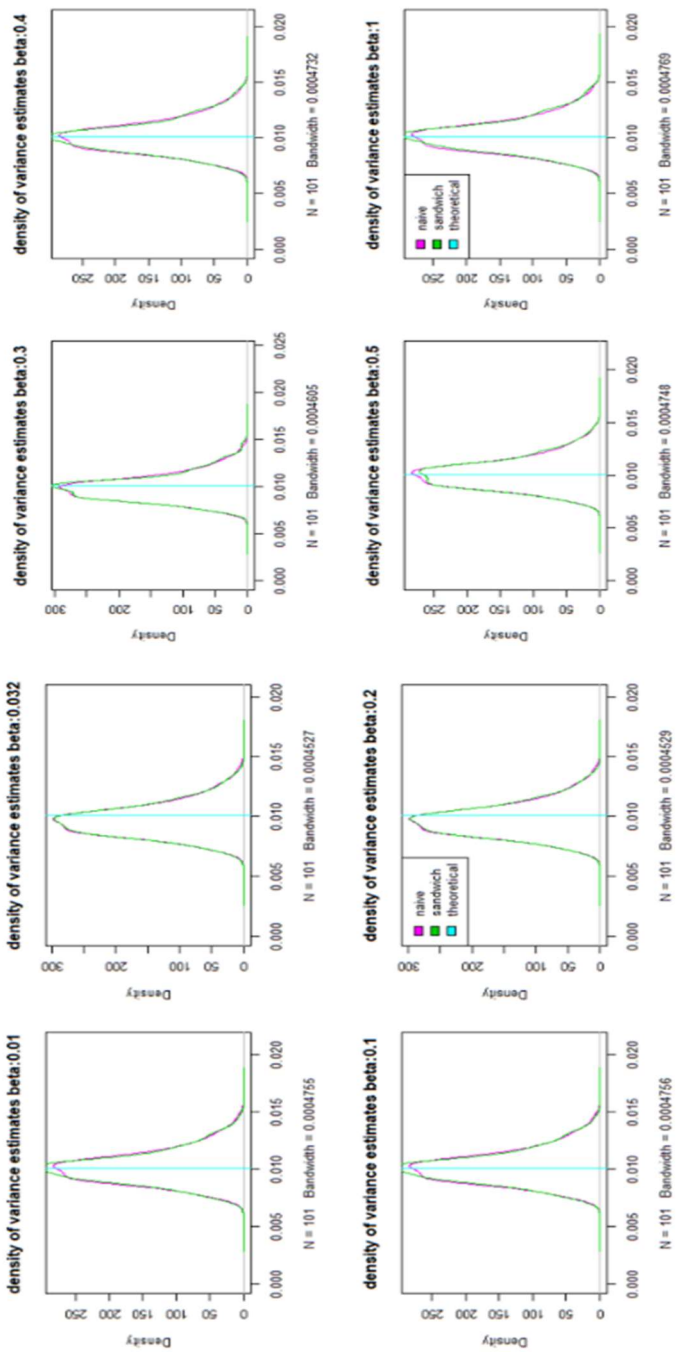


Figure 2 Variance estimate Comparison on different effect size

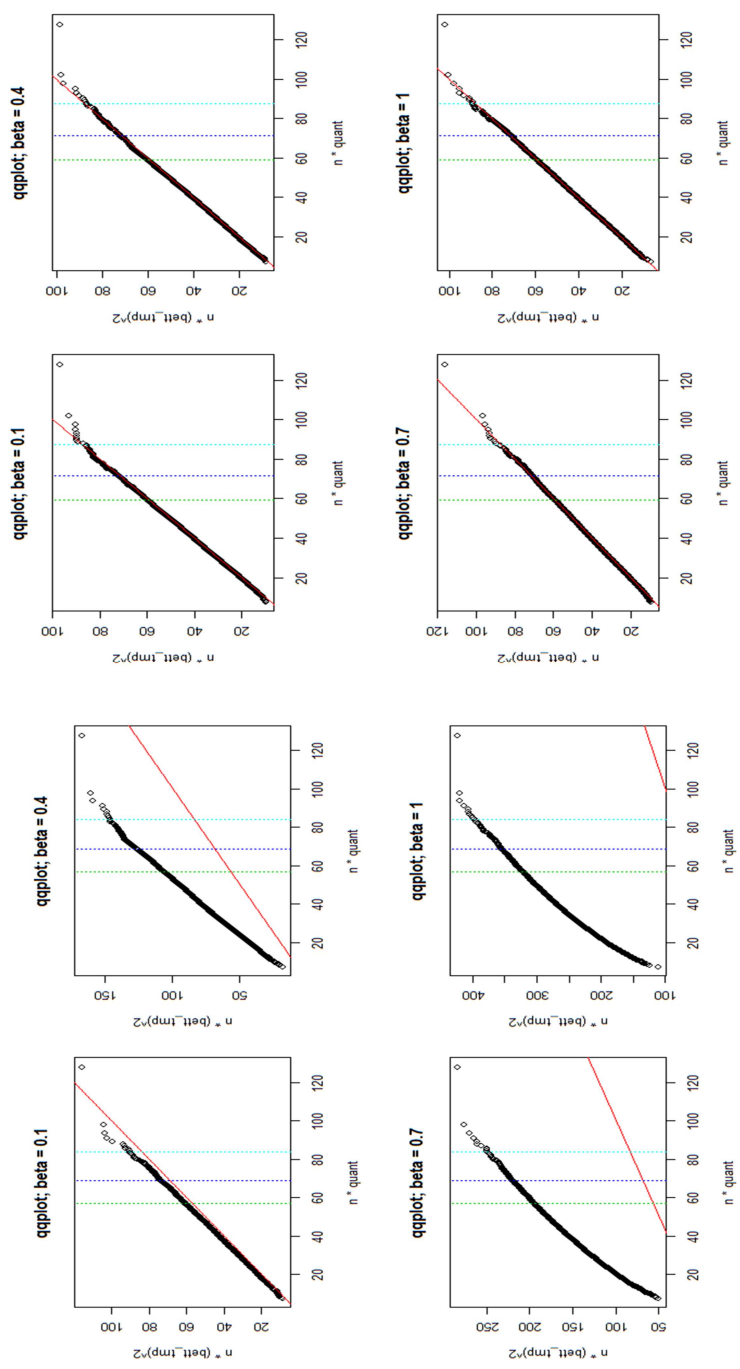


Figure 3 (Left) qq plot for Causal Pathway (Right) qqplot for noncausal Pathway

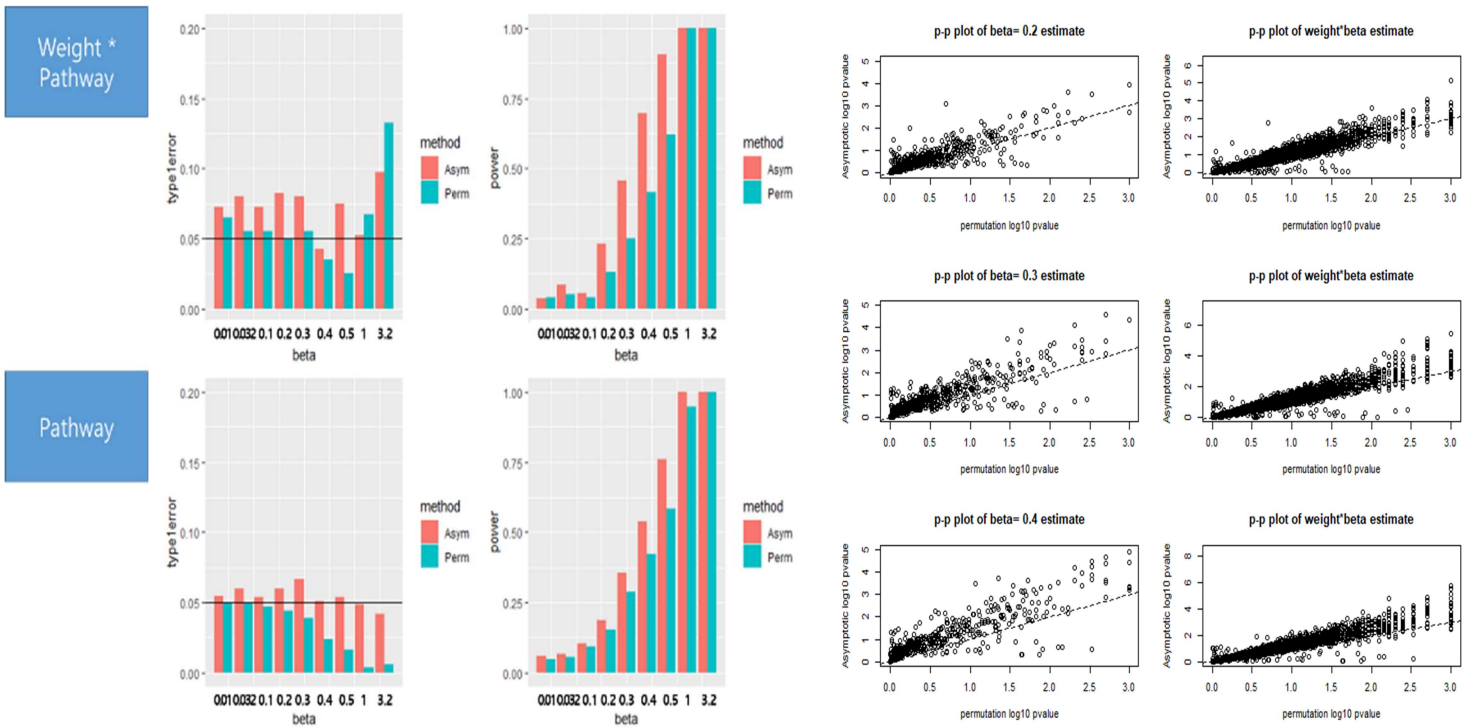


Figure 4 (left) Type I error and power of parametric test and permutation test

3.2 Real Data Analysis

Colorectal Cancer data

Colorectal Cancer(CRC) data by Baxter et al. were used to for real data analysis on binary phenotype which consists 120 CRC patients and 172 normal control group [15]. Associated microbiota consists of 335 OTUs with 33 identified families. One of the family, *Fusobacteriaceae* has been known to be related with the development of CRC. It is reported to enhance the upregulation of CARD3 expression, leading to the activation of the autophagy signaling [16]. From such prior information, other families were fitted to the regression model to discover additional association with controlling the effect of *Fusobacteriaceae.* Table 2 shows the significant families associated with the cancer status and controlled p-value of each family including *Fusobacteriaceae* ($\alpha = 0.01$). The result shows that the significant families from asymptotic test includes the significant families from the permutation test (5000 permutations). It is notable to see the *Fusobacteriaceae* family was not significant when fitted with only itself ($p = 0.17$). Figure 5 shows the log10 pvalue-pvalue plot of genus level effect and family(pathway) effect.

Table 2 Association significant results of CRC patients' microbiome

Family Name	Size	Perm. p value	Asym. p value
Acidaminococcaceae	5	0.000006	-
Actinomycetaceae	1	0.001148	-
Clostridiales_ Incertae_Sedis_XI	1	0.004657	-
Clostridiales_ Incertae_Sedis_XIII	1	0.003768	-
Enterobacteriaceae	2	0.009365	-
Micrococcaceae	1	0.006853	-
Ruminococcaceae	48	0.000002	-
Lachnospiraceae	53	-	0.00160
Erysipelotrichaceae	8	<10e-16	0.00040
Oxalobacteraceae	1	0.002956	0.00760
Peptostreptococcaceae	2	0.003244	0.00600
Streptococcaceae	2	<10e-10	0.20626

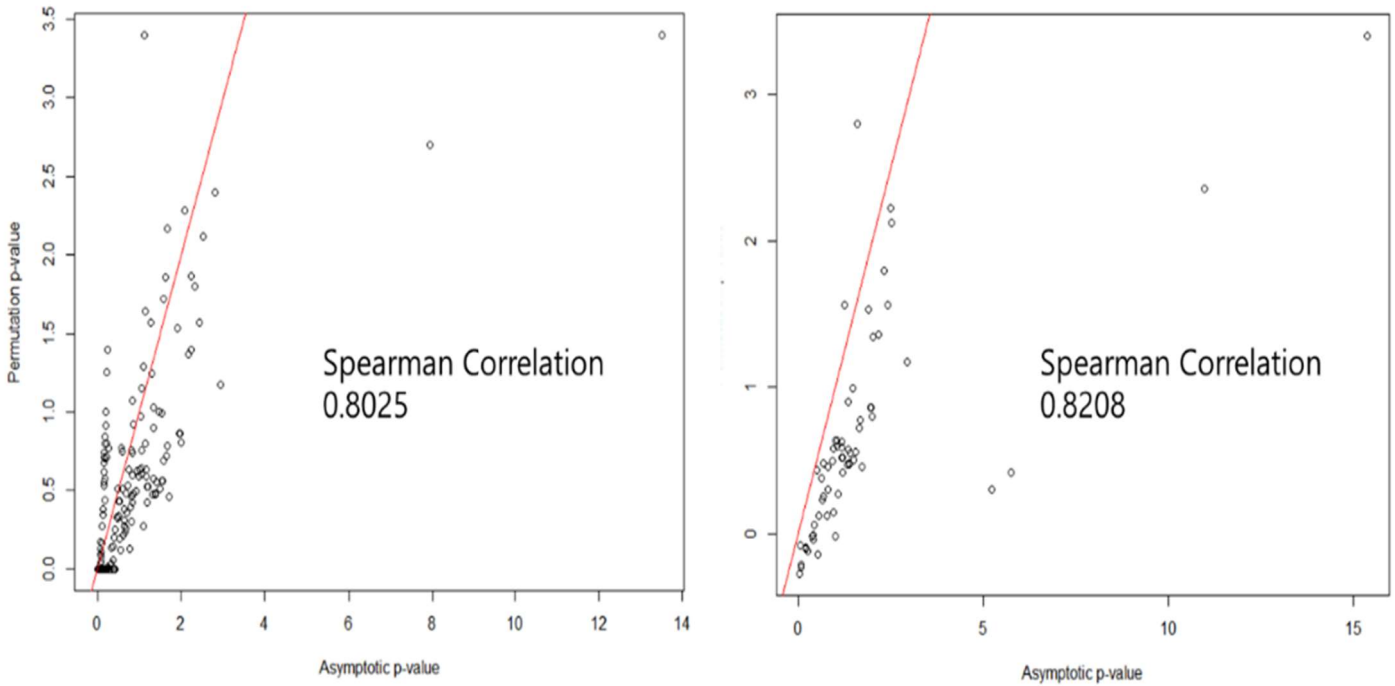


Figure 5 log10 pvalue-pvalue plot of OTU individual and family level effect

The Cancer Genome Atlas PDAC data

From The Cancer Genome Atlas(TCGA), RNA-seq gene expression data of 149 pancreatic adenocarcinoma(PDAC) patients were used [17]. Tumor cellularity is known to be closely related to subtypes of tumors, therefore tumor cellularity of each sample was chosen as the response variable. The top expressed 5000 genes were mapped into pathways using KEGG, with 185 pathways were identified. One of the main objectives of this pathway analysis is to identify pathways that are associated with time to death (overall survival). The test rests of HisCoM are summarized in Table 3 showing the significant pathways associated with the survival phenotypes and tumor cellularity from the permutation test and permutation test. The result shows that many significant pathways are similar using both tests. Pathways that showed different results sometimes.

Table 3 Associated pathway results of survival groups

Pathway	Size	Perm p-val.	Asym. p value
KEGG_TAURINE_AND_ HYPOTAURINE_METABOLISM	2	0.0016	0.043
KEGG_PROPANOATE_METABOLISM	12	0.038	0.066
KEGG_ONE_CARBON_POOL_BY_ FOLATE	5	0.0355	0.042
KEGG_HOMOLOGOUS_RECOMBINATION	7	0.0305	0.281
KEGG_RIG_I_LIKE_RECEPTOR_SIGNALING_PATHWY	15	0.0235	0.009

Chapter 4.

Discussions

Conclusion

In this study, a parametric test on the pathway and individual effect of Hierarchical structured component model was proposed. Implementation of the proposed test on general distributions by real data analysis on both continuous and binary phenotypes was made. On real data analysis of TCGA PDAC data, both simulation studies and real data analysis showed that the proposed parametric test yielded high efficiency in computation time compared to the previous permutation test. Asymptotic test on HisCoM with binary phenotype was also performed on the CRC microbiome data. Comparison between asymptotic

test and permutation test showed similarity when the significance is relatively small. Also, under the right distribution assumption, the power of the new proposed test performed well. The proposed method can be effective especially for the data that consist of many features or samples. Test procedures for extended HisCoMs are suggested correspondingly.

Further Extensions to other HisCoM models

HisCoM-PAGE is an implementation of Cox-regression model onto HisCoM [9][12]. The proportional hazard model of HisCoM-PAGE and its corresponding partial likelihood is given as (13) and (14)

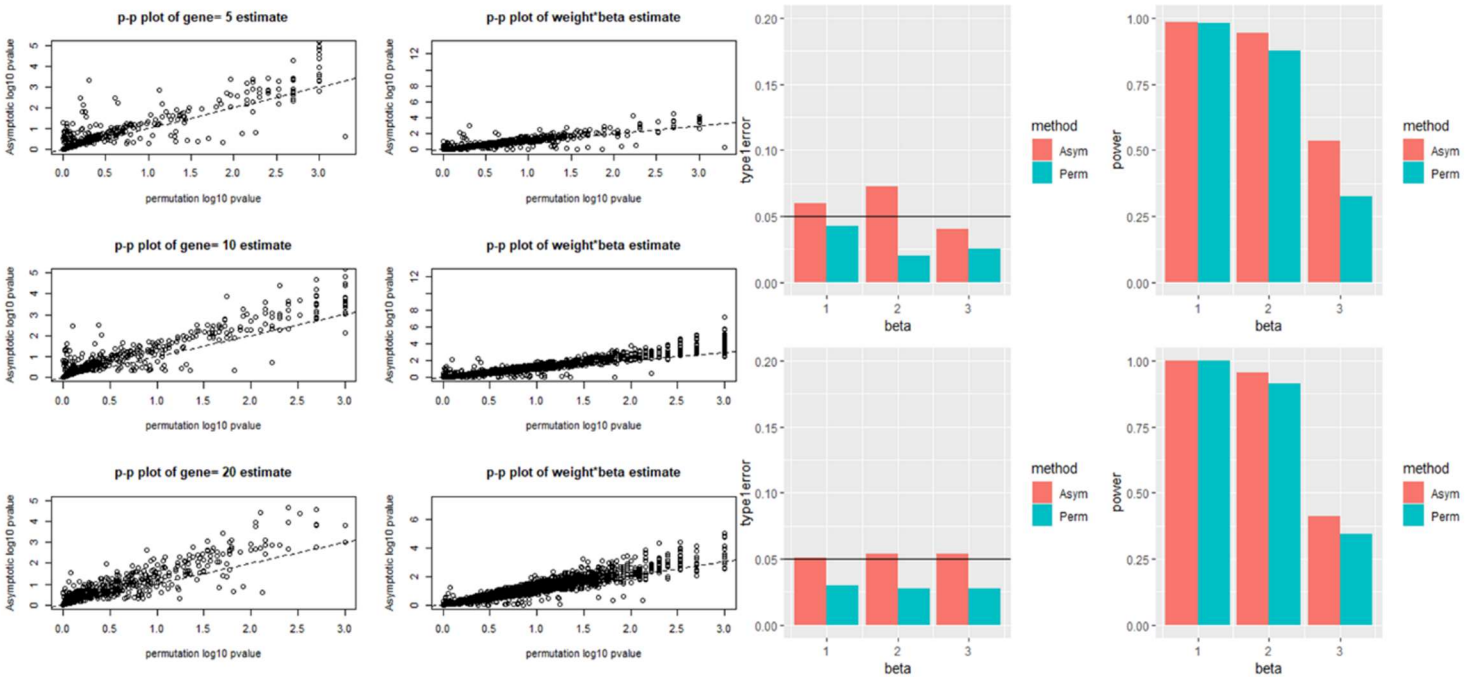
$$h(y_i|F_i) = h_0(y_i)exp\sum_{k=1}^{K}\left(\sum_{j\epsilon G_k} x_{ijk}w_{jk}\right)\beta_k \qquad (13)$$

$$\phi = \sum_{i:C_i=1}[\sum_{k=1}^{K} f_{ik}\,\beta_k - log\sum_{l\epsilon R(y_i)}\exp(\sum_{k=1}^{K} f_{ik}\beta_k)] \qquad (14)$$

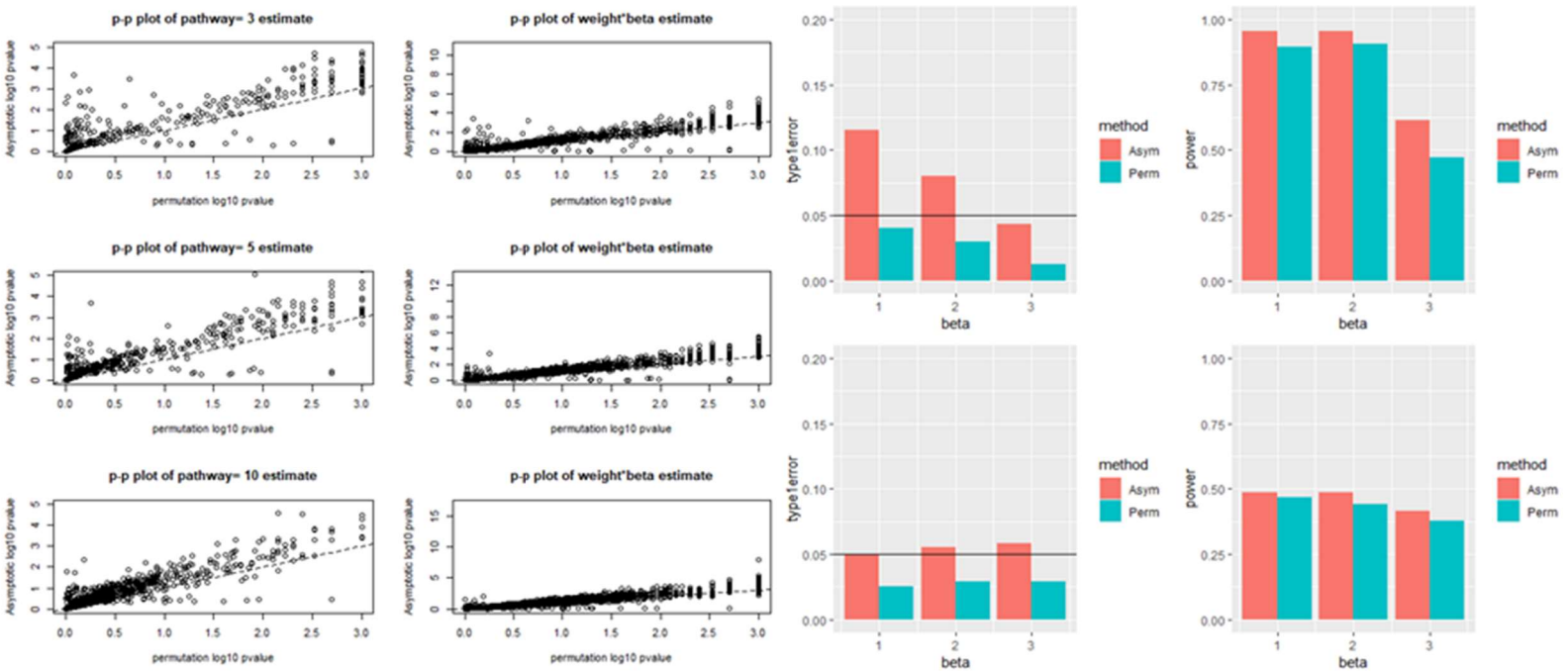
where $F_i = (f_{i1}, f_{i2}, \ldots, f_{i1K})^t$ are the pathway values of an i^{th} sample, C being censored status and $R(y_i)$ being the risk set at y_i. Using the asymptotic property of partial likelihood estimator, the covariance of linear estimators can be approximated using the information matrix derived from (14). Using the same procedure from (9) ~ (12), a test on pathway effect can be performed. PHARAOH-Gee

is the implementation of the cluster phenotype to the PHARAOH method by considering the correlation matrix of the phenotypes. To implement the model, a sandwich type covariance estimator formula as (3) can be used. Sandwich type covariance estimator is robust in the way it well approximates covariance in the wrong assumption of the correlation matrix. Additional simulation studies need to be performed for the non-Gaussian likelihood including partial likelihood to investigate the behavior of parametric tests

HisCoM with L2 regularization or also known as ridge regularization can also be tested using correct estimation of coefficients covariance. Cule et al. suggested significance testing in ridge regression using the covariance estimate [18]. With a proper estimation of degree of freedom, we can expect covariance matrix can be estimated correctly. In the near future, a testing procedure on the regression model with L2 regularization is planned to be developed and verified theoretically and experimentally.



Supplementary figure 2: Type I error and power, log10p-value according to number of pathways and number of genes in each pathway



Supplementary figure 1: Type I error and power, log10p-value according to number of pathways and number of genes in each pathway

Bibliography

1 Ionita-Laza, Iuliana, et al. "Sequence kernel association tests for the combined effect of rare and common variants." *The American Journal of Human Genetics* 92.6 (2013): 841-853.

2 Zhao, Ni, et al. "Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test." *The American Journal of Human Genetics* 96.5 (2015): 797-807.

3 Subramanian, Aravind, et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." *Proceedings of the National Academy of Sciences* 102.43 (2005): 15545-15550.

4 Lee, Sungyoung, et al. "Pathway-based approach using hierarchical components of rare variants to analyze multiple phenotypes." *BMC bioinformatics* 19.4 (2018): 79.

5 Hwang, Heungsun, and Yoshio Takane. "Generalized structured component analysis." *Psychometrika* 69.1 (2004): 81-99.

6 Lee, Sungyoung, et al. "Pathway analysis of rare variants for the clustered phenotypes by using hierarchical structured components analysis." *BMC medical genomics* 12.5 (2019): 100.

7 Lee, Sungyoung, et al. "Pathway-based approach using hierarchical components of rare variants to analyze multiple phenotypes." *BMC bioinformatics* 19.4 (2018): 79.

8 Kim, Yongkang, et al. "Hierarchical structural component modeling of microRNA-mRNA integration analysis." *BMC bioinformatics* 19.4 (2018): 75.

9 Mok, Lydia, et al. "HisCoM-PAGE: Hierarchical Structural Component Models for Pathway Analysis of Gene Expression Data." *Genes* 10.11 (2019): 931.

10 Kim, Sungtae, et al. "Drug response prediction model using a hierarchical structural component modeling method." *BMC bioinformatics* 19.9 (2018): 117.

11 R Choi, Sungkyoung, et al. "HisCoM-GGI: Hierarchical structural component analysis of gene-gene interactions." *Journal of bioinformatics and computational biology* 16.6 (2018): 1840026-1840026.

12 Bickel, Peter J., and Kjell A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics, Volumes I-II Package*. Chapman and Hall/CRC, 2015.

13 Provost, Serge B., and A. M. Mathai. Quadratic forms in random variables: theory and applications. M. Dekker, 1992.

14 Liu, Huan, Yongqiang Tang, and Hao Helen Zhang. "A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables." Computational Statistics & Data Analysis 53.4 (2009): 853-856.

15 Baxter, Nielson T., et al. "Structure of the gut microbiome following colonization with human feces determines colonic tumor burden." *Microbiome* 2.1 (2014): 20.

16 Chen, Yongyu, et al. "Fusobacterium nucleatum Promotes Metastasis in Colorectal Cancer by Activating Autophagy Signaling via the Upregulation of CARD3 Expression." *Theranostics* 10.1 (2020): 323.

17 Raphael, Benjamin J., et al. "Integrated genomic characterization of pancreatic ductal adenocarcinoma." Cancer cell 32.2 (2017): 185-203.

18 Cule, Erika, Paolo Vineis, and Maria De Iorio. "Significance testing in ridge regression for genetic data." *BMC bioinformatics* 12.1 (2011): 372.

초 록

HisCoM (Hierarchical Structural Component Model)은 생물학적 데이터의 특성을 나타내기 위해 일반화선형모델에 계층 구조를 추가하여 분석하는 방법이다. 이 방법의 확장은 임상 데이터에서부터 인체 마이크로바이옴 데이터에 이르기까지 다양한 생물학적 데이터 유형에 맞게 이루어졌다. HisCoM의 회귀 계수 검정을 위해 주로 순열 검정법을 이용해 왔으나, 이 검정법은 high-throughput 기술의 발전 및 데이터 크기의 확장으로 인한 계산량 증가에 따른 속도의 한계가 있다. 이 연구에서는 더 빠른 계산을 위해 순열을 활용하지 않는 모수적 검정을 제시하였다. 시뮬레이션 연구 결과 옳은 분포 하에서 더 빠른 속도 및 더 높은 검정력을 확인하였다. 또한, 암환자들의 RNA-seq 자료 및 인체 마이크로바이움 자료를 통한 실제 자료 분석을 진행하였다.

주요어: 계층적 구조 모형, 모수적 검정, 패스웨이분석, 인체 미생물

학 번: 2018-24959