Hindawi
BioMed Research International
Volume 2019, Article ID 4578983, 10 pages
https://doi.org/10.1155/2019/4578983

Research Article

Multivariate Cluster-Based Multifactor Dimensionality Reduction to Identify Genetic Interactions for Multiple Quantitative Phenotypes

Hyein Kim,[1] Hoe-Bin Jeong,[1] Hye-Young Jung,[2] Taesung Park,[2] and Mira Park[3]

[1] *Department of Statistics, Korea University, Seoul,02841, Republic of Korea*
[2] *Department of Statistics, Seoul National University, Seoul, 08826, Republic of Korea*
[3] *Department of Preventive Medicine, Eulji University, Daejeon, 34824, Republic of Korea*

Correspondence should be addressed to Mira Park; mira@eulji.ac.kr

Received 8 April 2019; Revised 19 June 2019; Accepted 26 June 2019; Published 11 July 2019

Academic Editor: Bilal Alatas

To understand the pathophysiology of complex diseases, including hypertension, diabetes, and autism, deleterious phenotypes are unlikely due to the effects of single genes, but rather, gene-gene interactions (GGIs), which are widely analyzed by multifactor dimensionality reduction (MDR). Early MDR methods mainly focused on binary traits. More recently, several extensions of MDR have been developed for analyzing various traits such as quantitative traits and survival times. Newer technologies, such as genome-wide association studies (GWAS), have now been developed for assessing multiple traits, to simultaneously identify genetic variants associated with various pathological phenotypes. It has also been well demonstrated that analyzing multiple traits has several advantages over single trait analysis. While there remains a need to find GGIs for multiple traits, such studies have become more difficult, due to a lack of novel methods and software. Herein, we propose a novel multi-CMDR method, by combining fuzzy clustering and MDR, to find GGIs for multiple traits. Multi-CMDR showed similar power to existing methods, when phenotypes followed bivariate normal distributions, and showed better power than others for skewed distributions. The validity of multi-CMDR was confirmed by analyzing real-life Korean GWAS data.

1. Introduction

In genome-wide association studies (GWAS), genotype data from a large number of single nucleotide polymorphisms (SNPs) are collected, to associate SNPs with traits of interest [1]. Not only single gene effects, but also interaction effects, between genes, play important roles in complex diseases such as hypertension, diabetes, and autism. By identifying gene-gene interactions (GGIs), we expect to increase statistical power, to detect associations. Moreover, we also hope to clarify the biological pathways underlying human diseases, by detecting interactions between loci [2].

In many cases, a phenotype is considered, and there are various studies on statistical methods for finding GGIs, for univariate phenotypes. For studying qualitative traits, as in the case-control studies, one simple way for identifying genetic interaction is to fit a logistic regression model (LRM) that includes main effects and relevant interaction terms. However, LRMs perform poorly when there is a dimensionality problem. Another well-known approach is a multifactor dimensionality reduction (MDR) method [3, 4], which reduces dimensions by converting a high-dimensional to a one-dimensional model. The genotype combinations are classified as either “high-risk” or “low-risk,” depending on the ratio of cases to controls, for each genotype combination. Thus, an MDR can avoid the issues of sparse data cells and overparameterization of models [2] and can outperform LRMs, for detecting higher order GGIs [5]. Recently, various approaches such as using multiple contingency table (MODENDR) [6] or particle swarm optimization method (PBMDR) have been developed [7].

Due to its superior performance there are now various extensions of MDR, including ordinal phenotypes, quantitative phenotypes, survival information, and odds-ratio-based analysis [8–11]. One specific extension of MDR, generalized MDR, which is applicable to both dichotomous and continuous traits, was proposed [12]. However, GMDR does not provide a computationally efficient algorithm that is easy to implement, and it still requires a dichotomous outcome in the data file [9]. As an alternative, quantitative MDR (QMDR) modified MDR's constructive induction algorithm, which assigns a genotype to either the high- or low-risk groups by comparing the local and global means and then applies a t-test to compare the means of the two groups. More recently, cluster-based MDR (CL-MDR), which is less sensitive to outliers and distributional assumptions, was also developed [13, 14]. Compared to QMDR, CL-MDR was shown to yield higher power, when the phenotype distribution is skewed. However, CL-MDR was developed only for univariate phenotype rather than multivariate phenotypes.

When considering multiple phenotypes, it becomes more difficult to find GGIs. Thus, most GWAS studies still focus on one trait to identify genetic variants associated with common complex traits, even though multiple phenotypes or repeated measurements of phenotypes are available. However, in the study of a complex disease, several correlated traits are often measured at the same time as risk factors for the disease. For example, it is known that intermediately correlated phenotypes, such as Factors VII, VIII, IX, XI, and XII and von Willebrand factor, jointly predict the risk of developing thrombosis [1, 9, 20]. By modeling multivariate disease-related traits, the power to detect associations between genes and diseases is expected to increase. Analyses of multiple traits have been successful in analyzing various complex diseases. In general, the multivariate approach has several advantages over the univariate approach considering one trait at a time. For example, the multivariate approach can consider several traits simultaneously in one model and hence it can take into account the correlation among traits. As a result, the multivariate approach would have higher power to detect pleiotropic genes and it can identify genetic variants not easily detected by the univariate approach [21].

There is relatively less GGI research on multivariate traits case. To deal with multiple phenotypes, generalized estimating equations (GEE)-GMDR is an extension of GMDR method, using the GEE model [22]. Multi-QMDR, which extends QMDR to multivariate cases, has also been proposed [5]. Multi-QMDR classifies samples into high- vs. low-risk groups, by using summary statistics, based mainly on principal component scores. After classification, the two groups' mean vectors are compared, using Hotelling's T^2 statistic. While this approach is simple and intuitive, it is not appropriate when the distribution of phenotypes is not symmetric and/or skewed and is also sensitive to outliers.

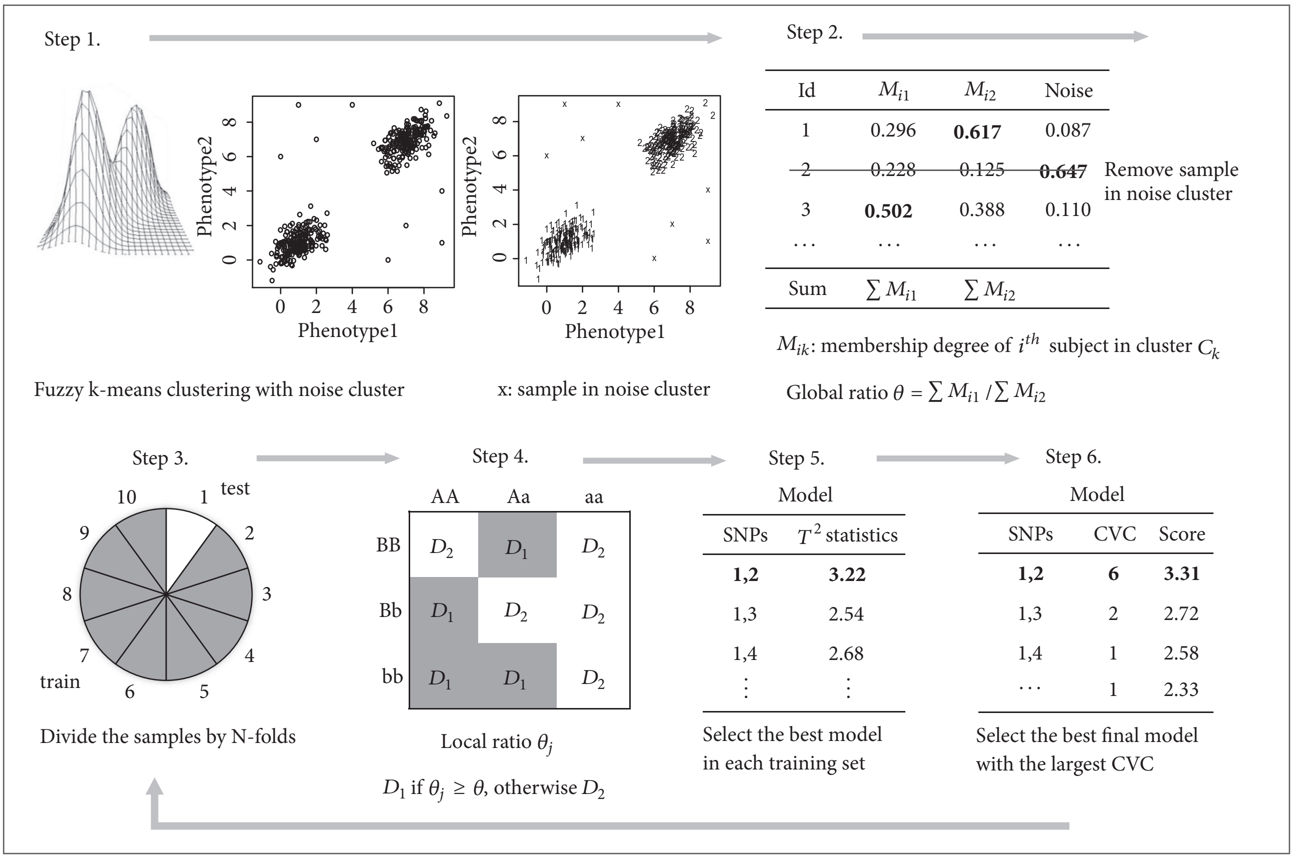
Recently, several MDR extensions were proposed using the fuzzy set theory [23–27]. Such fuzzy set-based MDR methods classify high-risk or low-risk groups as equivalent to defining the degree of membership in high- and low-risk groups. By adopting the fuzzy set theory, fuzzy set-based MDR methods take into account the uncertainty of this binary classification. Fuzzy set-based MDR methods allow the possibility of partial membership into high- and low-risk groups, through a membership function, which transforms the degree of uncertainty into a $[0, 1]$ scale. Then, the best genotype combinations can be selected, maximizing a new fuzzy set-based accuracy measure. Specifically, fuzzy MDR [23] was proposed to detect GGIs for a binary trait and was shown to yield higher power than the original MDR. Furthermore, an empirical fuzzy MDR (EF-MDR) model [24] was proposed to overcome the selection problem of tuning parameters in the original fuzzy MDR, while a fuzzy set-based generalized multifactor dimensionality reduction (FGMDR) model [25] was also proposed for covariate adjustment, for both quantitative and binary traits. More recently, a faster version of EFMDR was developed [26]. Fuzzy C-means-based entropy approach [27] was proposed as the method to detect GGIs for binary trait. It uses two measures: correct classification rate (FCMEMDR-CCR) and likelihood ratio (FCMEMDR-LR).

Here, we propose a new method to detect GGIs for multiple quantitative traits. The main idea of our method to detect GGIs for multiple quantitative traits lies in combining fuzzy clustering with a modified multifactor dimensionality reduction (MDR) approach, named "multivariate cluster MDR" (multi-CMDR). Like other MDR-based methods, multi-CMDR also pools multiple genotype combinations into two groups and uses them as a new attribute, reducing multidimensional space into one dimension. To classify genotype combinations, we first performed fuzzy k-means clustering and computed a threshold, representing the ratio of the sum of the membership degrees of the two groups. Each multilocus genotype is labeled by comparing the local ratio, in each multilocus genotype, to the global ratio. Then, multi-CMDR identifies the best genotype model, using Hotelling's T^2 statistic. To find the overall best model, 10-fold cross-validation (CV) is performed and the best model is chosen which has the largest CV consistency. Unlike other GGI methods for multiple quantitative traits, multi-CMDR is robust to outliers and underlying distributions.

We first introduce the multi-CMDR method in detail in Section 2. We next present a simulation study in Section 3, to show the performance of the proposed methods by comparing them to other methods, such as multi-QMDR. For a phenotype distribution, multivariate normal and multivariate gamma distributions are considered. In Section 4, we apply our method to three lipid-related phenotypes data extracted from the GWA study of the Korean Association Resource (KARE) project, as an illustration. We end with some conclusions in Section 5.

2. Materials and Methods

In this section, we introduce a new procedure, multi-CMDR, for finding GGIs for multiple continuous phenotypes. Similar to other MDR-based methods, multi-CMDR pools multiple genotype combinations into two groups and uses them as a new attribute that reduces a multidimensional space into only one dimension. The detailed algorithm is described in Figure 1 and the multi-CMDR pseudocode is presented in Pseudocode 1.


Step 1.
Phenotype2
Phenotype1
0 2 4 6 8
Fuzzy k-means clustering with noise cluster
x: sample in noise cluster
Step 2.
Remove sample in noise cluster
M_{ik}: membership degree of i^{th} subject in cluster C_k
Global ratio $\theta = \sum M_{i1} / \sum M_{i2}$
Step 3.
1 test
2
3
4
5
6
7 train
8
9
10
Divide the samples by N-folds
Step 4.
Local ratio θ_j
D_1 if $\theta_j \geq \theta$, otherwise D_2
Step 5.
Model
Select the best model in each training set
Step 6.
Model
Select the best final model with the largest CVC

Id	M_{i1}	M_{i2}	Noise
1	0.296	**0.617**	0.087
~~2~~	~~0.228~~	~~0.125~~	~~**0.647**~~
3	**0.502**	0.388	0.110
⋯	⋯	⋯	⋯
Sum	$\sum M_{i1}$	$\sum M_{i2}$	

	AA	Aa	aa
BB	D_2	D_1	D_2
Bb	D_1	D_2	D_2
bb	D_1	D_1	D_2

SNPs	T^2 statistics
1,2	**3.22**
1,3	2.54
1,4	2.68
⋮	⋮

SNPs	CVC	Score
1,2	**6**	**3.31**
1,3	2	2.72
1,4	1	2.58
⋯	1	2.33

Figure 1: Summary of the multi-CMDR algorithm in the case of 10-fold and 2^{nd}-order gene-gene interactions.

```
perform fuzzy k-means clustering with noise cluster for phenotypes
remove samples in noise cluster
compute global ratio $\theta$
get all combinations of SNPs
divide samples into N folds
for k = 1 to N
    set samples in kth folds as test dataset and the other samples as training data
    for i = 1 to number of all combinations of SNPs
        get all combination of genotypes
        for j = 1 to number of all combination of genotypes
            compute local ratio $\theta_j$
            classify each genotype combination as $D_1$ if $\theta_j \geq \theta$, otherwise $D_2$
        end j
        compute Hotelling's $T^2$ statistics for training and test data
    end i
    select the best SNP combination at $i^{th}$ fold by comparing Hotelling's $T^2$ statistics for training data
end k
compute CVC and select SNP combination with highest CVC as the best SNP combination
compute p-value by permutation test for the best SNP combination
```

Pseudocode 1: Pseudocode of multi-CMDR.

Step 0. Preprocessing.

(i) Suppose there are n samples, with p SNP data points and q continuous phenotypes. Let $Y_i = (y_{i1}, y_{i2}, \cdots, y_{iq})^T$ be the phenotype vector and let $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$ be the genotype vector for the i-th subject, respectively, $(i = 1, \cdots, n)$.

(ii) Standardize all the phenotypes to have a mean of zero and no unit variance.

Step 1. Perform fuzzy k-means clustering.

(i) Perform fuzzy k–means clustering with $k = 2$ using phenotype information. Here, we make an additional pseudocluster (i.e., "noise cluster") during the process of clustering [28]. Samples are then allocated into one of three clusters: two good cluster groups and one noise cluster. In this study, we set the noise cluster threshold value to equal the average squared Euclidean distance between samples. Two good clusters and one noise cluster are obtained by minimizing the following J_{Noise}:

$$J_{Noise} = \sum_{i=1}^{n}\sum_{k=1}^{2} M_{ik}{}^{m}\left(x_i - c_k\right)^2 + \sum_{i=1}^{n}\delta^2\left(1 - \sum_{k=1}^{2} M_{ik}\right)^m \quad (1)$$

such that $M_{ik} \in [0,1]$, $\sum_{k=1}^{3} M_{ik} = 1$. M_{ik} is the membership degree of the i^{th} subject in group C_k, c_k is the center of the cluster C_k, M_{i3} is the membership degree of the noise cluster, $m(>1)$ is the fuzzifier parameter which defines the group's fuzziness (usually $m = 2$), and δ is a squared distance of each data point to the noise cluster.

Step 2. Trim the data and calculate the global ratio.

(i) Data are trimmed by removing all the samples in the noise cluster. The remaining samples have membership degrees for each of the two groups. Denote these two groups as C_1 and C_2. The membership degree of the i^{th} subject in group C_k $(k = 1, 2)$ is given by

$$M_{ik} = \frac{1}{\sum_{j=1}^{2}\left[\left(x_i - c_k\right)^2 / \left(x_j - c_j\right)^2\right]^{1/(m-1)}} \quad (2)$$

(ii) Calculate global ratio θ:

$$\theta = \frac{\sum_{i=1}^{n} M_{i1}}{\sum_{i=1}^{n} M_{i2}}, \quad (3)$$

where M_{ik} is the membership degree of the i^{th} subject in cluster C_k.

Step 3. Divide the samples by N-folds.

(i) For N-folds, split the cross-validation (CV) samples randomly into N subgroups of equal size. Let N-1 sets of samples be the training dataset and let the remaining dataset be the test dataset used for evaluating the model.

Step 4. Calculate the local ratio.

(i) To find the m^{th}-order gene-gene interactions, select a set of m SNPs from a pool of SNPs. Calculate the local ratio θ_j for the j^{th} genotype combination in the training set. θ_j is the ratio of the sum of membership degrees of the samples belonging to C_1 to that belonging to C_2:

$$\theta_j = \frac{\sum_{i=1}^{n_j} M_{ij1}}{\sum_{i=1}^{n_j} M_{ij2}}, \quad \left(j = 1, \cdots, 3^m\right) \quad (4)$$

where M_{ijk} is the membership degree of the i^{th} subject with the j^{th} genotype combination, in cluster C_k.

(ii) Label each genotype combination either "D_1," if $\theta_j \geq \theta$, or "D_2," if $\theta_j < \theta$.

Step 5. Calculate the test statistic.

(i) Calculate Hotelling's T^2 statistic, for both training and testing datasets, to test differences of the mean vectors between the D_1 and D_2 groups:

$$T^2 = \left(\overline{x}_1 - \overline{x}_2\right)^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_{pooled}\right]^{-1} \left(\overline{x}_1 - \overline{x}_2\right) \quad (5)$$

where n_1 is the number of observations in group D_1 and n_2 is the number of observations in group D_2; x_{1i} is i^{th} observation of D_1; x_{2j} is j^{th} observation of D_2.

$$\begin{aligned}
\overline{x}_1 &= \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}, \\
\overline{x}_2 &= \frac{1}{n_2}\sum_{j=1}^{n_2} x_{2j}, \\
S_1 &= \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(x_{1i} - \overline{x}_1\right)\left(x_{1i} - \overline{x}_1\right)^T, \quad (6) \\
S_2 &= \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}\left(x_{2j} - \overline{x}_2\right)\left(x_{2j} - \overline{x}_2\right)^T, \\
S_{pooled} &= \frac{n_1 - 1}{n_1 + n_2 - 2} S_1 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2
\end{aligned}$$

(ii) The model with the largest statistic in the training data is chosen as the best model. Statistics for the test data will be performed later.

Step 6. Find the final best model and obtain the empirical p-value.

(i) Repeat Steps 4 and 5 N times, for each fold, and count the number of specific SNP combinations for the best model. We call this cross-validation consistency (CVC).

(ii) Find the best final interaction model, i.e., the one with the largest CVC.

(iii) Derive the final statistic for the best model by averaging N T^2 statistics for the test data and let this statistic be ${T^2}_{test}$.

(iv) To evaluate the statistical significance of the best model, perform a permutation test and obtain the empirical p-value. Generate B permuted datasets by shuffling only the phenotype vector Y_i across individuals while fixing the genotype vector X_i. This way of shuffling nullifies the association between the phenotype and genotype vectors, while preserving the correlation structures within their components. Perform the multi-CMDR and calculate T^2 statistics for each permuted dataset. B test statistics are in ${T^2}_{null}$. The empirical p-value is calculated as

$$p - value = \frac{1}{B}\sum_{i=1}^{B} I\left({T^2}_{null} > {T^2}_{test}\right) \tag{7}$$

where $I(x)$ is indicator function, returning 1 if x is true, otherwise 0.

3. Results and Discussion

3.1. Simulation Analysis. In this section, we conducted simulations to compare the performance of the proposed multi-CMDR method, with multi-QMDR and univariate QMDR methods. We also compared the performance of the two versions of multi-CMDR. One version is a nontrimmed version of multi-CMDR. That is, the noise cluster is not generated in the fuzzy clustering step. The other version uses k-means clustering, without considering membership score. For multi-QMDR methods, the First Principal Component (FPC) was used to classify each cell into high- or low-risk groups, as previously described [5]. For a univariate approach, QMDR was performed for each phenotype, separately. All of these methods were compared in terms of their hit-ratios, representing the ratio at which the true causal SNP pair is identified by the best model.

We then generated a multivariate normal distribution and a multivariate gamma distribution for phenotypes. We used 70 different penetrance functions that define a probabilistic relationship with disease-causal interaction. The models consisted of 7 different heritability values (0.01, 0.025, 0.05, 0.1, 0.2, 0.3, and 0.4) and 2 different minor allele frequencies (MAFs, 0.2 and 0.4). A total of 5 models for each of the 14 heritability-minor allele frequency combinations were considered. Thus, a total of 70 models were generated. The details of the 70 penetrance functions are given in [29]. For every 70 models, 100 datasets were generated. For each dataset, the sample size was 400, and we considered 20 SNPs and 2 continuous phenotypes. SNP1 and SNP2 denoted disease-causal SNP interactions. We used 10-fold cross-validation to determine best overall model.

3.1.1. Multivariate Normal Distribution. For the multivariate normal distributed case, two continuous phenotype values, $Y = (Y_1, Y_2)^T$, were associated with SNP_1 and SNP_2, respectively, and were generated from the bivariate normal distribution,

$$Y \mid (SNP_1 = i, SNP_2 = j) \sim MN\left(\mu_{ij}, \Sigma\right), \tag{8}$$

where $\mu_{ij} = \left(\begin{smallmatrix} f_{ij} \\ f_{ij} \end{smallmatrix}\right)$ and $\Sigma = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$, and f_{ij} is the element from the i^{th} row and j^{th} column of a penetrance function, representing the two functional interacting SNPs. From this, we considered 3 different ρ s : $\rho = 0, 0.25, 0.5$. We used R software to generate simulation data. For multivariate normal distributed cases, we used `mvrnorm()` function in `MASS` package in R.

The hit-ratios for each heritability values are reported in Figure 2. In the bivariate normal distribution case, all the multivariate methods were generally more powerful than the univariate QMDR methods. As the correlation increased, however, the difference between multivariate and univariate methods decreased. All multivariate methods showed similar performance. In the case of zero correlation, multi-QMDR showed slightly better performance than multi-CMDR. The hit-ratios of multi-CMDR, with trimming, were similar to those of multi-CMDR without trimming. That is, there was no effect of trimming outliers in multi-CMDR for the bivariate normal distribution case. The lower the correlation, the higher the hit-ratio, when the values of heritability were 0.05, 0.1, and 0.2. This is because the lower the correlation, the more unique information for each variable. In a similar context, when the correlation was high, the hit-ratios of the multivariate and univariate methods were similar.

3.1.2. Multivariate Gamma Distribution. For the skewed distribution, we generated bivariate gamma distribution using Gaussian copula [30]. In the Gaussian copula, the correlation matrix is responsible for the dependence. We used the same correlation structure, for the bivariate normal case. When the marginal distributions were continuous, a bivariate distribution could be defined by the density of the following form:

$$g\left(y_1, y_2; \Sigma\right) = c\left(u, \Sigma\right) f_1\left(y_1\right) f_2\left(y_2\right), \tag{9}$$

where $c(u, \Sigma)$ represents the copula density, $\Sigma = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$, f_1, f_2 are marginal probability density functions, and g is joint density function. The Gaussian copula density is then defined as follows:

$$c\left(u, R\right) = |\Sigma|^{1/2} \exp\left[-\frac{\tilde{u}^T\left(\Sigma^{-1} - I\right)\tilde{u}}{2}\right] \tag{10}$$

where $\tilde{u} = (\Phi^{-1}(u_1), \Phi^{-1}(u_2))^T$, $u_i = F_i(y_i)$, $i = 1, 2$, and Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution; F_1, F_2 are marginal cumulative distribution functions. The forms of two gamma distributions, $f_1(y_1)$ and $f_2(y_2)$, are as follows:

$$\begin{aligned} Y_1 \mid (SNP1 = i, SNP2 = j) &\sim Gamma\left({f_{ij}}^2, \frac{1}{f_{ij}}\right), \\ Y_2 \mid (SNP1 = i, SNP2 = j) &\sim Gamma\left({f_{ij}}^2, \frac{1}{f_{ij}}\right) \end{aligned} \tag{11}$$

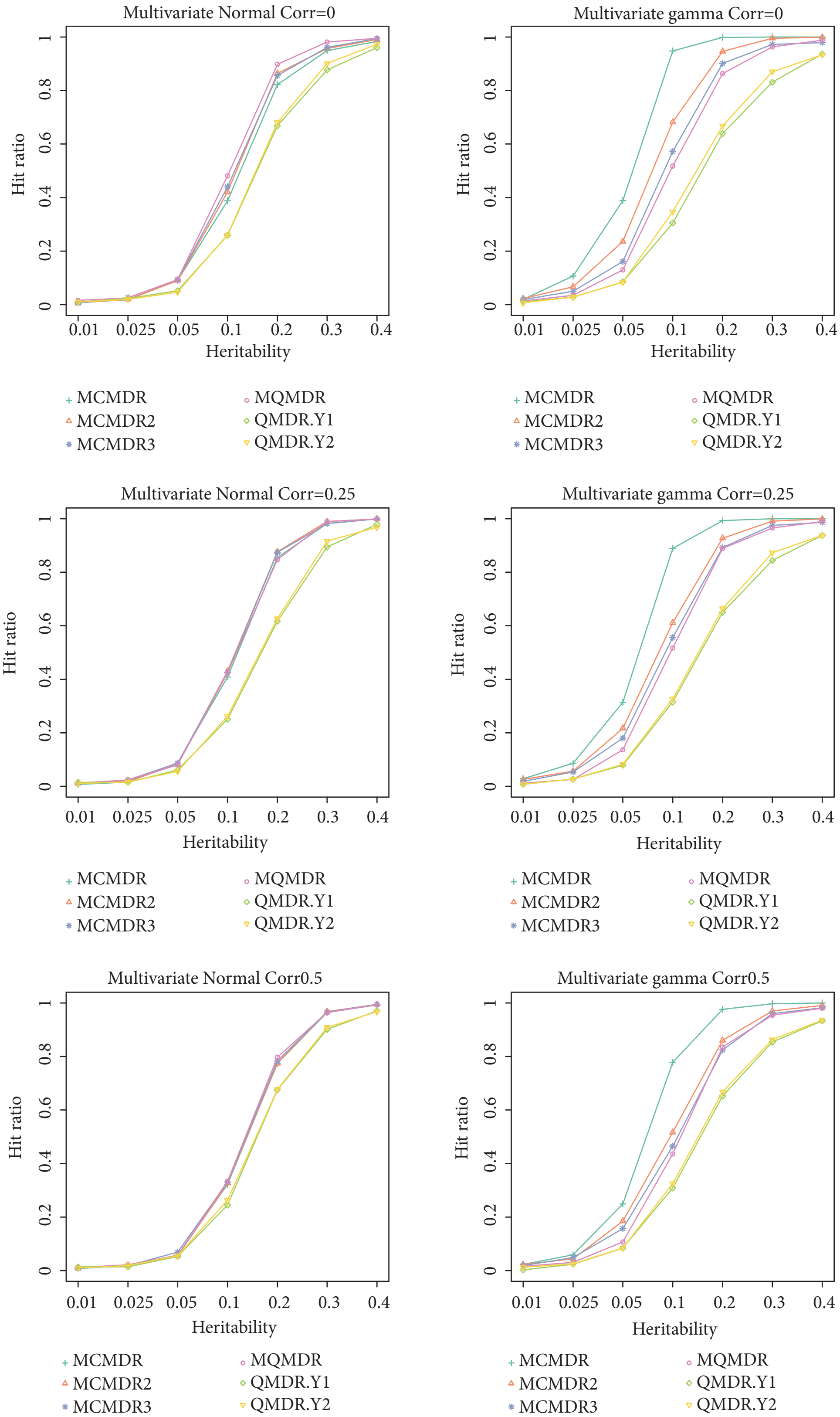


Figure 2: Hit-ratios for a multivariate normal distribution and multivariate gamma distribution. MCMDR (multi-CMDR), MCMDR2 (multi-CMDR without trimming, MCMDR3 (multi-CMDR, without membership score), MQMDR (multi-QMDR), QMDR.Y1 (QMDR with Y_1), and QMDR.Y2 (QMDR with Y_2).

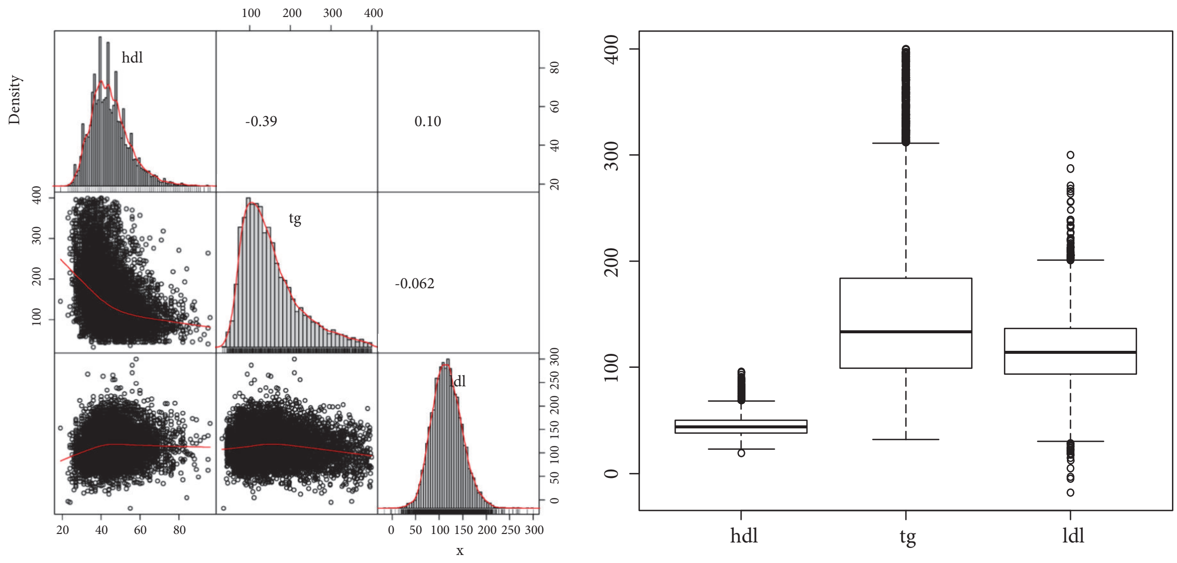


Figure 3: (Left) Scatter plots, histograms, and correlations between phenotypes. (Right) Box plots of phenotypes.

From this, we considered 3 different ρ s : ρ = 0, 0.25, 0.5. For multivariate gamma distributed cases, we used `mvdc(),normalCopula(),rMvdc()` functions in `copula` package in R.

In Figure 2, we observed that the proposed multi-CMDR outperformed the QMDR and the multi-QMDR, for all ranges of heritability, for the bivariate gamma distribution case. Also, multi-CMDR, without trimming, performed better than multi-QMDR. For the bivariate gamma distribution, the lower the correlation, the higher the overall hit ratio. The difference of hit-ratios between multi-CMDR and other methods was greatest when the heritability was 0.1. As the correlation increases, the differences between hit ratios of the multivariate methods, except multi-CMDR, decrease.

To sum up, the power of proposed multi-CMDR is similar to that of multi-QMDR, for symmetric distribution while it outperformed multi-QMDR for the skewed distribution. Moreover, the powers of the two different versions of multi-CMDR were also slightly better than those of multi-QMDR, in skewed phenotype distributions. For all situations, multivariate methods performed better than univariate methods. Results for each combination of two minor allele frequency (MAF) values and 5 models are presented in the supplemental materials (Supplemental Figures 1-6).

3.1.3. Empirical False Positive Rate. We computed empirical false positive rate. To compute empirical false positive rate, we permuted phenotypes over individuals for each case to generate null data. The selection rate of each SNP pair in null data is $1/\binom{20}{2}$= 0.0053. To compute empirical false positive rate, we counted the number of detecting a specific SNP combination, SNP1 and SNP2, as the best model. Overall, empirical false positive rates of each method are closed to the expected value 0.0053. Results for empirical false positive rates of each method are presented in the supplemental materials (Supplemental Tables 1-6).

3.2. Real Biological Data Analysis. For real-life data analysis, three lipid-related phenotypes' data, retrieved from the Korean Association Resource (KARE) project [31], were considered to evaluate the proposed multi-CMDR. Three lipid-related phenotypes consisted of high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL), and triglyceride (TG). After removing those observations with at least one missing phenotype value, there were 8,581 samples remaining. The largest absolute value of correlation between three phenotypes was 0.39 (Figure 3). Among 344,596 SNPs, we used 324 SNPs selected in [5] for this analysis.

We then applied the proposed multi-CMDR to search for the best second interaction model, again by using 10-fold CV. Table 1 displays the best 1^{st} and 2^{nd}-order SNP combinations, identified by the proposed multi-CMDR. In addition to the best model, which has the highest CVC, Table 1 shows other candidate models selected from the best models, in every 10 training datasets. To see if these SNP combinations have been previously detected, one previous study [5] reported the best SNP combinations found in this study, including those described in Table 1.

For 1^{st}-order analysis, rs1106280 was selected as the best model with the highest CVC. rs11066280 was identified as significantly associated with metabolism, TGs, and HDLs [5, 15] and was selected as the best lipid-related phenotypes in a 2^{nd}-order analysis from univariate analysis of HDL using QMDR [5]. The second best model, rs10503669, has been reported to associate with LPL [16]. The third best model, rs2074356, associated with HDL [1]. All p-values selected by the multi-CMDR method were < 10^{-3}.

For 2^{nd}-order analysis, the proposed multi-CMDR identified the best two SNP combinations, rs11216126 and rs4244457, where rs11216126 is reported to be related to HDL [17]. rs4244457 (LPL) occurs in the gene for the key enzyme responsible for the lipolytic processing of TG-rich

Table 1: Best models from 1^{st}- and 2^{nd}-order interaction analysis. T^2 statistics were calculated from the test set.

Order	rs ID	Chr.	CVC	Hotelling's T^2	p-value	Ref.
1^{st}	rs11066280	12	4	2.86	<0.001	[5, 15]
	rs10503669	8	4	2.79	<0.001	[16]
	rs2074356	12	2	2.82	<0.001	[1]
2^{nd}	rs11216126, rs4244457	11, 8	4	3.86	<0.001	[5, 17]
	rs11600380, rs10503669	11, 8	3	3.54	<0.001	[16, 18]
	rs11216126, rs10503669	11, 8	1	3.29	<0.001	[17, 18]
	rs16940212, rs10503669	15, 8	1	3.57	<0.001	[18, 19]
	rs16940212, rs4244457	15, 8	1	2.78	<0.001	[5, 19]

lipoproteins [5]. Note that rs4244457 was selected as the most lipid-related SNP in a 1^{st}- and 2^{nd}-order analysis, using a multi-QMDR method for testing association with LDL [5]. Moreover, rs11600380, rs10503669, and rs16940212 were previously reported to relate to TG, LDL, and HDL, respectively [16, 18, 19]. Each of those three SNPs was also reported in previous studies, but as far as we know, there were no simultaneously reported 2^{nd}-order interactions.

4. Discussion

For GGI analysis for multiple quantitative traits, we proposed multi-CMDR. Analyzing correlated multivariate phenotypes was shown to have higher power to detect susceptible genes and GGIs, by using more information from data [32]. The main feature differences between multi-QMDR and multi-CMDR lies in how to define groups for each combination cell. Multi-QMDR uses summary scores obtained by principal component analysis to classify high-risk and low-risk groups. The observations of each cell are assigned to the high-risk group if the local mean is greater than or equal to the global mean; otherwise the observations are assigned to the low-risk group. On the other hand, multi-CMDR divides groups using clustering. By comparing the global and local ratios, as calculated by using the membership degrees obtained through fuzzy k-means clustering, the observations of each cell are assigned to D_1, if the local ratio is greater than or equal to the global ratio; otherwise the observations are assigned to D_2.

This proposed multi-CMDR was shown to be less sensitive for outliers and nonsymmetric distributions than other methods. 10-fold cross-validation and Hotelling's T^2 statistic were used to select the best model. In the simulation study, we showed that the proposed multi-CMDR could be used effectively in case of bivariate gamma distribution. While the proposed method did not seem to have advantage of computing time over the multi-QMDR method, it was higher for the skewed distribution. In real-life data analysis, multi-CMDR detected the best SNPs and 2-way interactions for lipid-related traits (HDL, TG, and LDL). The best SNPs, selected by our method, have been reported to associate with similar traits [1, 5, 15–19]. While our proposed method performs well for nonsymmetric distributions, it would be always worth to try appropriate transformations to make nonsymmetric distributions symmetric.

In terms of computation time efficiency, multi-QMDR was slightly faster than multi-CMDR. Using an AMD Ryzen 2700x desktop machine with 16G RAM, multi-QMDR took 145.8841 seconds on average (100 repetitions) to conduct real data analysis for the first-order interaction, whereas multi-CMDR took 162.7906 seconds on average. For simulation dataset with 400 sample size and 20 SNPs, multi-QMDR took 17.3334 seconds on average to conduct the 2^{nd}-order interaction, while multi-CMDR took 19.3947 seconds on average. That is, when the number of SNPs is small, the difference in computation time is small. R program to conduct multi-CMDR is available at our github repository (https://github.com/stat17-hb/Multi-CMDR).

5. Conclusion

For the analysis of GGIs associated with multiple quantitative traits, we proposed a new extension of the MDR algorithm that includes clustering. Using fuzzy k-means clustering, we divided samples into two groups and trimmed outliers in noise cluster. By fuzzy k-means clustering, we can capture numerous attributes of multivariate data. Therefore, this is a very productive way to use values calculated from clusters to set thresholds to assign observations to specific groups, in that the proposed multi-CMDR uses a fuzzy k-means clustering method. Unlike k-means clustering, where each observation is assigned to only one cluster, fuzzy k-means clustering provides each observation with a degree of membership to each cluster. Fuzzy k-means clustering is especially useful when the cluster boundary is not clear, and it also allows outliers to be clustered into a noise cluster and reflects individual membership degrees of elements in the same cluster. We expect that multi-CMDR would improve the identification of gene-gene interactions associated with numerous multifactorial human pathologies.

Data Availability

The Korea Association Resource (KARE) project data will be publicly distributed by the Distribution Desk of Korea Biobank Network (https://koreabiobank.re.kr/). The data request should be made directly to Distribution Desk of Korea Biobank Network. Any inquiries should be sent to admin@koreabiobank.re.kr.

Disclosure

This paper has been presented at 2018 annual meeting of the Western North American Region of the International Biometric Society (WNAR), Edmonton, Canada. Our earlier work on univariate CL-MDR was presented at 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, USA.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2017R1A2B4011504, 2013M3A9C4078158). This research was also supported by the Bio & Medical Technology Development Program of the NRF funded by the Korean government, MSIP (No. 2016M3A9B694241).

Supplementary Materials

There were 10 combinations of minor allele frequencies (MAFs) and 5 models for each simulation setup. The MAF was 0.2, up to model 5. From models 6 to 10, the MAF was 0.4. We also considered correlation values of 0, 0.25, and 0.5. CMDR (multi-CMDR), MCMDR2 (multi-CMDR without trimming), MCMDR3 (multi-CMDR without membership score), MQMDR (multi-QMDR), QMDR. Y1 (QMDR with Y_1), and QMDR Y2 (QMDR with Y_2). *(Supplementary Materials)*

References

[1] S. Basu, Y. Zhang, D. Ray, M. B. Miller, W. G. Iacono, and M. McGue, "A Rapid Gene-Based Genome-Wide Association Test with Multivariate Traits," *Human Heredity*, vol. 76, no. 2, pp. 53–63, 2013.

[2] H. J. Cordell, "Detecting gene-gene interactions that underlie human diseases," *Nature Reviews Genetics*, vol. 10, no. 6, pp. 392–404, 2009.

[3] M. D. Ritchie, L. W. Hahn, N. Roodi et al., "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *American Journal of Human Genetics*, vol. 69, no. 1, pp. 138–147, 2001.

[4] L. W. Hahn, M. D. Ritchie, and J. H. Moore, "Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions," *Bioinformatics*, vol. 19, no. 3, pp. 376–382, 2003.

[5] W. Yu, M. Kwon, and T. Park, "Multivariate Quantitative Multifactor Dimensionality Reduction for Detecting Gene-Gene Interactions," *Human Heredity*, vol. 79, no. 3-4, pp. 168–181, 2015.

[6] C. Yang, L. Chuang, and Y. Lin, "Multiobjective differential evolution-based multifactor dimensionality reduction for detecting gene-gene interactions," 2017.

[7] C. Yang, H. Yang, and L. Chuang, "PBMDR: A particle swarm optimization-based multifactor dimensionality reduction for the detection of multilocus interactions," *Journal of Theoretical Biology*, vol. 461, pp. 68–75, 2019.

[8] D. Gola, J. M. Mahachie John, K. van Steen, and I. R. König, "A roadmap to multifactor dimensionality reduction methods," *Briefings in Bioinformatics*, vol. 17, no. 2, pp. 293–308, 2016.

[9] M. Germain, N. Saut, N. Greliche et al., "Genetics of venous thrombosis: insights from a new genome wide association study," *PLoS ONE*, vol. 6, no. 9, 2011.

[10] Y. Chung, S. Y. Lee, R. C. Elston, and T. Park, "Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions," *Bioinformatics*, vol. 23, no. 1, pp. 71–76, 2007.

[11] S. Yeoun Lee, Y. Chung, R. C. Elston, Y. Kim, and T. Park, "Log-linear model-based multifactor dimensionality reduction method to detect gene-gene interactions," *Bioinformatics*, vol. 23, no. 19, pp. 2589–2595, 2007.

[12] X.-Y. Lou, G.-B. Chen, L. Yan et al., "A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence," *American Journal of Human Genetics*, vol. 80, no. 6, pp. 1125–1137, 2007.

[13] Y. Lee, H. Kim, T. Park, and M. Park, "Cluster-based multifactor dimensionality reduction method to identify gene-gene interactions for quantitative traits in genome-wide studies," in *Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM '17)*, pp. 1772–1776, 2017.

[14] Y. Lee, H. Kim, T. Park, and M. Park, "Gene-gene interaction analysis for quantitative trait using cluster-based multifactor dimensionality reduction method," *International Journal of Data Mining and Bioinformatics*, vol. 20, no. 1, p. 1, 2018.

[15] N. Kato, F. Takeuchi, Y. Tabara et al., "Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in east Asians," *Nature Genetics*, vol. 43, no. 6, pp. 531–538, 2011.

[16] C. J. Willer, S. Sanna, A. U. Jackson et al., "Newly identified loci that influence lipid concentrations and risk of coronary artery disease," *Nature Genetics*, vol. 40, no. 2, pp. 161–169, 2008.

[17] Y. J. Kim, M. J. Go, C. Hu et al., "Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits," *Nature Genetics*, vol. 43, no. 10, pp. 990–995, 2011.

[18] F. Asselbergs, Y. Guo, E. van Iperen et al., "Large-Scale Gene-Centric Meta-analysis across 32 Studies Identifies Multiple Lipid Loci," *American Journal of Human Genetics*, vol. 91, no. 5, pp. 823–838, 2012.

[19] M. J. Go, J. Hwang, D. Kim et al., "Effect of Genetic Predisposition on Blood Lipid Traits," *Genomics & Informatics*, vol. 10, no. 2, pp. 99–105, 2012.

[20] J. C. Souto, L. Almasy, M. Borrell et al., "Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study. Genetic Analysis of Idiopathic Thrombophilia," *American Journal of Human Genetics*, vol. 67, no. 6, pp. 1452–1459, 2000.

[21] S. Oh, I. Huh, S. Y. Lee, and T. Park, "Analysis of multiple related phenotypes in genome-wide association studies," *Journal of Bioinformatics and Computational Biology*, vol. 14, no. 05, p. 1644005, 2016.

[22] H. Xu, X. Sun, T. Qi et al., "Multivariate Dimensionality Reduction Approaches to Identify Gene-Gene and Gene-Environment Interactions Underlying Multiple Complex Traits," *PLoS ONE*, vol. 9, no. 9, pp. 1–12, 2014.

[23] H. Jung, S. Leem, S. Lee, and T. Park, "A novel fuzzy set based multifactor dimensionality reduction method for detecting gene–gene interaction," *Computational Biology and Chemistry*, vol. 65, pp. 193–202, 2016.

[24] S. Leem and T. Park, "An empirical fuzzy multifactor dimensionality reduction method for detecting gene-gene interactions," *BMC Genomics*, vol. 18, 2, pp. 1–12, 2017.

[25] H. Jung, S. Leem, and T. Park, "Fuzzy set-based generalized multifactor dimensionality reduction analysis of gene-gene interactions," *BMC Medical Genomics*, vol. 11, no. S2, pp. 11–20, 2018.

[26] S. Leem and T. Park, "EFMDR-Fast: An Application of Empirical Fuzzy Multifactor Dimensionality Reduction for Fast Execution," *Genomics & Informatics*, vol. 16, no. 4, p. e37, 2018.

[27] C.-H. Yang, L.-Y. Chuang, and Y.-D. Lin, "Epistasis Analysis using an Improved Fuzzy C-means-based Entropy Approach," *IEEE Transactions on Fuzzy Systems*, vol. PP, no. L, p. 1, 2019.

[28] R. N. Davé, "Characterization and detection of noise in clustering," *Pattern Recognition Letters*, vol. 12, no. 11, pp. 657–664, 1991.

[29] D. R. Velez, B. C. White, A. A. Motsinger et al., "A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction," *Genetic Epidemiology*, vol. 30, no. 8, pp. 718–727, 2007.

[30] Y. Stitou, N. Lasmar, and Y. Berthoumieu, "Copulas based multivariate Gamma modeling for texture classification," in *Proceedings of the IEEE Int. Conf. Data Min*, pp. 1045–1048, 2009.

[31] Y. S. Cho, M. J. Go, Y. J. Kim et al., "A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits," *Nature Genetics*, vol. 41, no. 5, pp. 527–534, 2009.

[32] J. Choi and T. Park, "Multivariate generalized multifactor dimensionality reduction to detect gene-gene interactions," *BMC Systems Biology*, vol. 7, no. Suppl 6, pp. 1–11, 2013.