Conditional Distance Correlation Test for Gene Expression Level, DNA Methylation Level and Copy Number

Shanshan Zhang
University of Arkansas, Fayetteville

Conditional Distance Correlation Test for Gene Expression Level, DNA Methylation Level and Copy Number

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Statistics and Analytics

by

Shanshan Zhang
University of Arkansas
Bachelor of Business Administration, 2016

December 2020
University of Arkansas

This thesis is approved for recommendation to the Graduate Council.

Qingyang Zhang Ph.D.
Thesis Director

Mark Arnold Ph.D.
Committee Member

Jyotishka Datta Ph.D.
Committee Member

ABSTRACT

Over the past years, efforts have been devoted to the genome-wide analysis of genetic and epigenetic profiles to better understand the underlying biological mechanisms of complex diseases such as cancer. It is of great importance to unravel the complex dependence structure between biological factors, and many conditional dependence tests have been developed to meet this need. The traditional partial correlation method can only capture the linear partial correlation, but not the nonlinear correlation. To overcome this limitation, we propose to use the innovative conditional distance correlation (CDC), which measures the conditional dependence between random vectors and detect nonlinear relations. In this thesis, the CDC measure is applied to the rich Cancer Genome Atlas (TCGA) ovarian cancer data, and we identify a list of interesting genes with nonlinear features. We integrate three important types of molecular features including gene expression, DNA methylation and copy number variation, and implement the partial correlation test and CDC test to infer the relations between the three measurements for each gene. Out of 196 candidate oncogenes and tumor suppressors, we identify 19 genes in which two of the molecular features are nonlinearly dependent given the third variable. Of these 19 genes, many were reported to be associated with ovarian cancer or breast cancer in the literature. Our findings could shed new light on the biological relations between the three important molecular aspects.

This thesis is structured as follows: we begin with a brief introduction to ovarian cancer, TCGA data, the three molecular measurements, and two testing methods in Chapter 1. In the second chapter, we review different statistical methods including Pearson's partial correlation

and conditional distance correlation. In Chapter 3, we conduct an extensive simulation study to compare the empirical performance of different methods. In Chapter 4, we apply the new method to the TCGA ovarian data. We conclude the thesis with future directions in Chapter 5.

TABLE OF CONTENTS

1. INTRODUCTION

Ovarian cancer, ranking the fifth in cancer death among women, is one of the most common cancers in the United States. It is also one of the deadliest gynecologic cancer. Ovarian cancer accounts for 2.5 percent of cancers in women and the most common age range at diagnosis is 55-64 years old and the median age of death from ovarian cancer is 70 [22]. According to the American Cancer Society, it is estimated that there will be about 21,750 women receive a new diagnosis of ovarian cancer and there will be 13,940 about women die from it in the United States in 2020 [32].

Based on the statistics of the American Cancer Society, the majority of the patients are diagnosed in high-stage and usually treated with aggressive surgery followed by platinum-taxane chemotherapy. About 25% of patients recur platinum-resistant cancer within six months after chemotherapy and the overall five-year survival rate is 31%. Over the past years, studies have shown that many different factors may contribute to ovarian cancer. There is approximately 13% of high-grade serous ovarian cancer that can be attributed to germline mutations in *BRCA1* and *BRCA2*, and a smaller percentage of ovarian cancer can be attributed to other germline mutations [3].

As most previous studies focused on individual genes or single type of data, the analyses often fail to provide the accurate prediction of the status of ovarian cancer. Therefore, the ideal approach is to combine multiple genetic and epigenetic profiles together and perform an integrative analysis associated with ovarian cancer. The Cancer Genome Atlas (TCGA) program has profiled the most comprehensive genomic data resource from more than 30 types of cancers [20]. For instance, TCGA has collected and processed more than 500 high-quality samples from ovarian cancer and the data contains clinical information, metadata, histopathology slide images, and molecular information. The clinical profile derived from samples includes records on recurrence, survival, and treatment resistance. The metadata includes the weight of a sample portion, etc. The molecular profile derived from samples includes gene expression (microarray), genotype (SNP), exon expression, MircoRNA expression (microarray), copy number variation (CNV), DNA methylation, etc. Such massive dataset has motivated many studies to reveal the complex mechanisms of ovarian cancer by incorporating interactions between different genetic and epigenetic factors. In this thesis, we jointly modeled three important types of molecular data measurements including gene expression, DNA methylation, and copy number variation (CNV) and inferred their relations using Pearson's Correlation and Conditional Distance Correlation.

Gene expression is a fundamental process by which the genetic instructions in gene are converted into functional products, which are usually proteins. In a few cases, the genetic product is a small nuclear RNA, rather than protein. There are several basic steps in gene expression process towards the final product, including transcription, RNA splicing, translation, and post-translational modification of a protein. The two main steps involved in this process are transcription and translation. Transcription is the process in which DNA in a gene synthesize an RNA transcript called messenger RNA (mRNA) under the enzyme RNA polymerase action, therefore RNA usually has similar structure and properties with DNA. After carrying the genetic information from DNA, mRNA is read by a molecule called transfer RNA (tRNA). Using mRNA as a template and tRNA as a vehicle, the process of assembling activated amino acids on a ribosome into a protein polypeptide chain is called translation under the action of enzymes, cofactors and energy [37]. However, there may exist perturbations of transcription that affecting mRNA expression and protein synthesis, consequently leading to human pathological states, such as malignancy. DNA methylation and copy number variation are the two main reasons resulting the disruption of gene expression.

DNA methylation is a biochemical modifiable process by adding methyl groups to DNA molecule. It is an epigenetic mechanism that could change the activity of a DNA segment without changing its sequence. Cytosine and adenine are the only two bases in DNA that could be methylated, so the most common DNA methylation process occurs by addition of a methyl group to the 5 position of cytosine pyrimidine ring or the number 6 nitrogen of the adenine purine ring. It is commonly known that certain tumor suppressor genes are inactivated within the promotor region that lead to the consequence of abnormal hypermethylation [13]. In addition, a large number of studies have shown that there are many different types of genes silenced by aberrant DNA methylation are associated with different types of human cancers.

Copy number variation (CNV) refers to the variation in the number of repeats of a particular genetic region between individuals in human population. It is an important component of structural variation, which could be duplication, deletion, insertion or single nucleotide polymorphism. Copy number variation occurs during DNA replication and thus change the gene expression level and associated phenotypes. Previous studies have shown that CNV is associated with dozens of human diseases, especially on neurological disorders and cancers. Researchers have demonstrated that approximately 15% of those neurodevelopmental diseases are caused by CNV, such as autism and schizophrenia and a few neurodevelopmental related genes such as *A2BP1* are reported with mutational CNVs [27]. The reason why CNV and neurodevelopmental diseases are associated could be perturbation of genes involving in neurological disorders. Many cancers are also associated with copy number variation. Li at al. found that the CNVs of *AKT2, PIK3CA* etc. could result in breast cancer in late age, which is partly because copy number changed easily due to cellular stress [16]. Therefore, it is essential to integrate all these important aspects in the data analysis.

Since both DNA methylation and CNV can affect gene expression, it is necessary to illustrate the relationship between the three variables. In this paper, we introduce two methods to test conditional dependence including linear conditional relationship and nonlinear conditional relationship. Pearson's partial correlation method measures linear correlations between two multivariate variables given a third random variable. On the other hand, we can test nonlinear conditional independence through conditional distance correlation method.

Pearson's partial correlation is a measure of linear association between two random variables while controlling one or more additional variables. The assumptions and properties are analogous to Pearson's correlation. The graphical models based on Pearson's partial correlation is similar to Bayesian Graphical Model. For example, Xu et al. (2014) assume that the variables in graphical model are joint Gaussian. Xu et al. also applied the model along with MCMC estimation to the same TCGA data independently to describe the dependence structure of specific regulatory relationships [36]. For instance, in the graphical structure of gene *ERLIN2*, there is an edge between gene expression and CNV but no edge between gene expression and DNA methylation, indicating that the expression level of gene *ERLIN2* is correlated with CNV, but not its methylation level [36]. Freudenberg et al. (2009) analyzed the causal relationships between four variables in the human genome based on partial correlation method [4]. Poli et al. (2014) proposed the functional connectivity studies in neuronal network by partial correlation method and other information-based methods [23]. Bühlmann, Kalisch and Maathuis (2010) developed the partial correlation based on variable selection method for normal linear regression model [2]. Li, Liu and Lou (2017) addressed two significant issues regarding to partial correlation based variable selection method, namely the non-robustness to normality and high dimensionality [15]. Due to the limitation listed above, this method may not serve as a general measurement of conditional dependence test.

Existing association tests mostly focus on linear conditional correlations, thus not sensitive to nonlinear conditional relations especially to non-monotonic relation. Conditional distance correlation method is a better way when the relations between the two multivariate random variables with arbitrary dimensions conditioning on another random variable is not linear. Some examples of such conditional independence tests have been developed based on a weighted Hellinger distance [29] between the conditional densities or the difference between the conditional characteristic functions [28]. Huang (2010) used the maximal nonlinear conditional correlation to test conditional independence [8]. Gao and Zhao (2013) proposed to test the dependence structure of multivariate nonlinear times series based on conditional independence graph and applied the statistical mechanics to international financial markets [5]. We propose a nonparametric measure of conditional dependence and conditional correlation (covariance) for multivariate random variables. Especially, the conditional distance correlation coefficient being zero is equivalent to that the two multivariate random variables are conditionally independent given a third multivariate random variable [35]. The conditional distance correlation is defined by replacing characteristic function used in the correlation definition of Szekely et al. (2007) [30]. Concepts and properties of this novel dependence measure are provided in the following chapter.

2. METHODOLOGY

The Cancer Genome Atlas (TCGA) Research Network has provided the most comprehensive genomic data resources over more than 20 types of cancers. The TCGA project has examined 580 samples and 12,000 genes for ovarian cancer. In this work, we only consider three types of molecular data including gene expression, DNA methylation and copy number variation to develop the integrative network analysis for ovarian cancer and infer the differences between the three variables based on different methods. We denote gene expression as E, DNA methylation as M, and DNA copy number variation as C for the ease of notations. Specifically, let us use p_{lig} to represent the measurement of conditional p-value for the gene g, on $i\text{-}th$ sample, with the level l. Here, $l = 1$ denotes the level of E and M conditioning on C, $l = 2$ denotes the level of E and C conditioning on M, $l = 3$ denotes the level of M and C conditioning on E respectively, i indexes the $N = 580$ samples, g indexes the $G = 12{,}000$ genes.

The three methods being considered in this work for integrative analysis are Pearson's partial correlation, conditional mutual information and conditional distance correlation. Pearson's partial correlation is commonly used in assessing two quantitative variable correlations while eliminating the effect of one or more variables. Formally, let X, Y be random variables and Z be a specific quantitative variable. The formula of the partial correlation coefficient of X, Y given Z is defined as

$$\rho_{XY\cdot Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1-\rho_{XZ}^2} * \sqrt{1-\rho_{YZ}^2}},$$

where $\rho_{()}$ denotes Pearson's correlation coefficient between two variables.

The key assumptions that partial correlation relies upon are data normality and variable linearity, which are the same as Pearson's correlation. Note that if the joint distribution of the variables is bivariate normal then these other assumptions are necessarily met (Tabachnick and Fidell 2001,p.72). Like the marginal Pearson's correlation coefficient, Pearson's partial correlation coefficient, ρ, also has a value ranging between -1 and 1, where the larger the absolute value of coefficient, the stronger the association between the paired variables. It also shows "-1" means perfect negative association, "0" means no linear association, and "1" means perfect positive association [34]. The estimate of Pearson's partial correlation coefficient as well as p-value could be implemented using R package "ppcor" (https://cran.r-project.org/package=ppcor). The general idea behind the algorithm implemented in this package is that the derivation of a general matrix formula (inverse variance-covariance matrix) of the semi-partial correlation to resolve higher-order coefficient fast [11]. One could use it to statistically test linear conditional dependence between gene expression, DNA methylation, and CNV. When the p-value is less than the significace level α, we reject H_0 and accept H_α, i.e., two random variables given a third random variable are linearly dependent.

Conditional mutual information (CMI) measures conditional dependence between two variables given the third random variable. To begin with, we define entropy for a given random variable. For a discrete variable X, the entropy H(X) measures average expected uncertainty in variable X, and the formula of H(X) is defined as

$$H(X) = -\sum_{x \in X} p(x)\ \log p(x),$$

where $p(x)$ represents the probability mass function of each value x in the sampling space. The joint entropy H(X,Y) can be defined as

$$H(X,Y) = -\sum_{x \in X, y \epsilon Y} p(x,y) log p(x,y),$$

where $p(x,y)$ is the joint probability of X=x and Y=y.

Mutual information (MI) is the measure of marginal dependence between two random variables. The MI between two discrete variables X, Y can be expressed as

$$I(X,Y) = -\sum_{x \in X, y \epsilon Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}.$$

Similarly, the CMI of variables X and Y given Z can be written as

$$I(X,Y|Z) = -\sum_{x \in X, y \epsilon Y, z \epsilon Z} p(x,y,z) log \frac{p(x,y|z)}{p(x|z)p(y|z)}.$$

It can be shown that when variables X, Y given Z are conditionally independent, we have $I(X,Y|Z) = 0$. On the other hand, the higher value of CMI, the closer relationship between X and Y given Z will be. To perform the hypothesis test, Z-statistic was proposed by Kalisch and Bühlmann (2007) [10] and Satio et al., (2011) [26] and the CMIs are transformed using Fisher's Z transformation to approach normal distribution. First, the CMIs are normalized by

$$\hat{I}(X,Y|Z) = \frac{I(X,Y|Z)}{H(X,Z) + H(Y,Z)}.$$

The Fisher's Z-score can then be computed as following

$$Z_{X,Y|Z} = \frac{1}{2} \log\left(\frac{1 + \hat{I}(X,Y|Z)}{1 - \hat{I}(X,Y|Z)}\right).$$

where $\hat{I}(X, Y|Z)$ represents the normalized CMI, and $Z_{X,Y|Z}$ is the z-value of $\hat{I}(X, Y|Z)$. The p-value can be well approximated a normal distribution with mean 0 and variance $\frac{1}{n-4}$, which was developed by Zhang et al. (2011) [39].

We applied the R package "infotheo" to compute the conditional entropy and CMI [17]. The obtained CMI is then normalized by the entropy and the p-value can be calculated via Fisher's z transformation. The function "discretize" is used to discretize continuous data. Conditional entropy and conditional mutual information in natural logarithm are computed by using functions "condentropy" and "condinformation", and exponential transformation is taken to find the conditional entropy and conditional mutual information. Using the significance level of α, we reject the null hypothesis $H_0: Z_{X,Y|Z} = 0$ against the two-sided alternative hypothesis $H_\alpha: Z_{X,Y|Z} \neq 0$ if p-value is lower than the significance level α, otherwise, H_0 will be retained.

Conditional distance correlation is an innovative measure for conditional dependence, defined through weighted distance between $\emptyset_{X,Y|Z}$ and $\emptyset_{X|Z}\emptyset_{Y|Z}$, where $\emptyset_{X,Y|Z}$ is the conditional joint characteristic function of X, Y given Z and $\emptyset_{X|Z}, \emptyset_{Y|Z}$ are the conditional marginal characteristic function of X, Y given Z, respectively. The conditional joint characteristic function of X, Y given Z can be denoted as

$$\emptyset_{X,Y|Z}(t, s) = E[\exp(i\langle t, X\rangle + i\langle s, Y\rangle)\,|Z],$$

where E is the expectation, i is the imaginary unit, and $\langle\cdot,\cdot\rangle$ is the inner product of the two corresponding vectors. The conditional marginal characteristic functions of X, Y given Z are defined similarly as, respectively

$$\emptyset_{X|Z}(t) = \emptyset_{X|Z}(t, 0), and\ \emptyset_{Y|Z}(t) = \emptyset_{Y|Z}(0, s).$$

Here if X, Y given Z are conditionally independent, we denoted by $X \perp Y|Z$, and $\emptyset_{X,Y|Z} - \emptyset_{X|Z}\emptyset_{Y|Z} = 0$. Otherwise, the conditional distance correlation (covariance) for random variables could be defined by replacing the characteristic functions with conditional characteristic functions in the definition of distance correlation (covariance) [30]. The formulation of conditional distance covariance (CDCov) $\mathcal{D}(X, Y|Z)$ between random vectors X and Y with finite moments given Z can be expressed as the square root of

$$\mathcal{D}^2(X, Y|Z) = \| \emptyset_{X,Y|Z}(t, s) - \emptyset_{X|Z}(t)\emptyset_{Y|Z}(s) \|^2$$

$$= \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|\emptyset_{X,Y|Z}(t, s) - \emptyset_{X|Z}(t)\emptyset_{Y|Z}(s)|^2}{|t|_p^{P+1}|s|_q^{q+1}} dtds,$$

where $c_p = \frac{\pi^{(p+1)/2}}{\Gamma((p+1)/2)}, c_q = \frac{\pi^{(q+1)/2}}{\Gamma((q+1)/2)}$.

Similarly, the conditional distance variance is defined as the square root of

$$\mathcal{D}^2(X|Z) = \mathcal{D}^2(X, X|Z),$$

and the conditional distance correlation (CDCor) between random vectors X and Y with finite moments given Z is defined as the square root of

$$\rho^2(X, Y \mid Z) = \frac{\mathcal{D}^2(X, Y|Z)}{\sqrt{\mathcal{D}^2(X|Z)\, \mathcal{D}^2(Y|Z)}}.$$

if $\mathcal{D}^2(X|Z)\, \mathcal{D}^2(Y|Z) > 0$, or 0 otherwise. Hence, the conditional distance correlation is always nonnegative. Although the definitions are straightforward, the values are difficult to calculate by hand, so we use the R package “cdcsis”, developed by Hu et al. (2019) to find the conditional distance correlation for conditional independence inference [7].

Next, we review several desirable theoretical properties for conditional distance covariance, which are analogous to the properties of unconditional distance covariance [35]. On remarkable property of DC is that $\mathcal{D}^2(X, Y|Z) \geq 0$, and $\mathcal{D}^2(X, Y|Z) = 0$ if and only if X and Y are conditionally independent given Z, which means the necessary sufficient condition of conditional independence is that conditional distance covariance is nonnegative and that $\mathcal{D}^2(X, Y|Z) = 0$. In addition, if $E\left(|X|_p + |Y|_q \middle| Z\right) < \infty,\ then\ 0 \leq \rho(X, Y|Z) \leq 1$. If $\rho(X, Y|Z) = 0$ means X is conditionally independent of Y given Z, and as long as $\rho(X, Y|Z) = 1$, Y is a linear transformation of X conditioning on Z, which means $Y = AX + b$. They are valuable to do the conditional dependence test in the following chapters.

In this paper, we use permutation p-value for hypothesis testing as the sampling distribution of CDC is impractical to obtain [6]. We randomly shuffle the dataset and obtain the new test statistic for shuffled dataset and repeat the procedure for a predetermined number of times under null hypothesis. The doubled minimum ranking of old test statistic among the new test statistic for the shuffled dataset gives an approximate two-sided p-value. With the null hypothesis $H_0: \mathcal{D}^2(X, Y|Z) = 0$ and the alternative hypothesis $H_\alpha: \mathcal{D}^2(X, Y|Z) > 0$ under the significance level alpha, the classic decision theory follows a rule that the null hypothesis is rejected if p-value is lower than the significance level, otherwise the alternative hypothesis is rejected but the null hypothesis is not rejected.

All the three aforementioned methods are model-free, which do not rely on any model structure or assumptions, but it may lack power to detect the complex mechanisms of ovarian cancer formation by overlooking the interactions of gene expression, DNA methylation and copy number variation. To this end, Xu et al. (2014) applied a Bayesian graphical model to incorporate the complex interactions between different variables [36]. Bayesian graphical model by Xu et al. assumed a mixture model through Markov Chain Monte Carlo (MCMC) posterior distribution and simplified the conjugate priors to obtain the posterior estimates and the estimated graphs. For the model-based method, if the model could be found appropriately, the conclusion will be more powerful than based on model-free method. However, it is challenging to specify an appropriate model structure in real data analysis.

3. SIMULATION STUDY

3.1 Evaluation of Type I Error Rate

In this chapter, we conducted an extensive simulation study to evaluate the statistical performance of the three methods being compared, namely CDC test, partial correlation test and CMI test. The simulation settings are similar to the ones used in Wang et al., 2015 [35]. For each simulation, we generated 50 samples and use the significance level of 0.05 for test. Examples 1-3 considered the settings where X and Y are conditionally independent given Z and we evaluate the three models in terms of type I error rate. Type I error rate is the probability that rejection of the null hypothesis when the null hypothesis is true. Suppose we repeat the tests for 100 times and get 100 p-values, the proportion of the p-values that is less than the significance level is summarized as type I error rate.

Example 1 (X, Y, Z) follows a multivariate normal distribution with mean $\mu = (0,0,0)$ and covariance matrix $\Sigma = \begin{pmatrix} 1 & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{4} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix}$. Therefore, the conditional covariance matrix of X and Y given Z is $\Sigma(\mathrm{X}, \mathrm{Y} \mid Z) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \begin{pmatrix} \frac{3}{4} & 0 \\ 0 & \frac{3}{4} \end{pmatrix}$, where $\Sigma(\mathrm{X}, \mathrm{Y} \mid Z)$ is 2-by-2 diagonal matrix, so the element (1,2) and (2,1) are all zero, indicating that X is independent of Y given Z.

As X, Y, Z follow the multivariate normal distribution, they are linearly associated. In addition, X is independent of Y given Z, therefore, X, Y are linearly independent given Z.

Example 2 $X_1, Y_1, Z \sim N(0,1)$, and we define

$$Z_1 = 0.6 * \left(\frac{Z^3}{7} + \frac{Z}{2}\right), Z_2 = \frac{1}{4} * \left(\frac{Z^3}{2} + Z\right),$$

$$X_2 = Z_1 + \tanh(X_1)\,, X = X_2 + \frac{1}{4}(X_2)^3,$$

$$Y_2 = Z_2 + Y_1, Y = Y_2 + \tanh\left(\frac{Y_2}{4}\right).$$

Since X_1, Y_1, Z follow standard normal distribution independently, X and Y are polynomial transformation of X_1, Y_1, Z, therefore, X and Y are statistically independent given Z.

Example 3 $X_1, Y_1, Z_1, Z_2 \sim Binomial(12,0.4)$, and we define

$$X = X_1 + Z_1 + Z_2,$$

$$Y = Y_1 + Z_1 + Z_2,$$

$$Z = (Z_1, Z_2).$$

Here X_1, Y_1, Z_1, Z_2 are independent binomial variables, and X, Y are linear transformation of X_1, Y_1, Z_1, Z_2, Z is a vector containing Z_1, Z_2, therefore X and Y are linearly independent given Z.

Note that all X, Y, Z in examples 1 and 2 are univariate. In example 3, X, Y are univariate, and Z is multivariate. Table 1 shows the type I error rate for examples 1-3 based on three methods.

Table 1. Type I error rate for examples 1-3 with three different methods (n=50).

Type I Error	Example 1	Example 2	Example 3
Pearson’s Correlation	0.05	0.25	0.02
Conditional Mutual Information	0	0	0
Conditional Distance Correlation with Index=1	0.04	0.14	0.32

We can see that the empirical type I errors of partial correlation method and CDC method are reasonable but CMI method is inconclusive with 50 random samples. With CMI method, all type I errors for the three examples are zeros, which means that p-values are consistently greater than the significance level, and the CMI test is extremely conservative. In addition, we can see the result under partial correlation method is a slightly better than the result under CDC method in example 1 because X, Y are linearly independent conditioning on Z. CDC method outperforms the other two methods for example 2, because it is a nonlinearly independent case and the type I error under CDC test is closer to 0.05 compared with partial correlation test. In example 3, partial correlation method is preferred because we already know it is a linearly independent case and type I error under partial correlation method is better than its result under CDC method.

As CMI method is conservative in all cases, we increase the sample size from 50 to 300 for possible improvement.

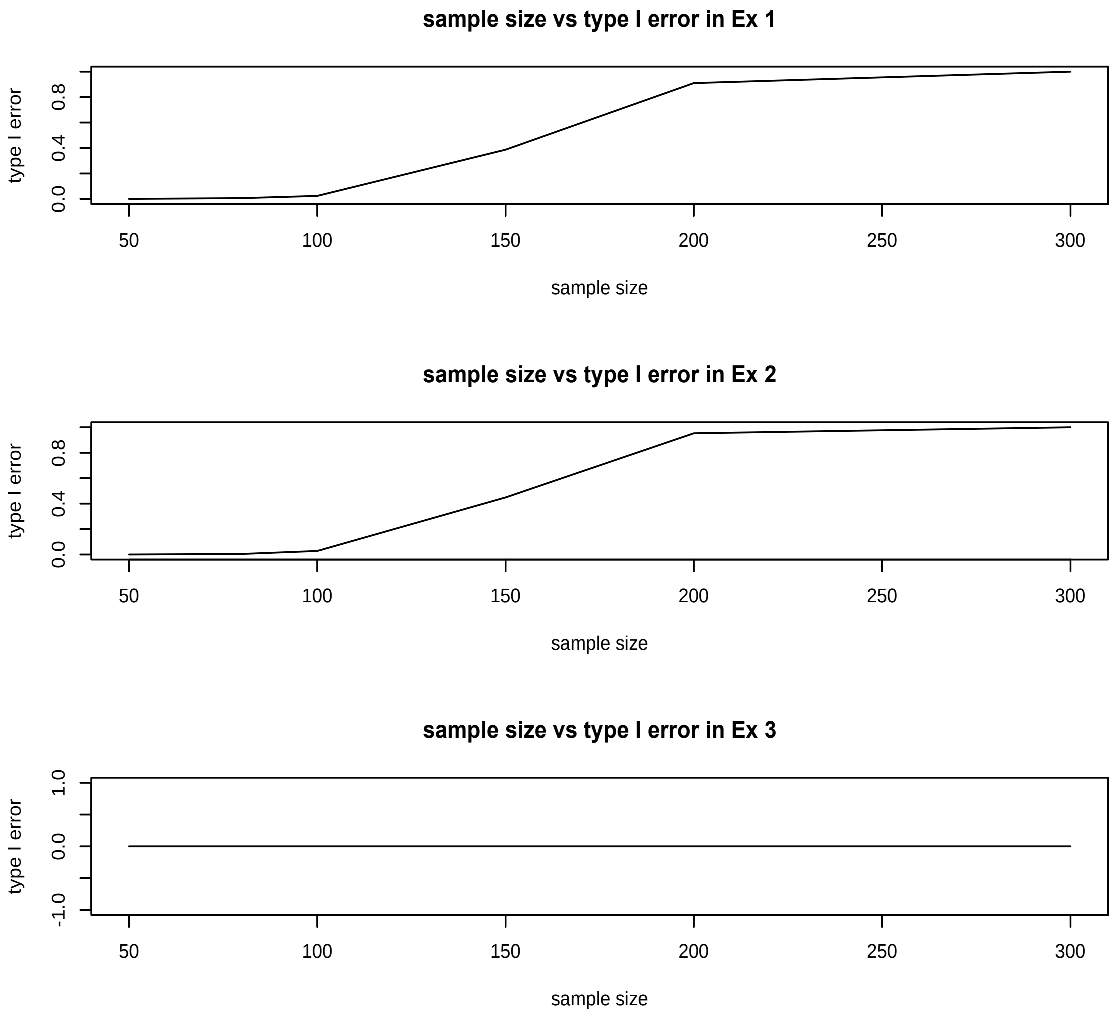

Figure 1. Sample size versus type I error rate for examples 1-3 with CMI method (nbins= $n^{\frac{1}{3}}$).

From Figure 1, we can see that sample size and type I error are positive correlated and the test is more powerful when the sample size is around 100. The type I error increases sharply as the sample size increases from 100 to 200. Once the sample size increases to 200, the type I error becomes one, which means all p-values are less than the significance level and the test severely underestimated p-values. However, type I error is always zero for all sample sizes in example 3. The CMI measure is problematic, and one possible reason is that CMI method is not well defined for continuous variables, and discretizing continuous data will more or less cause loss of information.

In the implementation of CMI, the “infotheo” package relies on the choice of bin size (argument “nbins” in function “discretize”) in order to discretize continuous data. Here, we investigate the effect of nbins on the type I error rate and statistical power.

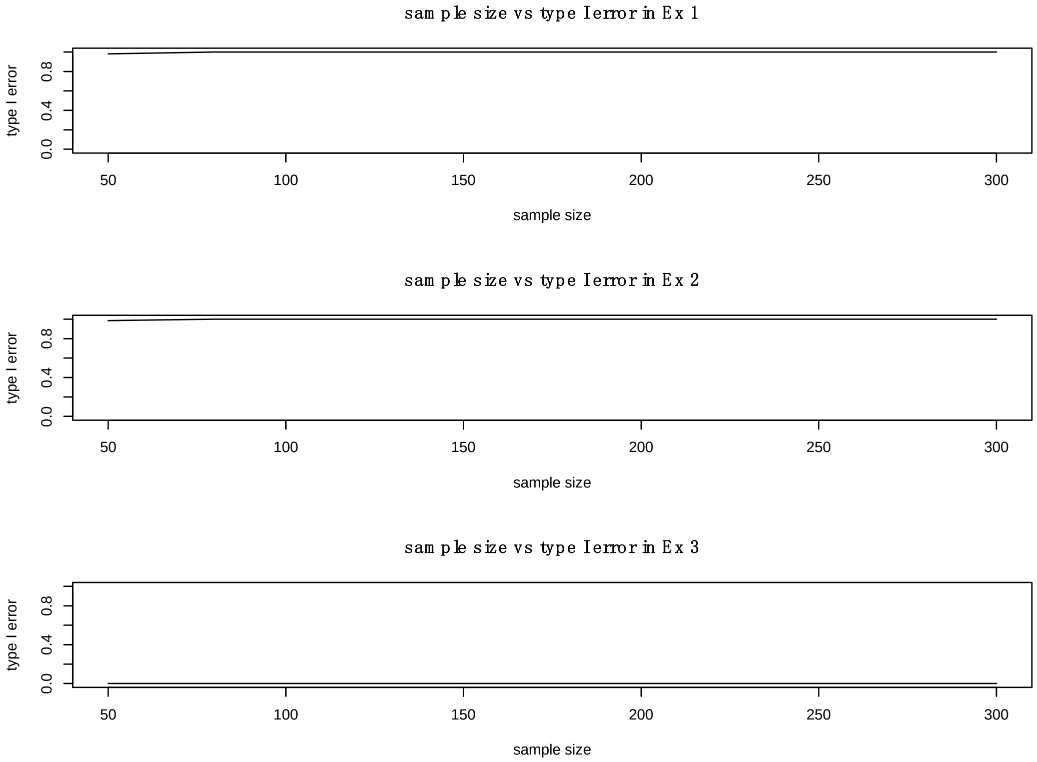

Figure 2. Sample size versus type I error rate for examples 1-3 with CMI method (nbins=n^1/2).

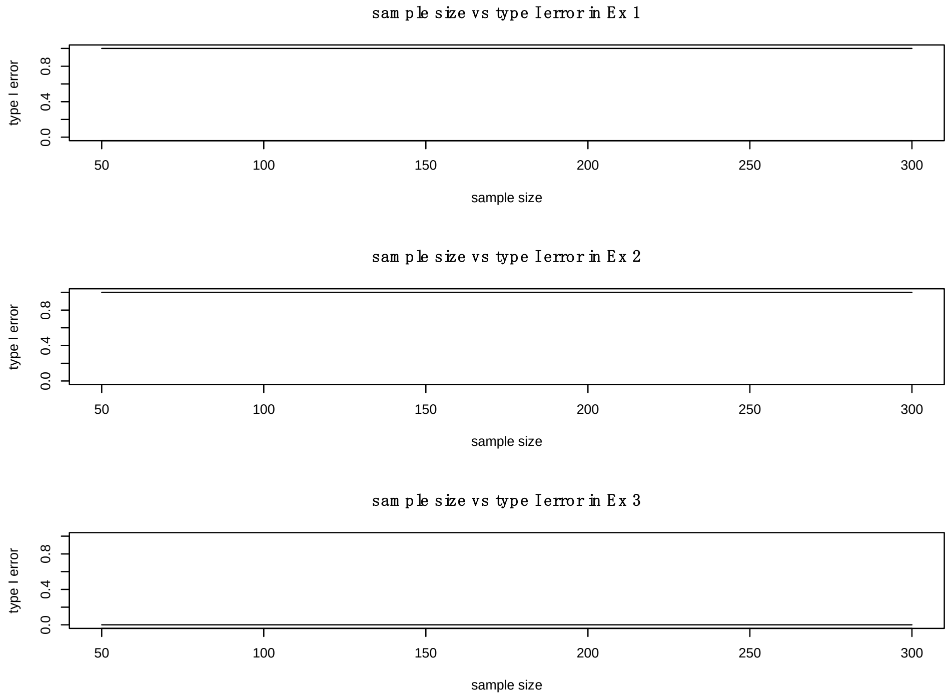

Figure 3. Sample size versus type I error rate for examples 1-3 with CMI method (nbins=n^2/3).

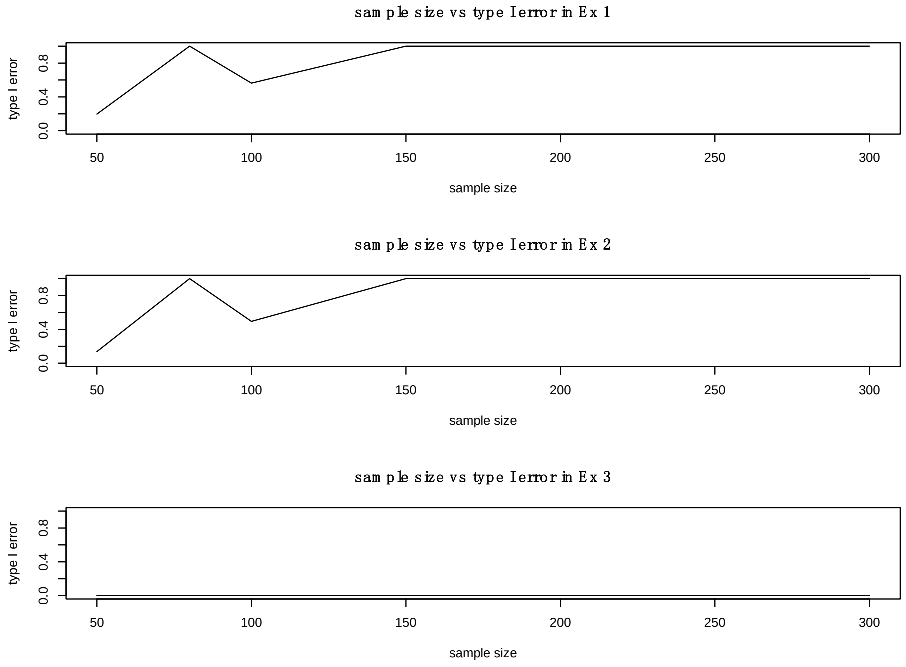

Figure 4. Sample size versus type I error rate for examples 1-3 with CMI method (nbins=n^1/4).

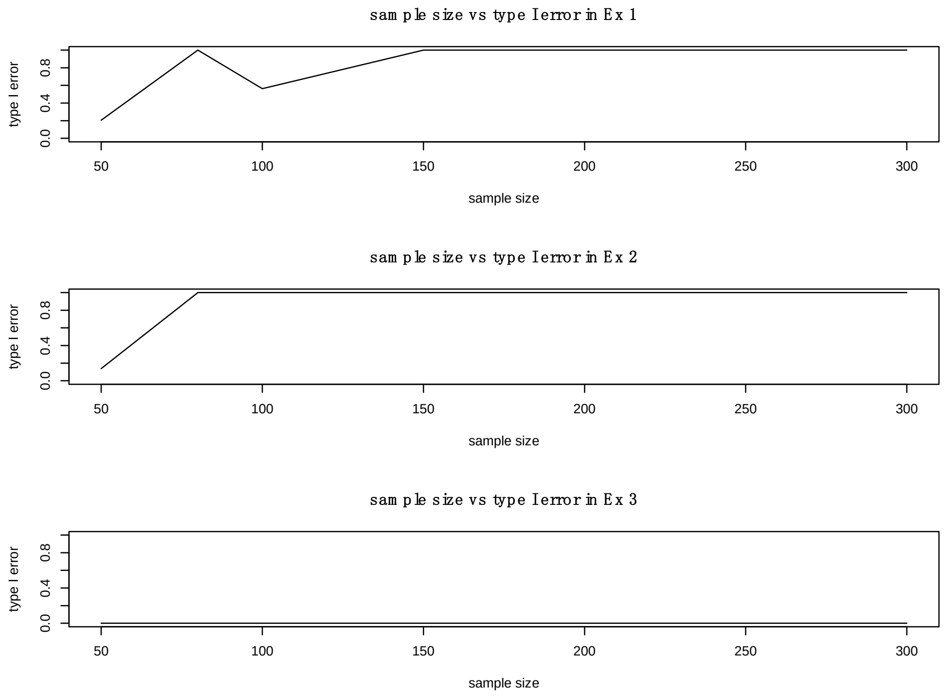

Figure 5. Sample size versus type I error rate for examples 1-3 with CMI method (nbins=n^1/5).

As we can see from Figures 2 and 3, the Type I error rates by different number of bins are comparable. Type I error rates are close to 1 for example 1 and 2 with increasing sample size, but it remains 0 for example 3. From Figure 4 and 5, we decrease the number of bins from the default 1/3, and type I error rate inflates to 1 as sample size increase for examples 1 and 2 and remains at 1 for example 3. CMI method fails to detect conditional dependency regardless of our choice for the number of bins when discretizing the continuous data. CMI may only apply to discrete variables because it will lose information when discretizing the continuous variables.

In R package “cdcsis”, the argument “index” specifies the exponent on Euclidean distance in $(0, 2]$ (default index is 1). We varied the index from 0.25 to 2 in example 1 and summarized how the type I error changes. Table 2 and Figure 6 summarized the results.

Table 2. Distance index versus type I error rate for example 1 with CDC method.

	Index									
	0.25	0.5	0.75	1	1.25	1.50	1.75	2.00	2.25	2.50
Type I Error Rate	0.08	0.08	0.04	0.04	0.06	0.16	0.04	0.12	0.06	0.08

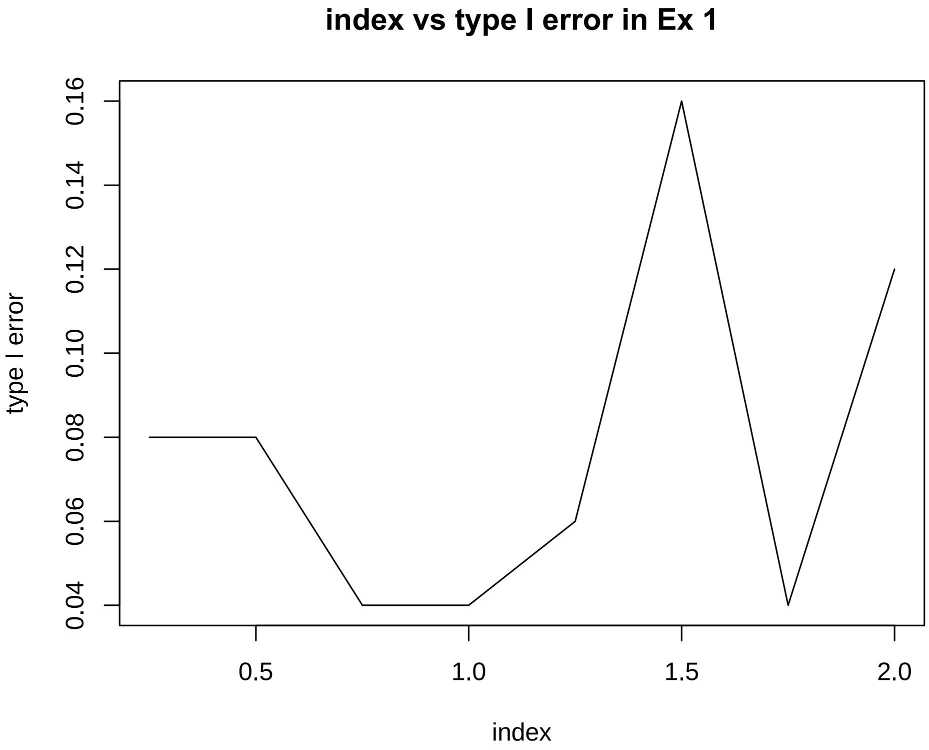

Figure 6. Distance index versus type I error for example 1 with CDC method.

From Table 2 and Figure 6, we can see that the default choice of index=1 performs overall better in example 1, thus we use index=1 for all our analyses.

3.2 Evaluation of Empirical Statistical Power

In comparison to examples 1-3, we considered another three examples for conditional dependence cases. The empirical power is summarized to compare the three methods. The statistical power is defined as the probability of rejecting a false null hypothesis (or equivalently, one minus the type II error rate). The simulation was repeated for 100 times and the power is computed as the proportion of p-values that are less than the pre-defined significance level.

Example 4. (X, Y, Z) follows a multivariate normal distribution with mean $\mu = (0,0,0)$ and covariance matrix $\Sigma = \begin{pmatrix} 1 & 0.6 & 0.5 \\ 0.6 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}$. Therefore, the conditional covariance matrix of X and Y given Z is $\Sigma(\mathrm{X},\mathrm{Y} \mid Z) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \begin{pmatrix} 0.75 & 0.35 \\ 0.35 & 0.75 \end{pmatrix}$, where the element (1,2) and (2,1) of $\Sigma(\mathrm{X},\mathrm{Y} \mid Z)$ are nonzero, indicating that X is conditionally dependent of Y given Z.

As X, Y, Z follow the multivariate normal distribution, X, Y, Z are linearly associated. In addition, X is dependent of Y given Z, and X, Y are linearly dependent given Z.

Example 5. $X_1, Y_1, Z, \varepsilon \sim N(0,1)$, and we define

$$Z_1 = 0.6 * \left(\frac{Z^3}{7} + \frac{Z}{2}\right), Z_2 = \frac{1}{4} * \left(\frac{Z^3}{2} + Z\right),$$

$$X_2 = Z_1 + \tanh(X_1), X_3 = X_2 + \frac{1}{4}(X_2)^3,$$

$$Y_2 = Z_2 + Y_1, Y_3 = Y_2 + \tanh\left(\frac{Y_2}{4}\right).$$

We standardize X_3, Y_3 and define

$$X = X_3 + cosh\varepsilon, Y = Y_3 + cosh\varepsilon^2.$$

Since X_1, Y_1, Z, ε follow normal distribution, X and Y are polynomial transformation of X_1, Y_1, Z, ε, therefore, X and Y are nonlinearly dependent given Z.

Example 6 $X_1, Z_1, Z_2 \sim Binomial(12{,}0.4)$, and we define

$$X = X_1 + Z_1 + Z_2,$$

$$Y = (X_1 - 5)^4 + Z_1 + Z_2,$$

$$Z = (Z_1, Z_2).$$

X_1, Z_1, Z_2 follow binomial distribution. X is linear transformation of X_1, Z_1, Z_2, Y is polynomial transformation of X_1, Z_1, Z_2, and Z is multivariate of Z_1, Z_2, therefore X and Y are nonlinearly dependent given Z.

Note that the variables X, Y, Z in examples 4 and 5 are all univariate variables. In example 6, X, Y are univariate, and Z is multivariate. Table 3 shows the power for examples 4-6 based on three methods.

Table 3. Empirical power for examples 4-6 with three different methods based on cutoff=0.05 (n=50).

	Example 4	Example 5	Example 6
Pearson’s Correlation	0.96	0.94	0.49
Conditional Mutual Information	0	0	0
Conditional Distance Correlation with Index=1	0.90	0.98	0.82

From Table 3, it can be seen that the CMI method fails to detect any conditional dependence, partially due to the required discretization. All empirical powers are zeros, indicating an extreme conservativeness. Partial correlation method performs the best for example 4 because the data are generated from a multivariate normal distribution and the three variables are linearly correlated. In the nonlinear cases (such as examples 5 and 6), CDC method is more powerful than the other two methods as we expected.

In examples 4-6, we use the three different significance levels, 0.05, 0.1 and 0.2 in the following tables and see how the conclusion changes.

Table 4. Empirical power for examples 4-6 with three different methods based on cutoff=0.1 (n=50).

	Example 4	Example 5	Example 6
Pearson's Correlation	0.99	0.98	0.56
Conditional Mutual Information	0.236	0.36	0
Conditional Distance Correlation with Index=1	0.96	1	0.96

Table 5. Empirical power for examples 4-6 with three different methods based on cutoff=0.2 (n=50).

	Example 4	Example 5	Example 6
Pearson's Correlation	0.99	1	0.65
Conditional Mutual Information	0.996	0.998	0
Conditional Distance Correlation with Index=1	0.98	1	0.98

From Tables 4 and 5, we can see that the advantage of CDC over other methods persist under different significance level.

In addition, we investigated the power under a larger sample size. Figure 7 shows the results for examples 4-6.

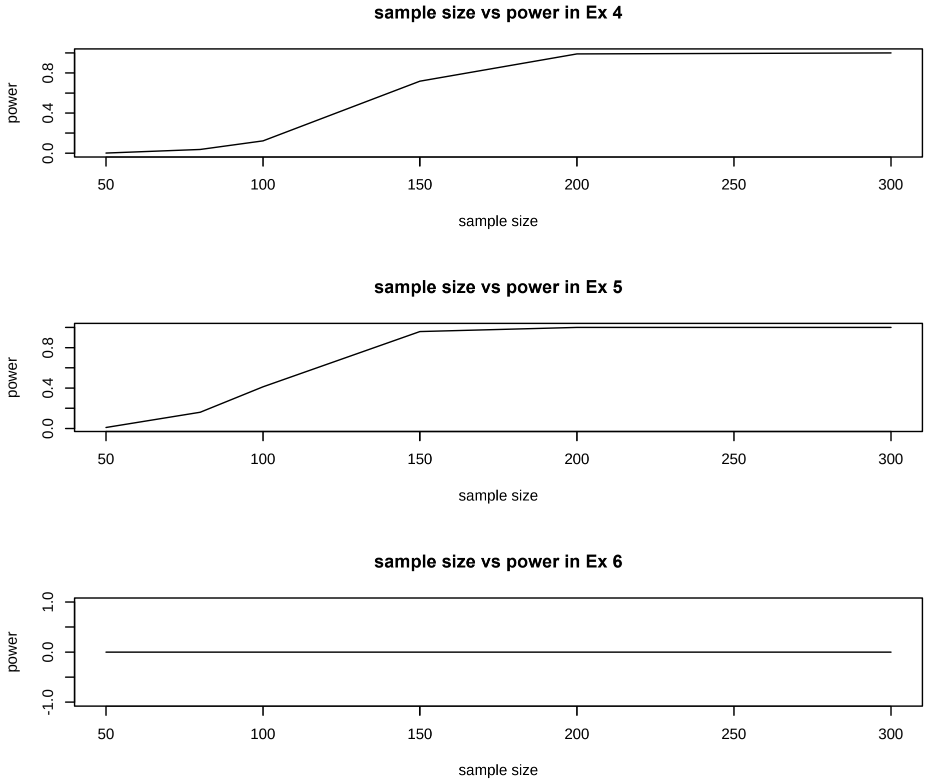

Figure 7. Sample size versus empirical power for examples 4-6 with CMI method.

Figure 7 presents the association between sample size and power for CMI method. In examples 4 and 5, the sample size is positively correlated with empirical power and the test is more powerful as the sample size increases. It shows all p-values are less than the significance level, so the CMI test is greatly underestimated p-values. However, in example 6, the power remains zero regardless of the sample size.

Our simulation studies suggested that the partial correlation test could better capture linear dependence while CDC test is more sensitive to nonlinear dependence than the two competitors.

In our real data application presented in next chapter, we will apply these two methods. As a comparison, we will pay attention to the genes detected by CDC but not by partial correlation.

4. RESULTS AND CONCLUSIONS

In this part, we apply two statistical tests, namely partial correlation test and CDC test, to study the conditional dependence between three important measurements in TCGA ovarian cancer data. Due to the high-speed development in next-generation sequencing technology, it is tremendously possible for us to perform genome-wide analysis of genetic and epigenetic features simultaneously. TCGA project characterized over 20,000 tumor and matched normal samples over 30 types of cancers (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga), which provided the most comprehensive cancer data resource.

In this analysis, we used the rich TCGA ovarian cancer data with measurements on 12,000 genes on 580 samples. Each sample was represented by three molecular measures including gene expression level, DNA methylation level and copy number variation. The samples consisted of 8 cancer-free controls, 15 early-stage (stage I) and 559 high-grade cancer samples. Among the 12,000 genes, there are several genes known to closely associated with the ovarian cancer, for instance, *TP53* (Zhang et al. (2016)), *PIK3CA* (Levine et al. (2005)), *BRCA1* and *BRCA2* (Ramus and Gayther (2009)) confer to a high life-time risk of ovarian cancer due to germline mutations [14, 25,40].

We applied CDC method to TCGA data to perform an integrated analysis between gene expression, methylation and CNV. Comparing with linear method (i.e., partial correlation method), we can infer non-linear relations based on CDC method. In addition, we can detect conditional non-linear dependence through hypothesis testing. If the p-value is above the significance level under partial correlation method, the alternative hypothesis is rejected. However, for the same gene given the same variable, the null hypothesis is rejected, and non-linear dependence is maintained if the p-value is lower than the significance level under CDC method.

We use the 245 oncogenes and tumor suppressors reported by Zhang, Burdette and Wang (2014) [38]. Out of 245 genes, there are 196 genes with the complete records of gene expression, methylation and CNV. We recorded the running time of one gene as the sample size increases from 30 to 150. Figure 8 shows the relation between sample size and the running time in R for one gene. To speed up our computation and investigate the performance of CDC under moderate sample size, for each gene, we used randomly selected 80 samples in the analysis. These 80 samples were used as a training set and we will test the robustness of our findings in the whole data set.

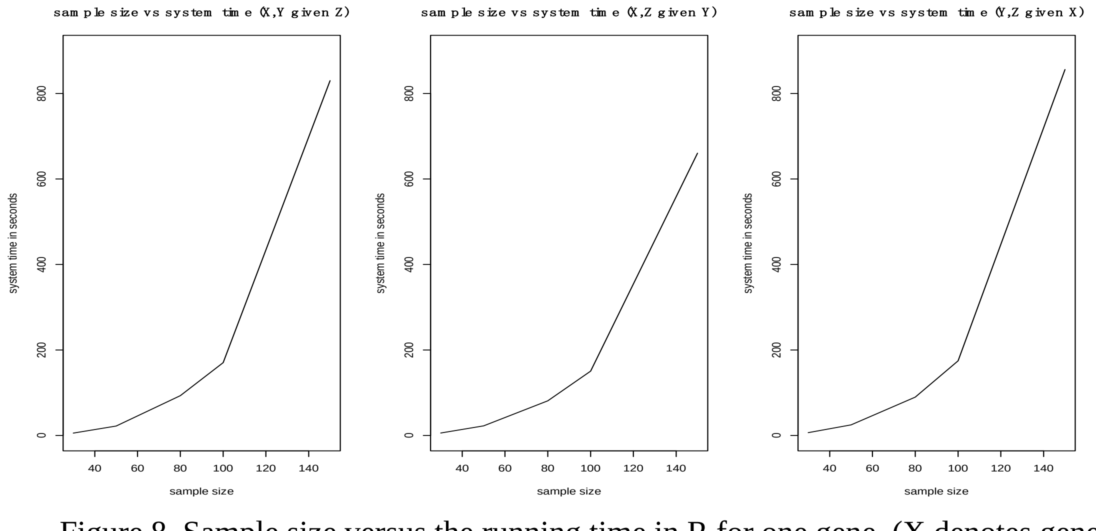

Figure 8. Sample size versus the running time in R for one gene. (X denotes gene expression, Y denotes DNA methylation, and Z denotes CNV)

To facilitate comparison, we used the same number of random samples to implement the partial correlation test. For every single gene, we obtained two p-values based on the two methods, for example, the gene *KRAS*, we did the conditional dependence test on gene expression, methylation given CNV, and found the p-value is 0.174 based on partial correlation method and the p-value is 0.002 based on CDC method. Choosing the significance level $\alpha = 0.05$ and comparing the two p-values with the significance level, we can say gene expression and methylation are linearly independent conditioning on CNV based on partial correlation method, but gene expression and methylation are nonlinearly dependent conditioning on CNV based on CDC method.

In comparison of two p-values with the same condition by the two methods clearly, I form three matrices with 196 genes (rows) and 2 p-values (columns) separately, including gene expression and methylation given CNV, gene expression and CNV given methylation, and methylation and CNV given gene expression. By choosing different cutoffs of significance level, we can detect the difference between the two methods for a few of genes conditioning on the same variable. Table 6 lists the names of genes with different decisions by two methods with different cutoffs of significance level.

Table 6, Gene selected by two methods with different significance level cutoffs.

	P-value of Pearson’s correlation method>0.15, p-pvalue of CDC method<0.02	Pvalue of Pearson’s correlation method>0.15, pvalue of CDC method <0.05	Pvalue of Pearson’s correlation method>0.10, pvalue of CDC method <0.05
X, Y given Z	"KRAS"	"DIRAS3", "MYC", "KRAS", "SNX5", "MELK"	"DIRAS3", "MYC", "KRAS", "SNX5", "MELK", "KLHL24"
X, Z given Y	"SHKBP1"	"MPHOSPH6", "SHKBP1"	"MPHOSPH6", "SHKBP1"
Y, Z given X	"PIK3CA", "ALG3", "SPEN", "MFN1", "CNP", "CROT", "MOCS3", "DPM1"	"PIK3CA", "ALG3", "HSF1", "POLR2H", "SPEN", "MFN1", "ZNF639", "CNP", "CROT", "MOCS3", "MRPL47", "DPM1"	"PIK3CA", "ALG3", "HSF1", "POLR2H", "SPEN", "MFN1", "ZNF639", "CNP", "PVRL4", "CROT", "MOCS3", "BOP1", "MRPL47", "DPM1"

where X denotes gene expression, Y denotes DNA methylation, and Z denotes CNV.

To show the difference between the two methods, we choose the second column of the table, in which the cutoff of significance level for partial correlation method is 0.15, and the cutoff of significance level for CDC method is 0.05. We found 19 genes in total that show nonlinear conditional dependence by CDC method but not detected by the partial correlation test.

In addition, we divide the 580 samples into three groups by sorting the conditioning variable in ascending order (first 200 samples with low level condition, the next 200 samples with medium level condition, the last 180 samples with high level condition) and use figure visualization to observe the relations between two default variables for each gene and 580 samples. Figure 9 - Figure 14 are the examples showing the sample clusters and variations with different levels of conditions.

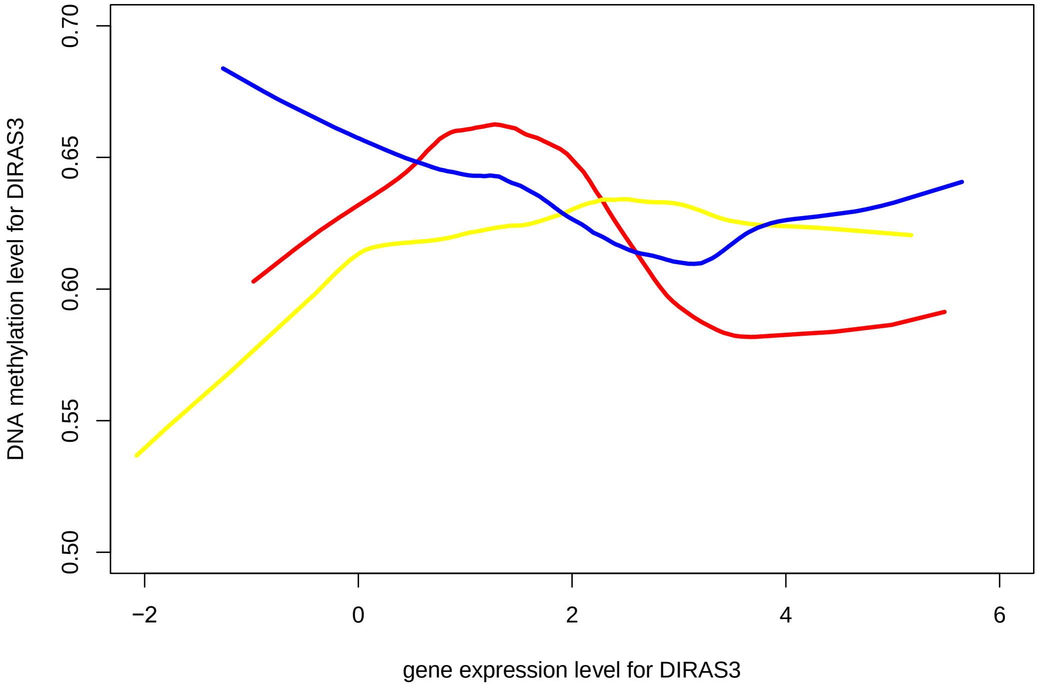

Figure 9. Gene expression level versus DNA methylation level for gene *DIRSA3* with 3 levels of CNV (red for low CNV level, yellow for medium CNV level, and blue high CNV level).

For gene *DIRSA3*, figure 9 shows the relation between gene expression and DNA methylation given CNV. For each CNV group, we fit a lowess line using smoothing parameter suggested by R. It can be seen that different CNV groups exhibited differential co-expression between gene expression and DNA methylation, with different non-linear patterns.

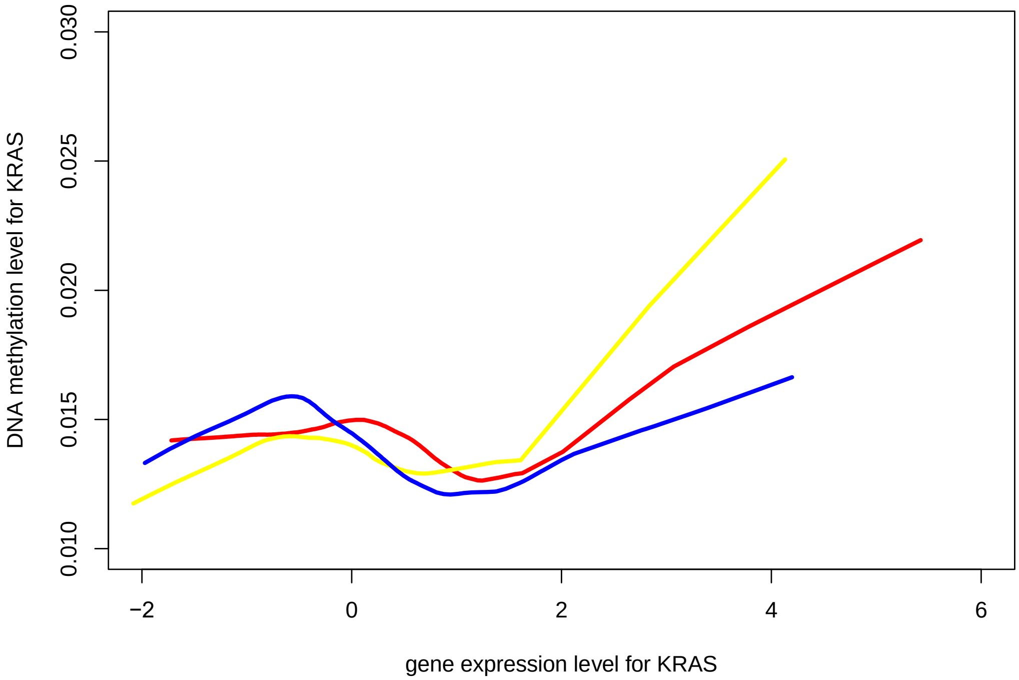

Figure 10. Gene expression level versus DNA methylation level for gene *KRAS* with 3 levels of CNV (red for low CNV level, yellow for medium CNV level, and blue for high CNV level).

Figure 10 shows the trend variation between gene expression and methylation conditioning on increasing level of CNV for gene *KRAS*. For each CNV group, a lowess line is fitted with smoothing parameter chosen by R. It shows different CNV groups displayed similar relations between gene expression level and DNA methylation in different non-linear patterns, but the trend increases sharply at the end under the medium level of CNV.

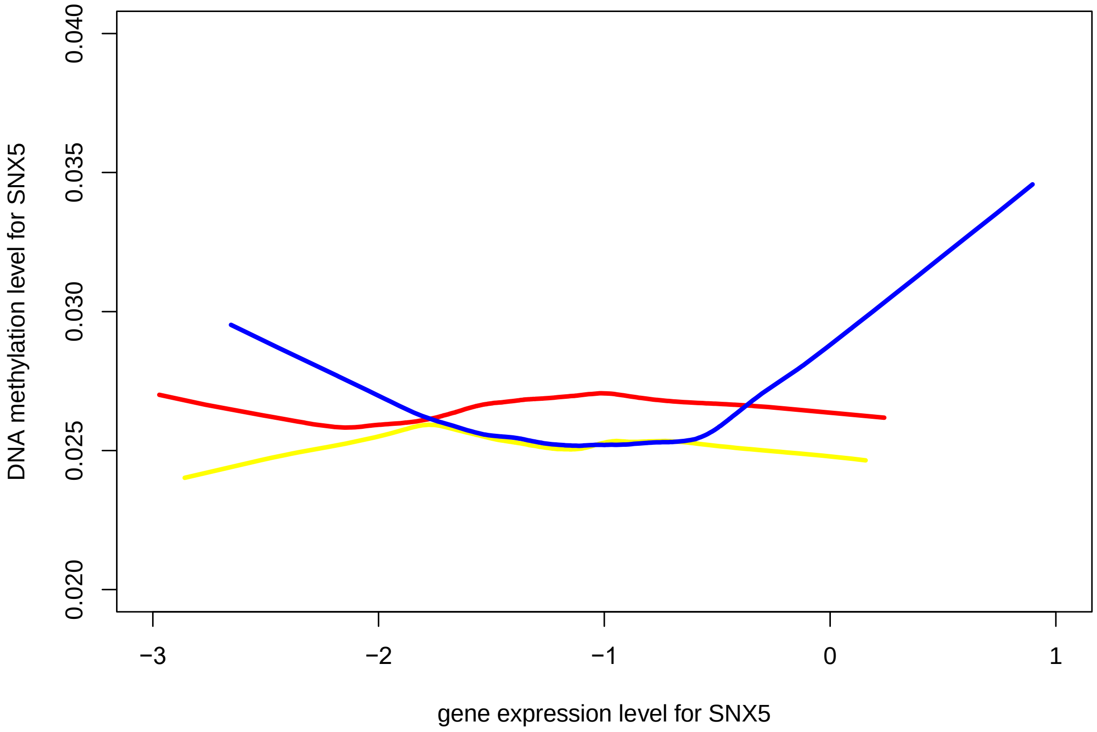

Figure 11. Gene expression level versus DNA methylation level for gene *SNX5* with 3 levels of CNV (red for low CNV level, yellow for medium CNV level, and blue for high CNV level).

For gene *SNX5*, figure 11 shows the trend variation between gene expression and methylation conditioning on increasing level of CNV. A lowess line is fitted with smoothing parameter chosen by R for each CNV group. For the low level and medium level of CNV, the trends are similar, and both different from the group with the high level of CNV.

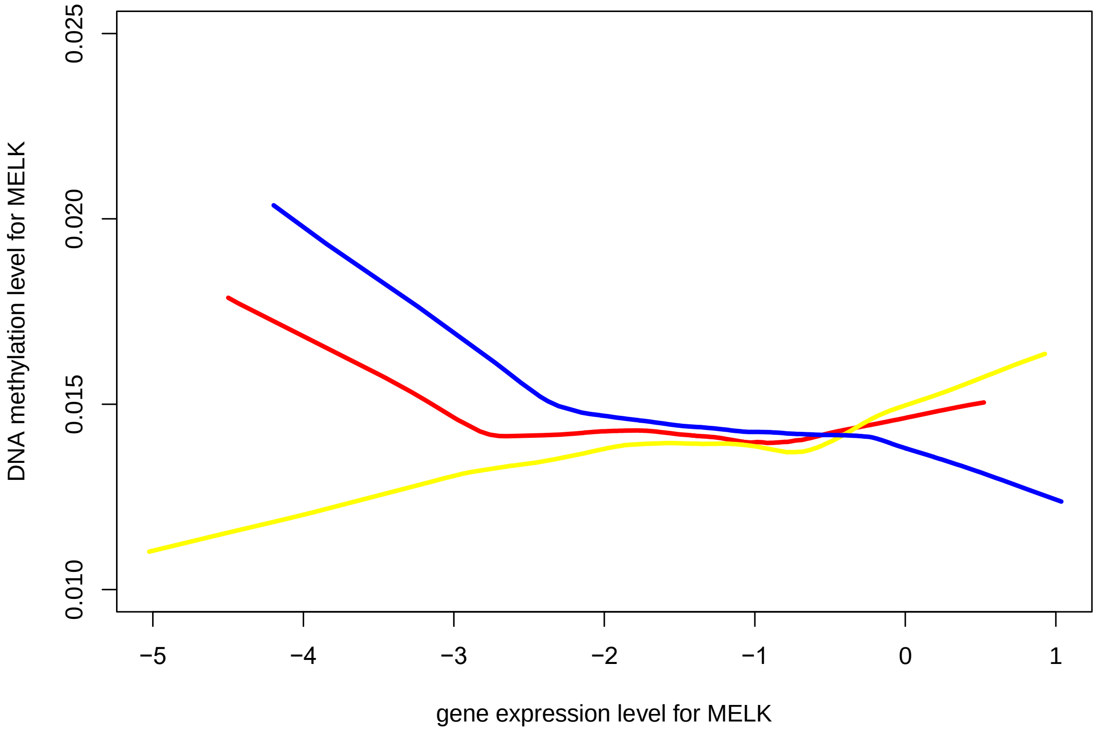

Figure 12. Gene expression level versus DNA methylation level for gene *MELK* with 3 levels of CNV (red for low CNV level, yellow for medium CNV level, and blue for high CNV level).

Figure 12 shows the relation between gene expression and methylation conditioning an increasing order of CNV for gene *MELK*. For each CNV group, a “lowess” function using smoothing parameter suggested by R is used to draw the trend variation. With different non-linear patterns, the trends for the low level of CNV (red line) and high level of CNV (blue line) are similar, which decreases slowly and then keeps constant. However, the trend seems like increases slowly based on medium CNV level (yellow line).

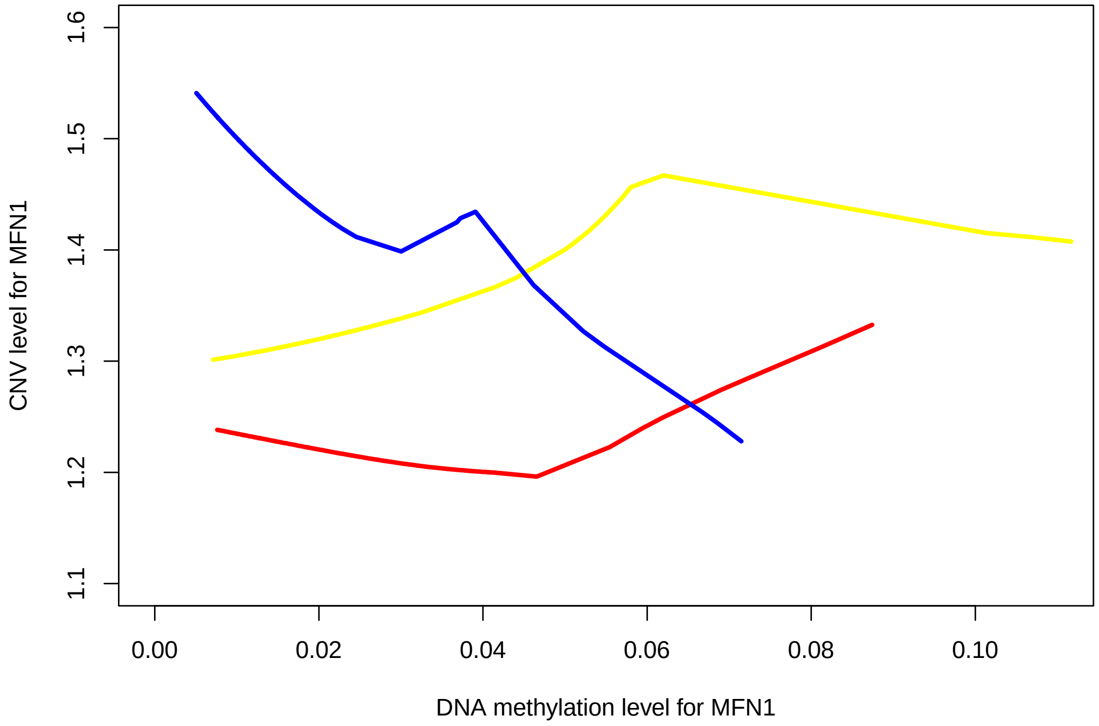

Figure 13, DNA methylation level versus CNV level for gene *MFN1* with 3 levels of gene expression (red for low gene expression level, yellow for medium gene expression level, and blue for high gene expression level).

For gene *MFN1*, figure 13 shows the relationship between methylation and CNV conditioning an increasing level of gene expression. We fit a smoothing lowess line suggested by R to draw the trend. In this figure, it shows the completely different co-expressions between DNA methylation level and CNV level based on diverse gene expression levels with different non-linear patterns. The red line keeps the same and then increases on low gene expression level. The yellow line increases and then almost remains the same on medium gene expression level. The blue line decreases and keep the same for a while and then decreases again on high gene expression level.

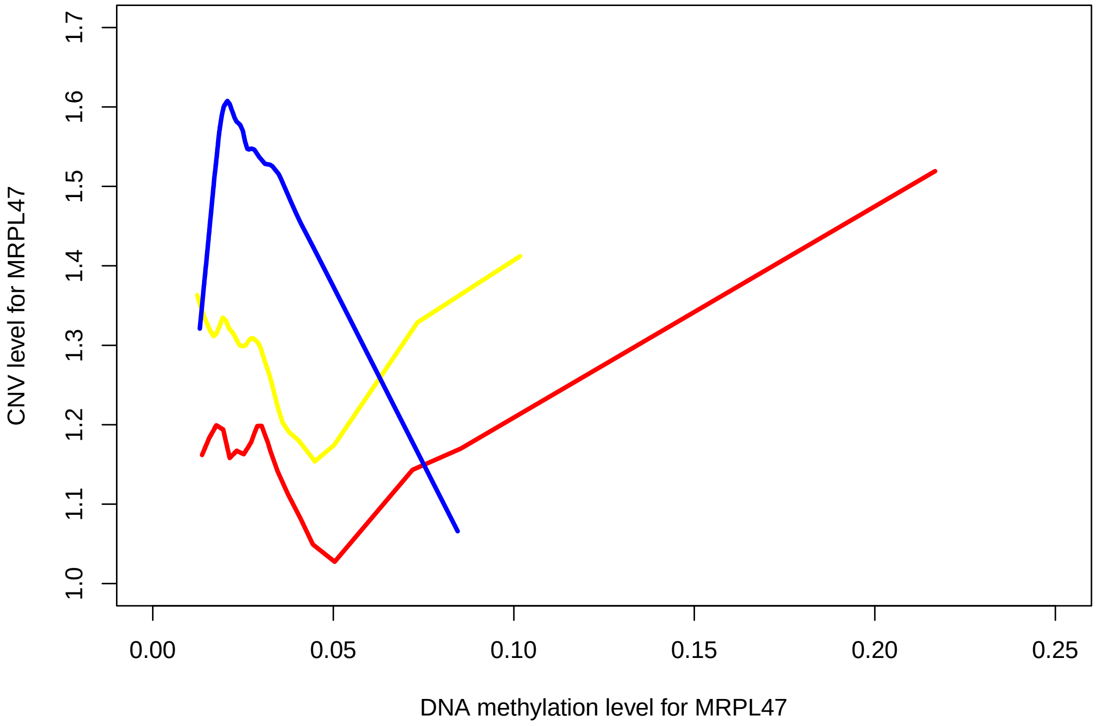

Figure 14. DNA methylation level versus CNV level for gene *MRPL47* with 3 levels of gene expression (red for low gene expression level, yellow for medium gene expression level, and blue for high gene expression level).

Figure 14 shows the relationship between DNA methylation and CNV given an increasing order of gene expression for gene *MRPL47*. In this figure, a "lowess" line with default setting suggested by R is implemented for each gene expression group. It shows different gene expression group exhibited differential co-expression between DNA methylation and CNV based on different non-linear patterns. The trends of red line and yellow line are similar based on low gene expression level and medium gene expression level, which decrease first and then increase. The trend of blue line based on high gene expression level is completely opposite to the red line and yellow line, which increases first and then decreases.

Out of 196 oncogenes or tumor suppressors related to cancers, we detected 19 genes in which the two variables are nonlinearly dependent given the third variable by CDC method but linearly independent by partial correlation method. In fact, over the past years, many of the existing studies have been made to point out these genes are related to multiple types of cancers by many researchers. To name a few, gene *KRAS* is one of the particular genes in these studies. It provides the instruction to make protein involving in signaling pathways that control cell growth, cell maturation and cell death [19, 9]. DNA mutation of KRAS could be found in non-small cell lung cancer, colorectal cancer and pancreatic cancer [19]. Molnár et al. (2018) identified the mutation hot-spot areas of frequently mutated gene KRAS showed DNA methylation alterations in colorectal cancer [18]. They also proposed the change of promoter DNA methylation could affect mRNA expression level in special cases.

The gene *DIRAS3* is a protein encoding gene that is linked to breast cancer and ovarian cancer. Barrow et al. (2015) found *DIRAS3* is one of the imprinted tumor suppressor genes that its aberrant frequent methylation is associated with negative hormone receptor status in invasive breast cancer [1]. Novak et al. (2017) found that for genes *DIRAS3* and *STAT3*, DNA methyltransferase inhibitors have effect on gene expression and DNA methylation for gene *DIRAS3* in ovarian and breast carcinomas [21].

The gene *MFN1* is also a protein encoding gene that could result in non-small cell lung carcinoma. Two major categories include lung adenocarcinoma and lung squamous cell carcinoma. Qiu et al. (2017) analyzed the copy number variation pattern for the two subtypes with 33 signature genes including *MFN1* and developed an accurate CNV classifier to distinguish lung adenocarcinoma from lung squamous cell carcinoma for non-small cell lung carcinoma [24].

Based on the discussion above, we present a novel CDC test for inferring nonlinear conditional dependency between gene expression, DNA methylation and CNV for TCGA data in contrast with partial correlation method and identify a bunch of genes with nonlinear dependent relations that related to cancers. The figures 9-14 show clear relationships between two measurements given the third. For instance, the gene expression and DNA methylation of gene *KRAS* are nonlinearly conditional dependent on three levels of CNV. We are still in the process of detecting genes with nonlinear relationships between genetic features and expanding to multiple types of cancers.

5. SUMMARY AND FUTURE WORK

5.1 Summary

In this thesis, we proposed to use a recently developed measure, namely conditional distance correlation to detect nonlinear dependence between gene expression, DNA methylation and CNV for a set of oncogenes and tumor suppressors related to ovarian cancer using TCGA data as the traditional partial correlation may fail to detect nonlinear dependence. To reduce the dimensionality and improve the efficiency of CDC test, we only consider 196 genes that have been reported in literature to closely related to ovarian cancer and randomly selected 80 samples out of 580 samples. We performed hypothesis testing based on partial correlation method and CDC method and found a list of 19 genes with conditionally nonlinear dependence among the three features. For each of the 19 genes, we divided the 580 samples into 3 groups including low expression group, medium expression group and high expression group and illustrated the nonlinear difference by some example genes. Some of the identified 19 genes have been reported to be associated with multiple cancer procedures (e.g. DNA mutation or protein encoding) and they could extend the common function to multiple types of cancers.

5.2 Future Work

In this part, we discuss some limitations of the CDC test with future perspectives. First, in the real data analysis, we used a random sample of size 80 for two considerations: (1) save computing time (2) test the robustness of CDC under a moderate sample size. To use hundreds or even thousands of samples in a CDC test, we need rely on parallel computing. Second, although we have identified several genes with nonlinear conditional dependency among three variables and the results have demonstrated the CDC method is promising, a possible future direction is to extend the test for a single gene to a gene set or a biological pathway, in which the member genes share similar biological function. Biologically, it would be more interesting to test multiple genes together as it may provide functional insights on the underlying mechanism. In the context of gene set testing, the variables would be three vectors (gene expression, DNA methylation and CNV) for multiple genes, resulting in an integrative dataset. Moreover, CDC method could be applied to other problems or areas to study the conditional association between three or more random variables. For example, in economics, gross domestic product (GDP) is related to population and people's standard of living in economic aspect. We could test conditional dependency for any of two variables given the third variable based on CDC method. In epidemiology, Zika virus spreads by mosquito and it is epidemic in Central & South America and the Caribbean. Tosepu (2017) have presented ambient humidity is one of the influence factors for Zika virus infection [33]. Tesla et al. (2018) have demonstrated the temperature is another strong driver leading to Zika virus transmission [31]. For the three variables (Zika virus, humidity and temperature), we could use CDC method to analyze whether there exists nonlinear conditional dependence between any two variables conditioning on the last variable.

6. REFERENCES

1. Barrow, T. M., Barault, L., Ellsworth, R. E., Harris, H. R., Binder, A. M., Valente, A. L., . . . Michels, K. B. (2015). Aberrant methylation of imprinted genes is associated with negative hormone receptor status in invasive breast cancer. *International Journal of Cancer, 137*(3), 537-547.

2. Bühlmann, P., Kalisch,M., & Maathuis, M.H. (2010). Variable selection in high-dimensional linear models: Partially faithful distributions and the pc -simple algorithm. *Biometrika*, 97(2), 261-278.

3. Cancer Genome Atlas Research Network. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, *474*(7353), 609-615.

4. Freudenberg, J., Wang, M., Yang, Y., & Li, W. (2009). Partial correlation analysis indicates causal relationships between GC-content, exon density and recombination rate in the human genome. *BMC Bioinformatics, 10*(Suppl 1).

5. Gao, W., & Zhao, H. (2013). Conditional independence graph for nonlinear time series and its application to international financial markets. *Physica A: Statistical Mechanics and Its Applications, 392*(10), 2460-2469.

6. Higgins, J. J. (2004). *An introduction to modern nonparametric statistics*. Belmont, CA: Brooks-Cole.

7. Hu, W., Huang, M., Pan, W., Wang, X., Wen, C., Tian, Y., Zhang, H., & Zhu, J. (2019) cdcsis: Conditional Distance Correlation Based Feature Screening and Conditional Independence Inference. R package version: 2.0.2. http://cran.r-project.org/package=cdcsis

8. Huang, T. (2010). Testing conditional independence using maximal nonlinear conditional correlation. *The Annals of Statistics, 38*(4), 2047-2091.

9. Genetics Homes Reference. (2019). KRAS gene - Genetics Home Reference - NIH. https://ghr.nlm.nih.gov/gene/KRAS

10. Kalisch, M., & Bühlmann, P. (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.*, 8, 613-636

11. Kim, S. (2015). Ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. *Communications for Statistical Applications and Methods, 22*(6), 665-674.

12. Kim, S. (2015). ppcor: Partial and Semi-Partial (Part) Correlation. R package version 1.1. https://cran.r-project.org/package=ppcor

13. Kulis, M., & Esteller, M. (2010). DNA methylation and cancer.

14. Levine, D. A., Bogomolniy, F., Yee, C. J., Lash, A., Barakat, R. R., Borgen, P.I., & Boyd, J. (2005). Frequent Mutation of the PIK3CA Gene in Ovarian and Breast Cancers. *Clinical Cancer Research, 11*(8), 2875-2878.

15. Li, R., Liu, J., & Lou. L. (2017). Variable Selection via Partial Correlation. *Statistica Sinica*, 27(3), 983-996

16. Li, X., Liu, C., Huang, T., & Zhong, Y. (2016). The Occurrence of Genetic Alterations during the Progression of Breast Carcinoma. *BioMed Research International, 2016*, 1-5.

17. Meyer, P. E., (2014). infotheo: Information-Theoretic Measures. R package version: 1.2.0. http://cran.r-project.org/package=infotheo

18. Molnár, B., Galamb, O., Péterfia, B., Wichmann, B., Csabai, I., Bodor, A., . . . Tulassay, Z. (2018). Gene promoter and exon DNA methylation changes in colon cancer development – mRNA expression and tumor mutation alterations. *BMC Cancer, 18*(1).

19. National Cancer Institute. NCI Dictionary of Cancer Terms. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/kras-gene

20. National Cancer Institute. (2018). The Cancer Genome Atlas. https://tcga-data.nci.nih.gov/docs/publications/tcga/about.html

21. Nowak, E. M., Poczęta, M., Bieg, D., & Bednarek, I. (2017). DNA methyltransferase inhibitors influence on the DIRAS3 and STAT3 expression and in vitro migration of ovarian and breast cancer cells. *Ginekologia Polska, 88*(10), 543-551.

22. Ovarian Cancer Research Alliance. (2018). Ovarian Cancer Statistics. https://ocrahope.org/patients/about-ovarian-cancer/statistics/

23. Poli, D., Pastore, V. P., Martinoia, S., & Massobrio, P. (2014). Partial correlation analysis for functional connectivity studies in cortical networks. *BMC Neuroscience*,15(S1).

24. Qiu, Z., Bi, J., Gazdar, A. F., & Song, K. (2017). Genome-wide copy number variation pattern analysis and a classification signature for non-small cell lung cancer. *Genes, Chromosomes and Cancer*, *56*(7), 559-569.

25. Ramus, S. J., & Gayther, S. A. (2009). The Contribution of BRCA1 and BRCA2 to Ovarian Cancer. *Molecular Oncology, 3*(2), 138-150.

26. Saito, S., Zhou, X., Bae, T., Kim, S., & Horimoto, K. (2011). A procedure for identifying master regulators in conjunction with network screening and inference. *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*.

27. Spectrum. Copy number variation | Spectrum | Autism Research News. https://www.spectrumnews.org/wiki/copy-number-variation/

28. Su, L., & White, H. (2007). A consistent characteristic function-based test for conditional independence. *Journal of Econometrics, 141*(2), 807-834.

29. Su, L., & White, H. (2008). A Nonparametric Hellinger Metric Test For Conditional Independence. *Econometric Theory, 24*(4), 829-864.

30. Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics, 35*(6), 2769-2794.

31. Tesla, B., Demakovsky, L. R., Mordecai, E. A., Ryan, S. J., Bonds, M. H., Ngonghala, C. N., . . . Murdock, C. C. (2018). Temperature drives Zika virus transmission: Evidence from empirical and mathematical models. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1884), 20180795

32. The American Cancer Society medical and editorial content team (2018). Key Statistics for Ovarian Cancer. https://www.cancer.org/cancer/ovarian-cancer/about/key-statistics.html

33. Tosepu, R. (2017). Humidity is an ambient parameter to development of Zika virus: An Indonesian case. *African Health Sciences, 17*(2), 597.

34. Vargha, A., Bergman, L. R., & Delaney, H. D. (2012). Interpretation problems of the partial correlation with nonnormally distributed variables. *Quality & Quantity, 47*(6), 3391-3402.

35. Wang, X., Pan, W., Hu, W., Tian, Y., & Zhang, H. (2015). Conditional Distance Correlation. *Journal of the American Statistical Association, 110*(512), 1726-1734.

36. Xu, Y., Zhang, J., Yuan, Y., Mitra, R., Muller, P., & Ji, Y. (2012). A Bayesian graphical model for integrative analysis of TCGA data. *Proceedings 2012 IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS)*.

37. Yourgenome. (2016) What is gene expression? https://www.yourgenome.org/facts/what-is-gene-expression

38. Zhang, Q., Burdette, J. E., & Wang, J. (2014). Integrative network analysis of TCGA data for ovarian cancer. *BMC Systems Biology*, *8*(1).

39. Zhang, X., Zhao, X., He, K., Lu, L., Cao, Y., Liu, J., . . . Chen, L. (2011). Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics*, *28*(1), 98-104.

40. Zhang, Y., Cao, L., Nguyen, D., & Lu, H. (2016). TP53 mutations in epithelial ovarian cancer. *Translational Cancer Research*, *5*(6), 650-663.