8-2018

Comparison of Correlation, Partial Correlation, and Conditional Mutual Information for Interaction Effects Screening in Generalized Linear Models

Ji Li
University of Arkansas, Fayetteville

Comparison of Correlation, Partial Correlation, and Conditional Mutual Information for Interaction Effects Screening in Generalized Linear Models

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

by

Ji Li
East China University of Science and Technology
Bachelor of Science in Food Quality and Safety, 2008

August 2018
University of Arkansas

This thesis is approved for recommendation to the Graduate Council.

_____
Qingyang Zhang, PhD
Thesis Director

_____
Jyotishka Datta, PhD
Committee Member

_____
Avishek Chakraborty, PhD
Committee Member

ABSTRACT

Numerous screening techniques have been developed in recent years for genome-wide association studies (GWASs) (Moore et al., 2010). In this thesis, a novel model-free screening method was developed and validated by an extensive simulation study. Many screening methods were mainly focused on main effects, while very few studies considered the models containing both main effects and interaction effects. In this work, the interaction effects were fully considered and three different methods (Pearson's Correlation Coefficient, Partial Correlation, and Conditional Mutual Information) were tested and their prediction accuracies were compared.

Pearson's Correlation Coefficient method, which is a direct interaction screening (DIS) procedure, tended to incorrectly screen interaction terms as it omits the relationship between main effects and interaction effects. To this end, we proposed to use two new interaction screening procedures, namely Partial Correlation Interaction Screening (PCIS) method and Conditional Mutual Information Interaction Screening (CMIIS) method. The Partial Correlation (PC) could measure association between two variables, while adjusting the effect of one or more extra variables. The Conditional Mutual Information (CMI) is the expected value of the mutual information (MI) of two random variables given the value of a third (Wyner, 1978), while MI is a measure of general dependence. Finally, an illustration and performance comparison of the three screening procedures by simulation studies were made and these procedures were applied to real gene data.

Key words: Interaction effects, Pearson's correlation coefficient, Partial correlation, Conditional mutual information, Numerous screening

ACKNOWLEDGEMENTS

Firstly, I would like to thank God for His unlimited love and blessings to me and helping me recover from illness. Only through Him, I can overcome all odds and difficulties and complete my thesis.

Secondly, I would like to express my sincere appreciation to my thesis advisor, Dr. Qingyang Zhang, who has always given me help and advice with patience, encouragement, and understanding, especially when I was sick, he made allowance for my difficulties. I am truly honored and fortunate to be his student.

I also want to thank the remaining members in my thesis committee: Dr. Avishek Chakraborty and Jyotishka Datta, for instructing me with their proficiency during the last two years. I learned the importance of statistics.

Finally, I would like to thank my family for all their love and encouragement throughout these years. To my husband, Xuan Gu, for his unlimited support and love. To my mother, Mingming Ji, for her unconditional help and love.

Special thanks are extended to the staff of the Department of Mathematical Science and the University of Arkansas Graduate School for all their help with theses and dissertations.

TABLE OF CONTENTS

1. INTRODUCTION

Since the screening methods focusing on models which contain interaction terms are limited, we developed some new measurements in this work to better select interaction terms.

Our main interest is to evaluate the importance of interaction terms in generalized linear model, and to improve the accuracy rate of interaction screening is the main goal in our project. For this purpose, literatures of different screening methods were reviewed and a direct interaction screening (DIS) procedure was finally chosen to compare with the other two proposed approaches: Partial Correlation Interaction Screening (PCIS) and Conditional Mutual Information Interaction Screening (CMIIS).

As reported by the review of Moore et al. (2010), computational methods were the certain trend of bioinformatics. Numerous screening techniques have been discussed and developed today for genome-wide association studies (GWASs) (Moore et al., 2010). However, many screening methods mainly focused on models that only have main effects. The model, which contains both main effects and interaction effects, will result in a more accurate prediction, since the interaction terms could remedy the main-effect-only model's limitation and interpret the data more informatively.

There have been many indications in the literature on the important of interaction terms. For instance, Nelder (1994) pointed out that the selection of interaction terms must be taken into account on inference from the fitting of linear models; McCullagh (2002) also found interaction terms had influence on Box-Cox type statistical models. All those early studies indicated that interaction terms needed to be further studied. Cordell (2009) and Van Steen (2011) paid more attention to interaction effects in genetic association analysis which made valuable contributions to GWASs and overviewed recent interaction effects selection methods for high dimensional

gene data. Researchers are currently interested in detecting important interactions between genetic loci and understanding their influences on human genetic disease; however, interaction terms are very difficult to select and analyze especially in high dimensional data. Finding an efficient and powerful interaction screening method becomes very important and urgent.

There has been a vast volume of statistical literature on interaction screening. For instance, Wu et al. (2009) proposed a two-stage strategy in the framework of lasso penalized logistic regression for gene mapping. In the first stage, they aimed on marginal predictors; in the second stage, they focused on interaction predictors. Their method has a good balance between model completeness and computational speed. However, this method has some handicaps. For example, it might overlook weaker signals due to dominance of strong signals and it might have some difficulty dealing with high correlated predictors since the interaction effects cannot be easily found. Moreover, Wu et al. (2010) applied the same strategy in their new procedure called screen and clean (SC), which is a model selection tool for high-dimensional regression for identifying reliable loci and interactions. However, Bien et al. (2015) pointed out that the two-stage strategy would have drawbacks in some cases. There has been little consensus about how to threshold the main effects and interaction effects to determine the threshold for the main effect, which might depend on the strength of the interactions.

Another popular strategy is to fit a model containing both main and interaction effects together with different penalty constraints; for instance, some models allow an interaction only if the corresponding main effects are also in the model (Bien et al., 2013). The penalty constraints are known through various names, such as “heredity,” “marginality,” and “hierarchical clustering/ selection/ testing” (Chipman, 1996; Choi et al., 2010; Nelder, 1977; McCullagh and Neider, 1989; Neider, 1994; Park and Hastie, 2007; Zhao et al., 2009; Bien et al., 2013).

However, these methods are infeasible in a high-dimensional setting due to the prohibitively computational cost. For instance, Fan and Lv (2008) introduced Sure Independence Screening (SIS), a sure screening model selection method, which is based on correlation learning. Sure screening property ("all the important variables survive after variable screening with probability tending to 1") guarantees that no important variables would be screened out after getting through a variable screening procedure with probability tending to 1.

Consider a linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\mathbf{y} = (Y_1, \dots, Y_n)^{\mathrm{T}}$ is a n-dimensional vector of response, and $\mathbf{X} = (X_1, \dots, X_n)^{\mathrm{T}}$ is a $n \times p$ random matrix, which are independent and identically distributed (IID), $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^{\mathrm{T}}$ is a p-dimensional vector of parameters, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^{\mathrm{T}}$ is a n-dimensional vector of IID random errors. Then each input variable is centered so that the observed mean is 0, and each predictor is scaled so that the sample standard deviation is 1. Denote $\boldsymbol{\omega} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$, where the $\mathbf{X}$ is the first standardized $n \times p$ matrix as mentioned before. Hence, $\boldsymbol{\omega}$ is indeed a vector of marginal correlation coefficients of response with predictors, rescaled by the standard deviation of the response. Therefore, the marginal correlation coefficient $Corr(Y, X_m)$ is proportional to $\omega_m = X_m^T \boldsymbol{y}$. By ranking the component-wise magnitudes of $\boldsymbol{\omega}$ in descending order, we define a sub-model

$$\mathcal{M}_\alpha = \{1 \ll m \ll p: |\omega_m| \text{ is among the first } [\alpha n] \text{ largest of all}\}, \tag{1.2}$$

where [ ] denotes the integer part of the unit. This is a direct and easy approach to contract full model $\{1, \dots, p\}$ to a sub-model $\mathcal{M}_\alpha$ with size $d = [\alpha n]$. This correlation learning basically sorts the marginal correlation coefficients with responses from high to low, which indicates the importance of the features. Furthermore, it can screen out those predictors with weak marginal

correlation coefficients. This is the general idea and a conservative example of SIS. Depending on the order of sample size *n*, we may choose a different size of the sub-model, such as, *n* - 1 or *n*/log(*n*) (Fan and Lv, 2008; Niu et al., 2018).

Although SIS mainly focuses on main effects, we can still apply the general idea to the interaction screen, which is to keep important interaction parts, while filtering out unimportant ones. Therefore, $\mathbf{X}^{\Delta 2} = \mathbf{X} \Delta \mathbf{X}$ is defined where $\mathbf{X}^{\Delta 2}$ is a $n \times \frac{p(p+1)}{2}$ matrix and contains all pairwise product of all $\mathbf{X}$'s column vectors. Although $\mathbf{X}$ is standardized, $\mathbf{X}^{\Delta 2}$ is further column-wise standardized and $\mathbf{Z}$ denotes the column vector of standardized $\mathbf{X}^{\Delta 2}$, so that $\mathbf{Z}_{lm} = \mathbf{X}_l \Delta \mathbf{X}_m, 1 \ll l \ll m \ll p$. Therefore, an extension of the SIS would be screening interactions based on $\mathbf{\Omega} = \mathbf{Z}^{\mathrm{T}}\mathbf{y}$, where $\mathbf{\Omega}$ is a $\frac{p(p+1)}{2}$ dimensional vector. A direct interaction screening (DIS) procedure selects a model

$$\mathcal{N}_\alpha = \{1 \ll l \ll m \ll p: |\Omega_{lm}| \text{ is among the first } [\alpha n] \text{ largest of all}\}. \quad (1.3)$$

This DIS approach is essentially an interaction screening procedure based on Pearson's correlation coefficient between predictor and response variable, which has some distinct disadvantages. In particular, the relationship between main terms and interaction terms is critical which must be considered in practice, but the DIS method fails to take it into account. The details of proof will be presented in Chapter 2.

To compare with DIS, two new methods are proposed: Partial Correlation Interaction Screening (PCIS) and Conditional Mutual Information Interaction Screening (CMIIS). Our strategy is that the statistical feature of different variables is calculated and ranked in increasing order. The accuracy of the statistical feature is also evaluated and compared. To focus on the interaction effects only, the main terms and quadratic terms are not considered in this study.

To better understand the two new methods, two concepts should be introduced first. Partial Correlation (PC) is an association measurement to screen out the effect of indirect paths, and it is computed as $\rho_{X,Y|Z} = \frac{\rho_{X,Y} - \rho_{X,Z}\rho_{Y,Z}}{\sqrt{1-\rho_{X,Y}^2}\sqrt{1-\rho_{Y,Z}^2}}$, when $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ represent as random variables from a linear model (Wikipedia contributors, 2018). There is a similar concept called partial faithfulness which was introduced by Bühlmann and Kalisch (2007). They proposed a variable selection method in high-dimensional linear models to solve the problem that the number of covariates are far larger than the sample size. Not limited to PC, the Distance correlation and the Pearson's correlation coefficient have also been studied before, such as Fan and Lv (2008), Zhu et al. (2011), Li et al. (2012), and Li et al. (2012). Most of these studies focus on high or ultra-high dimensional data.

Based on the property that PC can get rid of irrelevant effect, the PCIS is introduced as a new way to measure the correlation between interactions and the response since it considers their parental main effects. It is convenient and efficient to calculate PC by this new approach, especially for high dimensional data, since it does not need to consider the interaction effects. Moreover, whether the parental main effects are strong or not, they will not influence the detection of interactions, as it does not rely on the hierarchical model assumption of two-stage screening methods (Hao and Zhang, 2014).

Another concept is Conditional Mutual Information (CMI), defined as the expected value of the mutual information (MI) of two random variables given the value of a third, where MI is a measure of general dependence and able to detect both linear and non-linear dependencies (Sotoca and Pla, 2010). Cover and Thomas (1991) introduced several essential concepts in information theory, which are very important for us to understand the CMI, such as entropy and Shannon entropy. Entropy is another key measure of information based on the information

theory. Let $\mathbf{X} = (x_1, \ldots, x_n)$ be a discrete random variable with probability mass function P($\mathbf{X}$), and the Shannon entropy is defined as: $\mathrm{H}(\mathbf{X}) = \mathrm{E}[\mathrm{I}(\mathbf{X})] = \mathrm{E}\big[-\ln\big(\mathrm{P}(\mathbf{X})\big)\big]$, where E is the expected value operator, and I is the information content of $\mathbf{X}$. I($\mathbf{X}$) is itself a random variable. $\mathrm{H}(\mathbf{X})$ can also be written as: $\mathrm{H}(\mathbf{X}) = \sum_{i=1}^{n} \mathrm{P}(x_i)\mathrm{I}(x_i) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log_b \mathrm{P}(x_i)$, where b is the base of the logarithm used. Common values of b are 2, Euler's number e, and 10, and the corresponding units of entropy are bits for b = 2, nats for b = e, and bans for b = 10 (Schneider, 2007). The conditional entropy $\mathrm{H}(\mathbf{X}|\mathbf{Y})$ quantifies the amount of information needed to describe the outcome of a random variable $\mathbf{Y}$ given that the value of another random variable $\mathbf{X}$ is known. It is defined as

$$\mathrm{H}(\mathbf{X}|\mathbf{Y}) = -\sum_{x\in\boldsymbol{X},y\in\boldsymbol{Y}} p(x,y) \log \frac{p(x,y)}{p(y)}, \tag{1.4}$$

where $p(x, y)$ is the probability that $\mathrm{X} = x$, and $\mathrm{Y} = y$.

Cover and Thomas (1991) also introduced MI by defining

$$\mathrm{I}(\mathbf{X};\mathbf{Y}) = \sum_{y\in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = \mathrm{H}(\mathbf{X}) - \mathrm{H}(\mathbf{X}|\mathbf{Y}), \tag{1.5}$$

where $p(x, y)$ is joint distribution function of $\mathbf{X}$ and $\mathbf{Y}$, and $p(x)$ and $p(y)$ are marginal distribution of $\mathbf{X}$ and $\mathbf{Y}$, respectively.

By Wyner (1978) and Dobrushin (1963), CMI can be defined with discrete random variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ as

$$\mathrm{I}(\mathbf{X};\mathbf{Y}|\mathbf{Z}) = \mathbb{E}_Z(\mathrm{I}(\mathbf{X};\mathbf{Y})|\mathbf{Z}) = \sum_{z\in Z} p_z(z) \sum_{y\in Y} \sum_{x\in X} p_{\mathbf{X},\mathbf{Y}|\mathbf{Z}}(x,y|z) \log \frac{p_{\mathbf{X},\mathbf{Y}|\mathbf{Z}}(x,y|z)}{p_{\mathbf{X}|\mathbf{Z}}(x|z)p_{\mathbf{Y}|\mathbf{Z}}(y|z)}$$

$$= \sum_{z\in Z} \sum_{y\in Y} \sum_{x\in X} p_{\mathbf{X},\mathbf{Y},\mathbf{Z}}(x,y,z) \log \frac{p_{\mathbf{Z}}(z)p_{\mathbf{X},\mathbf{Y},\mathbf{Z}}(x,y,z)}{p_{\mathbf{X},\mathbf{Z}}(x,z)p_{\mathbf{Y},\mathbf{Z}}(y,z)}, \tag{1.6}$$

where p with appropriate subscript are the marginal, joint, and conditional probability mass functions. It can also be written in terms of joint and conditional entropies as

$$I(\mathbf{X};\mathbf{Y}|\mathbf{Z}) = H(\mathbf{X},\mathbf{Z}) + H(\mathbf{Y},\mathbf{Z}) - H(\mathbf{X},\mathbf{Y},\mathbf{Z}) - H(\mathbf{Z}) = H(\mathbf{X}|\mathbf{Z}) - H(\mathbf{X}|\mathbf{Y},\mathbf{Z}), \quad (1.7)$$

or in terms of mutual information as

$$I(\mathbf{X};\mathbf{Y}|\mathbf{Z}) = I(\mathbf{X};\mathbf{Y},\mathbf{Z}) - I(\mathbf{X};\mathbf{Z}),$$

where $I(\mathbf{X};\mathbf{Y},\mathbf{Z}) = I(\mathbf{X};\mathbf{Z}) + I(\mathbf{X};\mathbf{Y}|\mathbf{Z})$

$$= H(\mathbf{Z}|\mathbf{X}) + H(\mathbf{X}) + H(\mathbf{Z}|\mathbf{Y}) + H(\mathbf{Y}) - H(\mathbf{Z}|\mathbf{X},\mathbf{Y}) - H(\mathbf{X},\mathbf{Y}) - H(\mathbf{Z})$$

$$= I(\mathbf{X};\mathbf{Y}) + H(\mathbf{Z}|\mathbf{X}) + H(\mathbf{Z}|\mathbf{Y}) - H(\mathbf{Z}|\mathbf{X},\mathbf{Y}) - H(\mathbf{Z}).$$

Note that conditional mutual information $I(\mathbf{X};\mathbf{Y}|\mathbf{Z})$ is always nonnegative.

A logistic regression model is used for illustration and the rest of paper is organized as follows: in Chapter 2, we first provide details of the DIS, which is interaction screening by correlation, and discuss its deficiencies, then we introduce two new interaction screening methods based on partial correlation and conditional mutual information. A logistic regression model is used for illustration. In Chapter 3, we compare the performance of these three interaction screening methods namely Correlation values, PC values, and CMI values under various simulation settings. We also apply the methods for real data and interpret the results and discussion in Chapter 4. Eventually, the three methods are compared. We discussed and concluded this thesis in Chapter 5. R codes are presented in a separate R file.

2. METHODOLOGY

2.1 Logistic Model

Given a data set $\{y_i, x_{i1}, \dots, x_{ip}\}_{i=1}^{n}$ of n independent and identically distributed (IID) samples, $\mathbf{X} = (X_1, \dots, X_p)^{\mathrm{T}}$ is a p-dimensional predictor vector and $\boldsymbol{y}$ is the response. We consider a logistic model with two-way interaction terms and quadratic terms by assuming

$$\mathrm{Log}\frac{\pi}{1-\pi} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \gamma_{11} X_1^2 + \gamma_{12} X_1 X_2 + \cdots + \gamma_{pp} X_p^2. \quad (2.1.1)$$

In model (2.1.1), $\beta_0, \boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^{\mathrm{T}}, \boldsymbol{\gamma} = (\gamma_{11}, \gamma_{12}, \dots, \gamma_{pp})^{\mathrm{T}}$ are unknown parameters. $\{X_m\}_{m=1}^{p}, \{X_m^2\}_{m=1}^{p}, and\ \{X_m X_n\}_{1 \le m < n \le p}$ are main effects, quadratic effects, and two-way interaction effects, respectively. We used a simplified model, which excludes the quadratic terms in this study,

$$\mathrm{Log}\frac{\pi}{1-\pi} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \gamma_{12} X_1 X_2 + \cdots + \gamma_{(p-1)p} X_{p-1} X_p. \quad (2.1.2)$$

2.2 A Direct Interaction Screening by Pearson's Correlation Coefficient

As we discussed in previous chapter, DIS has some drawbacks, and we used a very simple example to illustrate. Let us consider the covariance between Y and $X_m X_n$, denoted by $Cov(Y, X_m X_n)$ or $\sigma_{Y, X_m X_n}$, and the correlation between Y and $X_m X_n$, denoted by $corr(Y, X_m X_n)$ or $\rho_{Y, X_m X_n}$. Consider the model $\mathrm{Log}\frac{\pi}{1-\pi} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + a X_1 X_2, X_1 \perp X_2, X_1 \sim Ber(p_1), X_2 \sim Ber(p_2), Y \sim Ber(\pi)$.

$E(X_1) = p_1, E(X_2) = p_2, E(X_1^{\,2}) = p_1, E(X_2^{\,2}) = p_2, E(X_1 X_2) = p_1 p_2$

$Var(X_1) = E(X_1^{\,2}) - E(X_1)^2 = p_1(1 - p_1)$

$Var(X_2) = E(X_2^{\,2}) - E(X_2)^2 = p_2(1 - p_2)$

$Var(X_1 X_2) = E(X_1^{\,2} X_2^{\,2}) - E(X_1 X_2)^2 = p_1 p_2 (1 - p_1 p_2)$

$$\pi = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + a X_1 X_2)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + a X_1 X_2)}$$

$$E(Y) = E(Y|X_1 = X_2 = 0)P(X_1 = X_2 = 0) + E(Y|X_1 = X_2 = 1)P(X_1 = X_2 = 1) +$$

$$E(Y|X_1 = 1, X_2 = 0)P(X_1 = 1, X_2 = 0) + E(Y|X_1 = 0, X_2 = 1)P(X_1 = 0, X_2 = 1)$$

$$= \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} * (1 - p_1)(1 - p_2) + \frac{\exp(\beta_0 + \beta_1 + \beta_2 + a)}{1 + \exp(\beta_0 + \beta_1 + \beta_2 + a)} * p_1 p_2 +$$

$$\frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)} * p_1(1 - p_2) + \frac{\exp(\beta_0 + \beta_2)}{1 + \exp(\beta_0 + \beta_2)} * (1 - p_1)p_2$$

$$E(YX_1X_2) = \sum YX_1X_2 * \ P(YX_1X_2) = 1 * \ P(YX_1X_2)$$

$$= P(Y|X_1 = 1, X_2 = 1)P(X_1 = 1, X_2 = 1)$$

$$= \frac{\exp(\beta_0 + \beta_1 + \beta_2 + a)}{1 + \exp(\beta_0 + \beta_1 + \beta_2 + a)} * p_1 p_2$$

$$Cov(Y, X_1X_2) = \sigma_{Y,X_1X_2} = E(YX_1X_2) - E(Y)E(X_1X_2)$$

$$= \frac{\exp(\beta_0 + \beta_1 + \beta_2 + a)}{1 + \exp(\beta_0 + \beta_1 + \beta_2 + a)} * p_1 p_2 - p_1 p_2 \cdot E(Y)$$

$$= -p_1 p_2 [\frac{\exp(\beta_0)}{1 + \exp(\beta_0)} * (1 - p_1)(1 - p_2) + \frac{\exp(\beta_0 + \beta_1 + \beta_2 + a)}{1 + \exp(\beta_0 + \beta_1 + \beta_2 + a)} *$$

$$(1 - p_1 p_2) + \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)} * p_1(1 - p_2) + \frac{\exp(\beta_0 + \beta_2)}{1 + \exp(\beta_0 + \beta_2)} *$$

$$(1 - p_1)p_2]$$

$$corr(Y, X_1X_2) = \rho_{Y,X_1X_2} = \frac{\sigma_{Y,X_1X_2}}{\sigma_Y \sigma_{X_1X_2}}$$

There are two facts:

(i) when $a = 0, Cov(Y, X_1X_2) \neq 0$ and $corr(Y, X_1X_2) \neq 0$;

(ii) when $Cov(Y, X_1X_2) = 0$ $(corr(Y, X_1X_2) = 0)$, $a = \exp\left(\frac{A}{1-A}\right) - \beta_0 - \beta_1 - \beta_2$, where A = $\frac{\exp(\beta_0)}{1+\exp(\beta_0)} * \frac{(1-p_1)(1-p_2)}{1-p_1p_2} + \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)} * \frac{p_1(1-p_2)}{1-p_1p_2} + \frac{\exp(\beta_0+\beta_2)}{1+\exp(\beta_0+\beta_2)} * \frac{(1-p_1)p_2}{1-p_1p_2}$

Fact (i) suggests that X_1X_2 is essentially not predictive to the response, but could be falsely detected as important predictor in some circumstances. Fact (ii) suggests that X_1X_2 is important and predictive to the response, but may be taken for irrelevant to the response by mistake. In either case, the interaction screening by correlation does not work for this simple example.

In short, the naive screening procedure DIS fails to account for intrinsic correlations between interaction terms and their parents. Here, in particular, when we considered the interaction term X_1X_2, we forgot to contemplate the main effect terms X_1 and X_2. As a result, when $Cov(X_1, X_1X_2) \neq 0$, this DIS procedure may fail to find significant interaction effects. This motivated us to develop an alternative method which takes into account main effects when evaluating interaction effects and can improve accuracy for interaction screening.

2.3 Partial Correlation Interaction Screening

To improve the correlation method, we considered the partial correlation between Y and X_mX_n conditional on X_m and X_n, denoted by $\rho_{Y,X_mX_n|X_m,X_n}$. To see advantages of the partial correlation approach, let us revisit the example in Section 2.2.

Consider the model, $\text{Log}\frac{\pi}{1-\pi} = \beta_0 + \beta_1X_1 + \beta_2X_2 + aX_1X_2,\ X_1 \perp X_2,\ X_1 \sim Ber(p_1)$, $X_2 \sim Ber(p_2), Y \sim Ber(\pi)$, and assume $a = 0$.

$\rho_{Y,X_1X_2|X_1,X_2} = Constant \cdot (\rho_{Y,X_1X_2|X_2} - \rho_{Y,X_1|X_2} \cdot \rho_{X_1,X_1X_2|X_2})$ based on the definition

where $\text{Part}(1) = \rho_{Y,X_1X_2|X_2} - \rho_{Y,X_1|X_2} \cdot \rho_{X_1,X_1X_2|X_2}$

$$\rho_{Y,X_1X_2|X_2} = \frac{\rho_{Y,X_1X_2} - \rho_{Y,X_2}\rho_{X_1X_2,X_2}}{\sqrt{1-\rho_{Y,X_2}^2}\sqrt{1-\rho_{X_1X_2,X_2}^2}}$$

$$\rho_{Y,X_1|X_2} = \frac{\rho_{Y,X_1} - \rho_{Y,X_2}\rho_{X_1,X_2}}{\sqrt{1-\rho_{Y,X_2}^2}\sqrt{1-\rho_{X_1,X_2}^2}} = \frac{\rho_{Y,X_1}}{\sqrt{1-\rho_{Y,X_2}^2}} \; since \; \rho_{X_1,X_2} = 0$$

$$\rho_{X_1,X_1X_2|X_2} = \frac{\rho_{X_1,X_1X_2} - \rho_{X_1,X_2}\rho_{X_1X_2,X_2}}{\sqrt{1-\rho_{X_1,X_2}^2}\sqrt{1-\rho_{X_1X_2,X_2}^2}} = \frac{\rho_{X_1,X_1X_2}}{\sqrt{1-\rho_{X_1X_2,X_2}^2}} \quad since \; \rho_{X_1,X_2} = 0$$

$$\text{Part}(1) = \frac{\rho_{Y,X_1X_2} - \rho_{Y,X_2}\rho_{X_1X_2,X_2} - \rho_{Y,X_1}\rho_{X_1,X_1X_2}}{Constant}$$

where $\text{Part}(2) = \rho_{Y,X_1X_2} - \rho_{Y,X_2}\rho_{X_1X_2,X_2} - \rho_{Y,X_1}\rho_{X_1,X_1X_2}$

$$= \frac{\sigma_{Y,X_1X_2}}{\sigma_Y\sigma_{X_1X_2}} - \frac{\sigma_{Y,X_2}\sigma_{X_1X_2,X_2}}{\sigma_Y\sigma_{X_2}^2\sigma_{X_1X_2}} - \frac{\sigma_{Y,X_1}\sigma_{X_1,X_1X_2}}{\sigma_Y\sigma_{X_1}^2\sigma_{X_1X_2}}$$

$$= Constant \cdot (\sigma_{Y,X_1X_2}\sigma_{X_1}^2\sigma_{X_2}^2 - \sigma_{Y,X_2}\sigma_{X_1X_2,X_2}\sigma_{X_1}^2 - \sigma_{Y,X_1}\sigma_{X_1,X_1X_2}\sigma_{X_2}^2)$$

where $\text{Part}(3) = \sigma_{Y,X_1X_2}\sigma_{X_1}^2\sigma_{X_2}^2 - \sigma_{Y,X_2}\sigma_{X_1X_2,X_2}\sigma_{X_1}^2 - \sigma_{Y,X_1}\sigma_{X_1,X_1X_2}\sigma_{X_2}^2$

$$= \sigma_{Y,X_1X_2}p_1(1-p_1)p_2(1-p_2) - \sigma_{Y,X_2}p_1(1-p_1)p_1p_2(1-p_2) -$$

$$\sigma_{Y,X_1}p_1p_2(1-p_1)p_2(1-p_2)$$

$$= Constant \cdot (\sigma_{Y,X_1X_2} - \sigma_{Y,X_2}p_1 - \sigma_{Y,X_1}p_2)$$

where $\text{Part}(4) = \sigma_{Y,X_1X_2} - \sigma_{Y,X_2}p_1 - \sigma_{Y,X_1}p_2$

The simplified $\text{Part}(4) = Constant \cdot (\frac{\exp(\beta_0+\beta_1+\beta_2)}{1+\exp(\beta_0+\beta_1+\beta_2)} - \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)} - \frac{\exp(\beta_0+\beta_2)}{1+\exp(\beta_0+\beta_2)} +$

$\frac{\exp(\beta_0)}{1+\exp(\beta_0)}) \neq 0$ unless $\beta_1 \; or \; \beta_2 = 0$

$$\rho_{Y,X_1X_2|X_1,X_2} = Constant \cdot (\frac{\exp(\beta_0+\beta_1+\beta_2)}{1+\exp(\beta_0+\beta_1+\beta_2)} - \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)} - \frac{\exp(\beta_0+\beta_2)}{1+\exp(\beta_0+\beta_2)} + \frac{\exp(\beta_0)}{1+\exp(\beta_0)}) \neq 0$$

unless $\beta_1 \; or \; \beta_2 = 0$

In particular, $\rho_{Y,X_1X_2|X_1,X_2} = 0$ when $\beta_1 = 0 \; or \; \beta_2 = 0$. This, together with the simulation study in the next chapter suggests that by using PC, we can partially eliminate the influence of main effects when conducting interaction screening.

2.4 Conditional Mutual Information Interaction Screening

Conditional mutual information was also applied to screen the interaction terms, and the accuracy of interaction screening was compared with the other two approaches.

$$\mathrm{I}(\mathrm{Y};\mathrm{X}_1\mathrm{X}_2|\mathrm{X}_1,\mathrm{X}_2) = \mathrm{H}(\mathrm{Y}|\mathrm{X}_1,\mathrm{X}_2) - \mathrm{H}(\mathrm{Y}|\mathrm{X}_1\mathrm{X}_2,\mathrm{X}_1,\mathrm{X}_2)$$

$$= -\sum p(y,x_1,x_2)\log\frac{p(y,x_1,x_2)}{p(x_1,x_2)} + \sum p(y,x_1x_2,x_1,x_2)\log\frac{p(y,x_1x_2,x_1,x_2)}{p(x_1x_2,x_1,x_2)}$$ by definition.

As we discussed in the previous section, DIS has some drawbacks, and we want to see if the use of CMI can improve. We will use a very simple example to illustrate. Consider the model $\mathrm{Log}\frac{\pi}{1-\pi} = \beta_0 + \beta_1X_1 + \beta_2X_2 + aX_1X_2,\ X_1 \perp X_2, X_1 \sim Ber(p_1), X_2 \sim Ber(\,p_2), Y \sim Ber(\pi)$, and assume $a = 0$. By calculating $\mathrm{I}(y;\mathrm{X}_1\mathrm{X}_2) = \sum_{y\in Y} p(y,\mathrm{X}_1\mathrm{X}_2)\log\frac{p(y,\mathrm{X}_1\mathrm{X}_2)}{p(y)p(\mathrm{X}_1\mathrm{X}_2)}$, we want to see if there is any special case that $\mathrm{I}(y;\mathrm{X}_1\mathrm{X}_2)$ could equal 0.

There are four situations:

(i) $y = 0, \mathrm{X}_1\mathrm{X}_2 = 0$;

(ii) $y = 0, \mathrm{X}_1\mathrm{X}_2 = 1$;

(iii) $y = 1, \mathrm{X}_1\mathrm{X}_2 = 0$;

(iv) $y = 1, \mathrm{X}_1\mathrm{X}_2 = 1$.

In situation (i),

$$p(y = 0, \mathrm{X}_1\mathrm{X}_2 = 0) = p(y = 0, \mathrm{X}_1 = 0, \mathrm{X}_2 = 0) + p(y = 0, \mathrm{X}_1 = 0, \mathrm{X}_2 = 1) +$$

$$p(y=0, X_1=1, X_2=0)$$

$$= p(y=0|X_1=0, X_2=0) * p(X_1=0) * p(X_2=0) +$$

$$p(y=0|X_1=0, X_2=1) * p(X_1=0) * p(X_2=1) +$$

$$p(y=0|X_1=1, X_2=0) * p(X_1=1) * p(X_2=0)$$

$$= \left(1 - \frac{\exp(\beta_0)}{1+\exp(\beta_0)}\right)(1-p_1)(1-p_2) +$$

$$\left(1 - \frac{\exp(\beta_0+\beta_2)}{1+\exp(\beta_0+\beta_2)}\right)(1-p_1)p_2 +$$

$$\left(1 - \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}\right)p_1(1-p_2);$$

$$p(y=0) = p(y=0|X_1=0, X_2=0) * p(X_1=0, X_2=0) +$$

$$p(y=0|X_1=0, X_2=1) * p(X_1=0, X_2=1) +$$

$$p(y=0|X_1=1, X_2=0) * p(X_1=1, X_2=0) +$$

$$p(y=0|X_1=1, X_2=1) * p(X_1=1, X_2=1)$$

$$= \left(1 - \frac{\exp(\beta_0)}{1+\exp(\beta_0)}\right)(1-p_1)(1-p_2) + \left(1 - \frac{\exp(\beta_0+\beta_2)}{1+\exp(\beta_0+\beta_2)}\right)(1-p_1)p_2 +$$

$$\left(1 - \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}\right)p_1(1-p_2) + (1 - \frac{\exp(\beta_0+\beta_1+\beta_2)}{1+\exp(\beta_0+\beta_1+\beta_2)})p_1p_2;$$

$$p(X_1X_2=0) = p(X_1X_2=0|X_1=0, X_2=0) + p(X_1X_2=0|X_1=0, X_2=1) +$$

$$p(X_1X_2=0|X_1=1, X_2=0)$$

$$= (1-p_1)(1-p_2) + (1-p_1)p_2 + p_1(1-p_2);$$

similar calculation in situation (ii),

$$p(y=0, X_1X_2=1) = p(y=0, X_1=0, X_2=0) + p(y=0, X_1=0, X_2=1) +$$

$$p(y=0, X_1=1, X_2=0)$$

$$= (1 - \frac{\exp(\beta_0 + \beta_1 + \beta_2)}{1 + \exp(\beta_0 + \beta_1 + \beta_2)})p_1p_2;$$

$$p(y = 0) = p(y = 0|\mathrm{X}_1 = 0, \mathrm{X}_2 = 0) * p(\mathrm{X}_1 = 0, \mathrm{X}_2 = 0) +$$

$$p(y = 0|\mathrm{X}_1 = 0, \mathrm{X}_2 = 1) * p(\mathrm{X}_1 = 0, \mathrm{X}_2 = 1) +$$

$$p(y = 0|\mathrm{X}_1 = 1, \mathrm{X}_2 = 0) * p(\mathrm{X}_1 = 1, \mathrm{X}_2 = 0) +$$

$$p(y = 0|\mathrm{X}_1 = 1, \mathrm{X}_2 = 1) * p(\mathrm{X}_1 = 1, \mathrm{X}_2 = 1)$$

$$= \left(1 - \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}\right)(1 - p_1)(1 - p_2) + \left(1 - \frac{\exp(\beta_0 + \beta_2)}{1 + \exp(\beta_0 + \beta_2)}\right)(1 - p_1)p_2 +$$

$$\left(1 - \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}\right)p_1(1 - p_2) + (1 - \frac{\exp(\beta_0 + \beta_1 + \beta_2)}{1 + \exp(\beta_0 + \beta_1 + \beta_2)})p_1p_2;$$

$$p(\mathrm{X}_1\mathrm{X}_2 = 1) = p(\mathrm{X}_1\mathrm{X}_2 = 1|\mathrm{X}_1 = 1, \mathrm{X}_2 = 1) = p_1p_2;$$

similar calculation in situation (iii),

$$p(y = 1, \mathrm{X}_1\mathrm{X}_2 = 0) = p(y = 1, \mathrm{X}_1 = 0, \mathrm{X}_2 = 0) + p(y = 1, \mathrm{X}_1 = 0, \mathrm{X}_2 = 1) +$$

$$p(y = 1, \mathrm{X}_1 = 1, \mathrm{X}_2 = 0)$$

$$= \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}(1 - p_1)(1 - p_2) + \frac{\exp(\beta_0 + \beta_2)}{1 + \exp(\beta_0 + \beta_2)}(1 - p_1)p_2 +$$

$$\frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}p_1(1 - p_2);$$

$$p(y = 1) = p(y = 1|\mathrm{X}_1 = 0, \mathrm{X}_2 = 0) * p(\mathrm{X}_1 = 0, \mathrm{X}_2 = 0) +$$

$$p(y = 1|\mathrm{X}_1 = 0, \mathrm{X}_2 = 1) * p(\mathrm{X}_1 = 0, \mathrm{X}_2 = 1) +$$

$$p(y = 1|\mathrm{X}_1 = 1, \mathrm{X}_2 = 0) * p(\mathrm{X}_1 = 1, \mathrm{X}_2 = 0) +$$

$$p(y = 1|\mathrm{X}_1 = 1, \mathrm{X}_2 = 1) * p(\mathrm{X}_1 = 1, \mathrm{X}_2 = 1)$$

$$= \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}(1 - p_1)(1 - p_2) + \frac{\exp(\beta_0 + \beta_2)}{1 + \exp(\beta_0 + \beta_2)}(1 - p_1)p_2 +$$

$$\frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}p_1(1 - p_2) + \frac{\exp(\beta_0 + \beta_1 + \beta_2)}{1 + \exp(\beta_0 + \beta_1 + \beta_2)}p_1p_2;$$

$$p(\mathrm{X}_1\mathrm{X}_2 = 0) = p(\mathrm{X}_1\mathrm{X}_2 = 0|\mathrm{X}_1 = 0, \mathrm{X}_2 = 0) + p(\mathrm{X}_1\mathrm{X}_2 = 0|\mathrm{X}_1 = 0, \mathrm{X}_2 = 1) +$$

$$p(\mathrm{X}_1\mathrm{X}_2 = 0|\mathrm{X}_1 = 1, \mathrm{X}_2 = 0)$$

$$= (1 - p_1)(1 - p_2) + (1 - p_1)p_2 + p_1(1 - p_2);$$

similar calculation in situation (iv),

$$p(y = 1, \mathrm{X}_1\mathrm{X}_2 = 1) = p(y = 1, \mathrm{X}_1 = 1, \mathrm{X}_2 = 1) = \frac{\exp(\beta_0 + \beta_1 + \beta_2)}{1 + \exp(\beta_0 + \beta_1 + \beta_2)} p_1 p_2;$$

$$p(y = 1) = p(y = 1|\mathrm{X}_1 = 0, \mathrm{X}_2 = 0) * p(\mathrm{X}_1 = 0, \mathrm{X}_2 = 0) +$$

$$p(y = 1|\mathrm{X}_1 = 0, \mathrm{X}_2 = 1) * p(\mathrm{X}_1 = 0, \mathrm{X}_2 = 1) +$$

$$p(y = 1|\mathrm{X}_1 = 1, \mathrm{X}_2 = 0) * p(\mathrm{X}_1 = 1, \mathrm{X}_2 = 0) +$$

$$p(y = 1|\mathrm{X}_1 = 1, \mathrm{X}_2 = 1) * p(\mathrm{X}_1 = 1, \mathrm{X}_2 = 1)$$

$$= \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}(1 - p_1)(1 - p_2) + \frac{\exp(\beta_0 + \beta_2)}{1 + \exp(\beta_0 + \beta_2)}(1 - p_1)p_2 +$$

$$\frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)} p_1(1 - p_2) + \frac{\exp(\beta_0 + \beta_1 + \beta_2)}{1 + \exp(\beta_0 + \beta_1 + \beta_2)} p_1 p_2;$$

$$p(\mathrm{X}_1\mathrm{X}_2 = 1) = p(\mathrm{X}_1\mathrm{X}_2 = 1|\mathrm{X}_1 = 1, \mathrm{X}_2 = 1) = p_1 p_2;$$

as a result, $\mathrm{I}(y; \mathrm{X}_1\mathrm{X}_2) = \sum_{y \in Y} p(y, \mathrm{X}_1\mathrm{X}_2) \log \frac{p(y, \mathrm{X}_1\mathrm{X}_2)}{p(y)p(\mathrm{X}_1\mathrm{X}_2)}$ is a combination of these four situations.

There are two facts:

(i) when $a = 0, \mathrm{I}(y; \mathrm{X}_1\mathrm{X}_2)$ is generally nonzero;

(ii) unless $p_1 = p_2 = 0.5$ and $\beta_1 = \beta_2 = 0$ simultaneously.

Fact (i) suggests that X_1X_2 is essentially not predictive to the response, but could be falsely considered as existent and important to the prediction. Fact (ii) suggests that X_1X_2 is important and predictive to the response, but may be taken for irrelevant to the response by mistake. In either case, the interaction screening by mutual information does not work for this

simple example. To this end, we considered a conditional mutual information measure to partially adjust for the main effects. This is same as the idea of partial correlation coefficient in spirit.

3. SIMULATION

To study the performance of the interaction screening methods that were proposed above, we present two simulations and one real data example. R packages: "ppcor" and "infotheo" are used in both simulation and real data analysis parts. The "ppcor" package calculates partial and semi-partial (part) correlations along with p-value. The "infotheo" package implements various measures of information theory based on several entropy estimators.

The key idea of the interaction screen is to sift out the important interaction terms based on different statistical features. We have already known the limitation of DIS, but for the other two approaches, we are not familiar with their performances. In this case, we decide to evaluate their accuracy rates by simulated data.

3.1 Simulated Setting I: Discrete Case

For the first simulation, we used logistic model (2.1.2) IID predictor variables generated from a discrete distribution: Bernoulli distribution, which is equivalent to binomial distribution of size 1. We considered two such models with (n, p) = (1000, 5), and (n, p) = (1000, 10). We set the number of main terms as p, and p-dimensional vectors β were randomly chosen, i.e. $\beta_1 = \beta_2 = \cdots = \beta_p = 0.25$. Let c denote the number of non-zero coefficients of interaction terms, and here, we set c = 3. $\beta_{12}, \beta_{25}, \beta_{45}$ are true interaction term's coefficients for model with 5 predictors, and $\beta_{12}, \beta_{45}, \beta_{69}$ are true interaction term's coefficients for model with 10 predictors. We set Y ~ Bin (1, prob), where $\text{prob} = \frac{\exp(\beta_0+\beta_1+\cdots+\beta_p+\beta_{ef}+\beta_{gh}+\beta_{ij})}{1+\exp(\beta_0+\beta_1+\cdots+\beta_p+\beta_{ef}+\beta_{gh}+\beta_{ij})}$ and $\beta_{ef}, \beta_{gh}, \beta_{ij}$ are true interaction term's coefficients. We also set three different numbers of simulations s (number of replications), which are 100, 500, and 1000. Finally, we summarize the number of times that

the three different methods correctly capture the true interactions for every simulation, which is noted by *t*. Accuracy rate is calculated by $\frac{\sum_{i=1}^{s} t_i}{3s}$. Seed was set as 12345. The results are shown in the table 3.1.1. In the table, Correlation, PC, and CMI are the evaluation terms for DIS, PCIS, and CMIIS, respectively.

Table 3.1.1

Simulation Results for Discrete Data

Number of simulations	Accuracy Rate (%)					
	Correlation		Partial Correlation		Conditional Mutual Information	
	5	10	5	10	5	10
100	46.00	13.33	31.00	8.00	62.33	19.33
500	50.47	14.67	31.2	7.4	66.73	19.93
1000	50.67	14.86	30.73	6.63	66.57	20.13

From the results, we can tell in discrete case, accuracy rate of CMI is superior to PC or Correlation methods for smaller number of predictors. The accuracy rates significantly drop as we change the number of predictors from 5 to 10 in all three number of simulation situations, although the size of true model does not change. Among the three methods, the performance of CMIIS is superior to DIS and PCIS. There is very little difference in accuracy rates between different simulation times. In two different settings of the model (p=5 or 10), the CMIIS is favorable and preferred.

3.2 Simulated Setting II: Continuous Case

For the second simulation, we used logistic model (2.1.2) IID predictor generated from a continuous distribution: N(0,1). We still considered two such models with (n, p) = (1000, 5), and (n, p) = (1000, 10). We set the number of main terms as p, and p-dimensional vectors β were randomly chosen, i.e. $\beta_1 = \beta_2 = \cdots = \beta_p = 0.25$. Let c denote the number of non-zero

coefficients of interaction terms, and here, we set $c = 3$. $\beta_{12}, \beta_{25}, \beta_{45}$ are true interaction term's coefficients for model with 5 predictors, and $\beta_{12}, \beta_{45}, \beta_{69}$ are true interaction term's coefficients for model with 10 predictors. We set Y ~ Bin (1, prob), where $\text{prob} = \frac{\exp(\beta_0+\beta_1+\cdots+\beta_p+\beta_{ef}+\beta_{gh}+\beta_{ij})}{1+\exp(\beta_0+\beta_1+\cdots+\beta_p+\beta_{ef}+\beta_{gh}+\beta_{ij})}$ and $\beta_{ef}, \beta_{gh}, \beta_{ij}$ are true interaction term's coefficients. We also set three different number of simulations (number of replications), which are 100, 500, and 1000. Finally, we summarize the number of times that the three different methods correctly capture the true interactions for every simulation, which is noted by *t*. Accuracy rate is calculated by $\frac{\sum_{i=1}^{s} t_i}{3s}$. Seed was set as 123456. The results are show in the table 3.2.1. In the table, Correlation, PC, and CMI are the evaluation terms for DIS, PCIS, and CMIIS, respectively.

Table 3.2.1

Simulation Results for Continuous Data

Number of simulations	Accuracy Rate					
	Correlation		Partial Correlation		Conditional Mutual Information	
	5	10	5	10	5	10
100	97.66	84.67	98.00	85.67	34.67	17.33
500	97.67	79.87	98.07	80.73	34.93	20.13
1000	97.57	79.67	98.20	80.77	34.87	19.97

Note: Different number of bins (5, 10, 20) were compared during calculating CMI, and there was no significant difference between them.

From the results, we can tell in continuous case, accuracy rate of Correlation, and PC are significantly larger than CMI. DIS and PCIS are comparable in two different settings of the model (*p*=5 or 10), and PCIS performs slightly better than DIS in all six conditions. Among the three methods, the performances of DIS and PCIS are superior to CMIIS. We cannot tell much difference in accuracy rates between 500 and 1000 simulations; however, when simulation number increases from 100 to 500, PC and Correlation values drop while CMI improving. We

suggest to use PCIS and DIS in continuous case, and PCIS is preferred. Furthermore, considering the time consumed, we suggest to use smaller number of simulations.

4. RESULTS AND DISCUSSION

RNA sequencing (RNA-Seq) is a formidable new technology in characterization and quantification for transcriptomes. Using sequencing technology, gene expression levels of all transcripts can be quantified digitally. However, the substantial biases in the data generated by RNA-Seq introduced great challenges to data analysis. Since RNA-Seq is more accurate the microarray and hold great promise to elucidate important information about the transcriptomes, in our analysis, we use RNA-seq data for cervical cancer to do interaction screening.

4.1 Dataset Description

This cervical cancer data set contains the expression level of 714 genes measured in normal group and cancer group. By data cleaning such as deleting poorly sequenced samples and genes with extremely low read count, a subset of 528 genes was selected from a total of 714 genes. In the data set, N stands for normal (healthy) subject, and T represents tumor (cancer) subject. For convenience, X was transferred to log (X^T+1) (X^T is the transpose of X) or sqrt (X^T+1), and then added a column of dependent variable y, which are set as N=0 and T=1 in our logistic model. For comparison, gene selection was also applied in log transformation case, and top 50 genes were selected based on the correlation between main effects and dependent variable y.

4.2 Result

After interaction effects screening by three different methods, three matrices were obtained by Correlation, PC and CMI of every interaction terms from the 528 genes for log transformation case and square root (sqrt) transformation case.

We ranked the result according to the magnitudes of Correlation, PC, CMI, and absolute value of difference of Correlation and PC in descending order. Only the top 10 gene interactions are shown in the tables as below.

4.2.1 Result with Log Transformation

Table 4.2.1.1

Interaction Effects Ranked According to Correlation in Descending Order with Log Transformation

	Xm	Xn	Corr	PC	CMI	|Corr-PC|
1	39	56	-0.69	-0.13	0.59	0.56
2	55	184	-0.69	-0.33	0.37	0.36
3	40	56	-0.69	-0.06	0.60	0.63
4	56	184	-0.69	-0.30	0.45	0.39
5	56	67	-0.69	-0.12	0.43	0.56
6	55	72	-0.69	-0.22	0.47	0.46
7	40	55	-0.69	-0.09	0.44	0.60
8	56	72	-0.68	-0.15	0.52	0.53
9	32	55	-0.68	0.11	0.42	0.79
10	56	464	-0.68	-0.11	0.44	0.56

From the Table 4.2.1.1, some genes show high frequencies in the first 10 interaction terms with top Correlation magnitudes, and these genes might be implicative to cervical cancer and deserve future research. Gene 56 appears 6 times, gene 55 appears 4 times, and gene 40, 72, and 184 appear twice. The first 10 interaction terms show a substantial discrepancy between Correlation and PC in magnitude, which are caused by ignoring the main effects in DIS.

Table 4.2.1.2

Interaction Effects Ranked According to PC in Descending Order with Log Transformation

	X_m	X_n	Corr	PC	CMI	|Corr-PC|
1	190	240	0.24	-0.56	0.43	0.80
2	34	432	0.06	0.53	0.45	0.47
3	146	240	0.37	-0.53	0.44	0.90
4	84	292	-0.07	0.52	0.48	0.60
5	70	292	-0.12	0.52	0.49	0.64
6	47	292	-0.10	0.52	0.46	0.63
7	432	438	-0.08	0.51	0.77	0.59
8	70	83	-0.05	0.50	0.45	0.56
9	110	430	0.16	0.50	0.40	0.34
10	94	292	-0.18	0.50	0.50	0.68

From the Table 4.2.1.2, some genes appear frequently in the first 10 interaction terms with top PC magnitudes. Gene 292 appears 4 times, and gene 240 and 432 appear 2 times. These 10 interaction terms have very high PC, however, their Correlations are much lower.

Table 4.2.1.3

Interaction Effects Ranked According to CMI in Descending Order with Log Transformation

	X_m	X_n	Corr	PC	CMI	|Corr-PC|
1	263	496	0.32	-0.10	0.90	0.42
2	221	496	0.34	-0.21	0.90	0.55
3	237	352	0.25	0.05	0.90	0.21
4	474	496	0.32	-0.09	0.89	0.41
5	351	352	0.20	-0.13	0.89	0.33
6	311	496	0.37	0.20	0.89	0.17
7	69	438	-0.23	-0.07	0.89	0.17
8	352	354	0.16	-0.09	0.89	0.25
9	84	496	0.35	0.08	0.89	0.27
10	121	352	0.28	0.07	0.88	0.21

From the Table 4.2.1.3, some genes also show high frequencies in the first 10 interaction terms with top CMI magnitudes. Gene 496 appears 5 times, and gene 352 appears 4 times. When CMI of these interaction terms are high, their PC are relatively low.

Table 4.2.1.4

Interaction Effects Ranked According to |Corr-PC| in Descending Order with Log Transformation

	X_m	X_n	Corr	PC	CMI	|Corr-PC|
1	32	282	-0.61	0.32	0.37	0.93
2	32	332	-0.57	0.34	0.41	0.91
3	146	240	0.37	-0.53	0.44	0.90
4	138	332	-0.53	0.34	0.54	0.87
5	4	32	-0.60	0.27	0.48	0.87
6	32	462	-0.63	0.24	0.33	0.87
7	60	332	-0.43	0.43	0.46	0.87
8	32	331	-0.62	0.25	0.37	0.87
9	154	205	-0.49	0.37	0.46	0.86
10	32	115	-0.60	0.25	0.40	0.85

From the Table 4.2.1.4, some genes appear frequently in the first 10 interaction terms with top absolute values of difference of correlation and partial correlation. Gene 32 appears 5 times, and gene 332 appears 3 times. The result shows that the difference between Correlation and PC could be as large as 0.93.

In statistics, hypotheses about the value of the population correlation coefficient ρ between variables X and Y can be tested using the Fisher z-transformation applied to the sample correlation coefficient. If (X, Y) has a bivariate normal distribution with correlation ρ and the pairs (X_i, Y_i) are independent and identically distributed, then z is approximately normally distributed with mean $0.5 * \ln \frac{1+\rho}{1-\rho}$, and standard error $\frac{1}{\sqrt{N-3}}$ where N is the sample size, and ρ is the true correlation coefficient (Wikipedia contributors, 2018). We applied Fisher z-transformation ($z = 0.5 * \ln \frac{1+r}{1-r}$) to Correlation, PC, CMI, and *p*-values for the three methods were also provided in the tables. The results are showed in figures and tables as below.

Table 4.2.1.5

Top 10 Interactions with Smallest p-values & Fisher z-transformation of Correlation with Log Transformation

	X_m	X_n	Corr	p-value
1	39	56	-0.85	< 0.001
2	55	184	-0.85	< 0.001
3	40	56	-0.84	< 0.001
4	56	184	-0.84	< 0.001
5	56	67	-0.84	< 0.001
6	55	72	-0.84	< 0.001
7	40	55	-0.84	< 0.001
8	56	72	-0.83	< 0.001
9	32	55	-0.83	< 0.001
10	56	464	-0.82	< 0.001

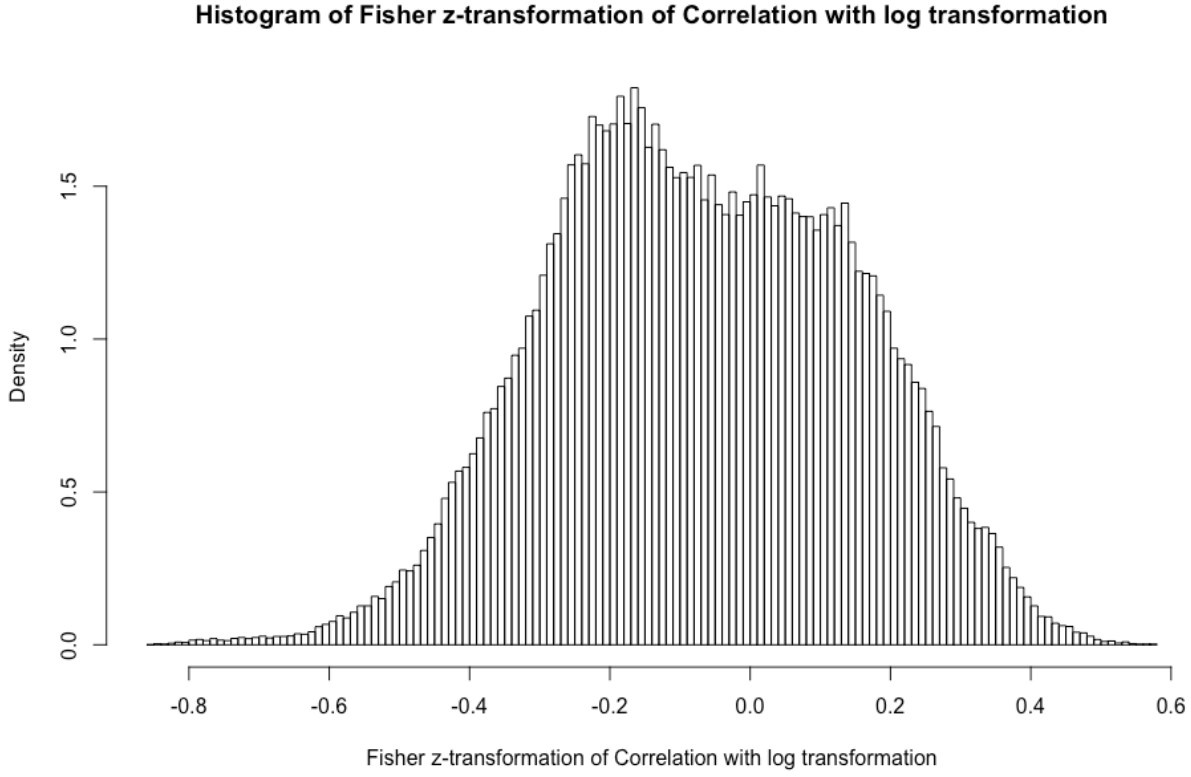

Figure 4.2.1.1 The figure shows the histogram of Fisher z-transformation of Correlation with log transformation, and it provides a distribution from at least two combining normal distributions.

From *Figure 4.2.1.1*, we can tell most information are from a normal distribution whose mean is around -0.2, and the rest information might from a normal distribution whose mean is around 0.1. However, the distributions are not separated very well. It can also prove that, during the screening, we lost a lot of information.

Table 4.2.1.6

Top 10 Interactions with Smallest p-values & Fisher z-transformation of Partial Correlation with Log Transformation

	X_m	X_n	PC	p-value
1	190	240	-0.63	< 0.001
2	34	432	0.59	< 0.001
3	146	240	-0.59	< 0.001
4	84	292	0.58	< 0.001
5	70	292	0.58	< 0.001
6	47	292	0.58	< 0.001
7	432	438	0.56	< 0.001
8	70	83	0.55	< 0.001
9	110	430	0.55	< 0.001
10	94	292	0.55	< 0.001

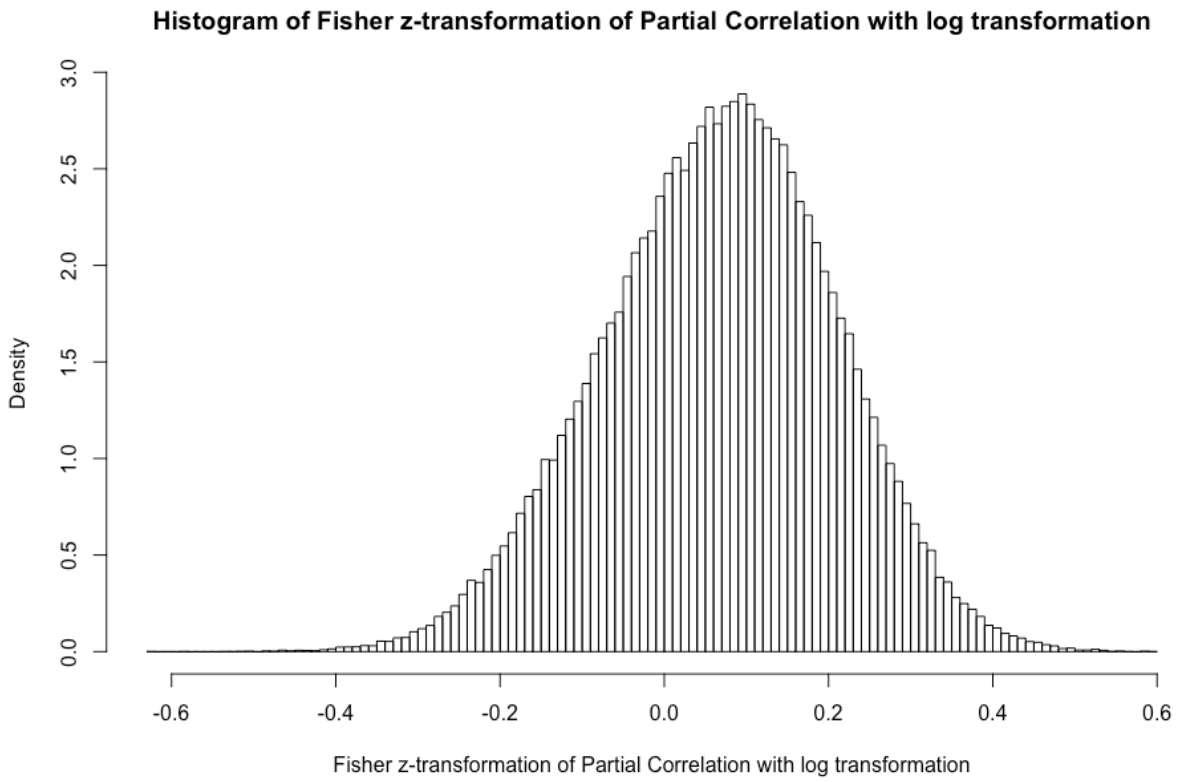

Figure 4.2.1.2 The figure shows the histogram of Fisher z-transformation of Partial Correlation with log transformation, and it provides a normal distribution whose mean is shifting from 0 to the right.

From *Figure 4.2.1.2*, we can tell Fisher z-transformations of PC are generally from a normal distribution whose mean is around 0.1, even though it is not a perfect normal distribution, since we might lose some information by screening.

Table 4.2.1.7

Top 10 Interactions with Smallest p-values & Fisher z-transformation of Conditional Mutual Information with Log Transformation

	Xm	Xn	CMI	p-value
1	69	438	0.15	0.26
2	226	438	0.15	0.27
3	32	168	0.15	0.27
4	237	352	0.15	0.28
5	351	352	0.15	0.28
6	352	354	0.15	0.28
7	56	168	0.15	0.28
8	121	352	0.15	0.28
9	55	168	0.14	0.28
10	263	496	0.14	0.28

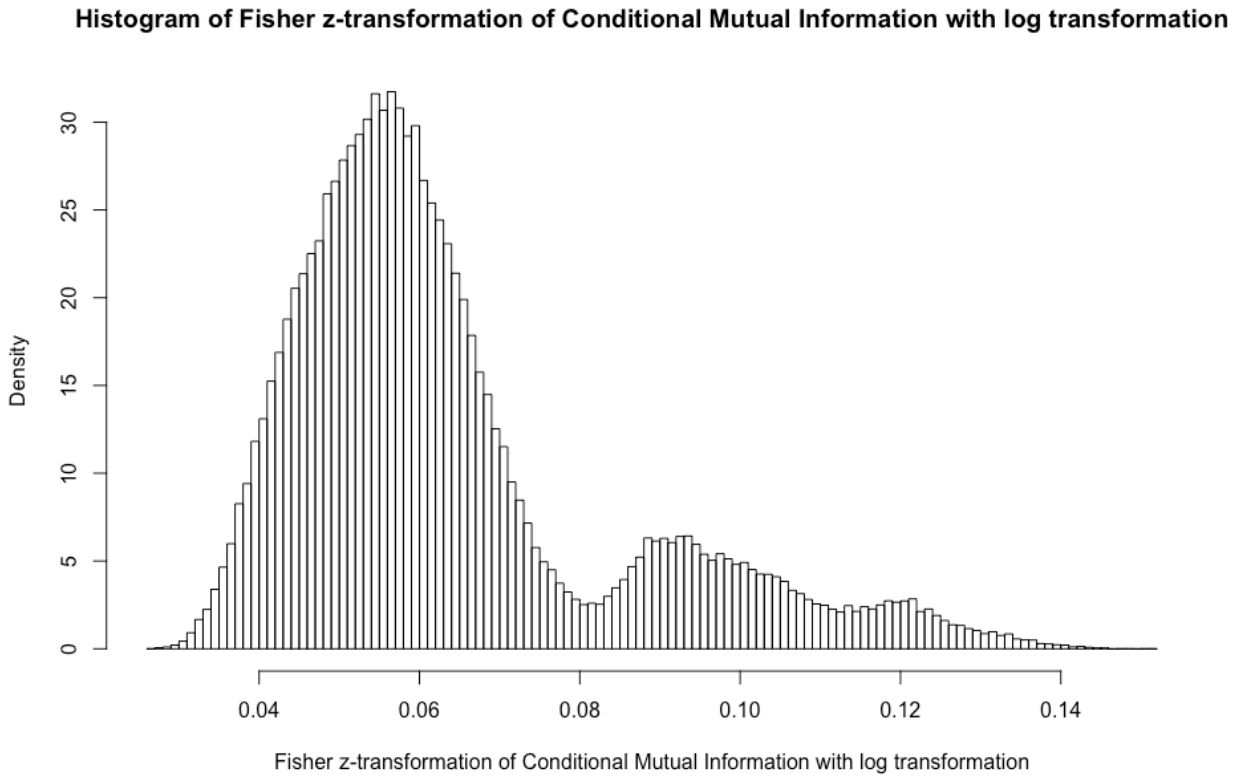

Figure 4.2.1.3 This figure shows the histogram of Fisher z-transformation of Conditional Mutual Information with log transformation, and it provides that this distribution comes from at least three normal distributions combined.

The histogram of Fisher z-transformation of CMI (*Figure 4.2.1.3*) shows us a distribution contained at least three peaks with three means that were far enough apart, and I believed it was a mixture distribution which means it was a mixture of at least three normal distributions with different means. The first pick means there were a lot interaction terms that contain similar amount information, and those interaction terms formed a bigger normal distribution compared with the second pick. It can also prove that, during the screening, we lost a lot of information.

4.2.2 Result with Gene Selection

Top 50 genes were selected based on the correlation between main effects and dependent variable y, and they are gene "34", "72", "73", "53", "52", "399", "18", "473", "188", "92", "87", "165", "279", "97", "618", "189", "54", "625", "641", "102", "138", "398", "22", "85", "562", "8", "19", "32", "207", "180", "585", "582", "186", "56", "33", "426", "86", "581", "271", "338", "353", "427", "690", "224", "84", "632", "57", "197", "171", "270".

Table 4.2.2.1

Interaction Effects Ranked According to Correlation in Descending Order with Gene Selection

	Xm	Xn	Corr	PC	CMI	|Corr-PC|
1	73	52	-0.69	-0.13	0.52	0.56
2	72	224	-0.69	-0.33	0.37	0.36
3	73	53	-0.69	-0.06	0.50	0.63
4	73	224	-0.69	-0.30	0.45	0.39
5	73	87	-0.69	-0.12	0.43	0.56
6	72	92	-0.69	-0.22	0.47	0.46
7	72	53	-0.69	-0.09	0.41	0.60
8	73	92	-0.68	-0.15	0.52	0.53
9	34	72	-0.68	0.11	0.42	0.79
10	73	625	-0.68	-0.11	0.44	0.56

From the Table 4.2.2.1, some genes show high frequencies in the first 10 interaction terms with top Correlation magnitudes, and these genes might be implicative to cervical cancer and deserve future research. Gene 73 appears 6 times, gene 72 appears 4 times, and gene 53 appear twice. The first 10 interaction terms show a substantial discrepancy between Correlation and PC in magnitude, which are caused by ignoring the main effects in DIS. Considering the log transformation case, gene 72 would be the one appears in both case.

Table 4.2.2.2

Interaction Effects Ranked According to PC in Descending Order with Gene Selection

	X_m	X_n	Corr	PC	CMI	|Corr-PC|
1	32	690	-0.35	0.46	0.69	0.81
2	33	690	-0.38	0.43	0.68	0.81
3	19	690	-0.38	0.42	0.66	0.81
4	56	690	-0.41	0.42	0.68	0.83
5	690	57	-0.40	0.42	0.42	0.82
6	427	690	-0.28	0.42	0.67	0.70
7	399	426	-0.40	0.42	0.61	0.82
8	207	180	0.18	-0.42	0.37	0.60
9	85	690	-0.36	0.41	0.64	0.77
10	399	427	-0.34	0.40	0.69	0.74

From the Table 4.2.2.2, some genes appear frequently in the first 10 interaction terms with top PC magnitudes. Gene 690 appears 7 times, and gene 399 and 427 appear 2 times. These 10 interaction terms have very high PC, however, different from log transformation case, their Correlations are quite close to PC. Comparing with log transformation case, there are no genes appear in top 10 interaction terms in two cases at same time.

Table 4.2.2.3

Interaction Effects Ranked According to CMI in Descending Order with Gene Selection

	X_m	X_n	Corr	PC	CMI	|Corr-PC|
1	52	165	-0.55	0.13	0.87	0.68
2	53	165	-0.55	0.10	0.81	0.65
3	18	165	-0.54	0.14	0.78	0.68
4	92	165	-0.55	0.11	0.78	0.66
5	87	585	-0.50	-0.04	0.77	0.45
6	87	165	-0.54	0.17	0.76	0.71
7	188	165	-0.52	0.18	0.76	0.70
8	72	165	-0.56	0.04	0.74	0.60
9	399	165	-0.49	0.29	0.73	0.78
10	426	427	-0.34	0.27	0.73	0.61

From the Table 4.2.2.3, some genes also show high frequencies in the first 10 interaction terms with top CMI magnitudes. Gene 165 appears 8 times, and gene 87 appears twice. Same in log transformation case, when CMI of these interaction terms are high, their PC are relatively low.

Table 4.2.2.4

Interaction Effects Ranked According to |Corr-PC| in Descending Order with Gene Selection

	X_m	X_n	Corr	PC	CMI	|Corr-PC|
1	34	399	-0.57	0.34	0.41	0.91
2	399	171	-0.53	0.34	0.47	0.87
3	34	398	-0.62	0.25	0.37	0.87
4	34	165	-0.52	0.33	0.69	0.85
5	399	279	-0.50	0.34	0.42	0.84
6	399	86	-0.52	0.32	0.47	0.84
7	34	86	-0.61	0.22	0.36	0.83
8	34	171	-0.61	0.22	0.31	0.83
9	34	625	-0.62	0.21	0.41	0.83
10	34	19	-0.64	0.20	0.35	0.83

From the Table 4.2.2.4, some genes appear frequently in the first 10 interaction terms with top absolute values of difference of correlation and partial correlation. Gene 34 appears 7

times, gene 399 appears 4 times, and gene 86 appears 2 times. Similar as log transformation case, the result shows that the difference between Correlation and PC could be as large as 0.91.

We applied Fisher z-transformation ($z = 0.5 * \ln \frac{1+r}{1-r}$) to Correlation, PC, CMI, and *p*-values for the three methods were also provided in the tables. The results are showed in figures and tables as below.

Table 4.2.2.5

Top 10 Interactions with Smallest p-values & Fisher z-transformation of Correlation with Gene Selection

	X_m	X_n	Corr	p-value
1	73	52	-0.85	< 0.001
2	72	224	-0.85	< 0.001
3	73	53	-0.84	< 0.001
4	73	224	-0.84	< 0.001
5	73	87	-0.84	< 0.001
6	72	92	-0.84	< 0.001
7	72	53	-0.84	< 0.001
8	73	92	-0.83	< 0.001
9	34	72	-0.83	< 0.001
10	73	625	-0.82	< 0.001

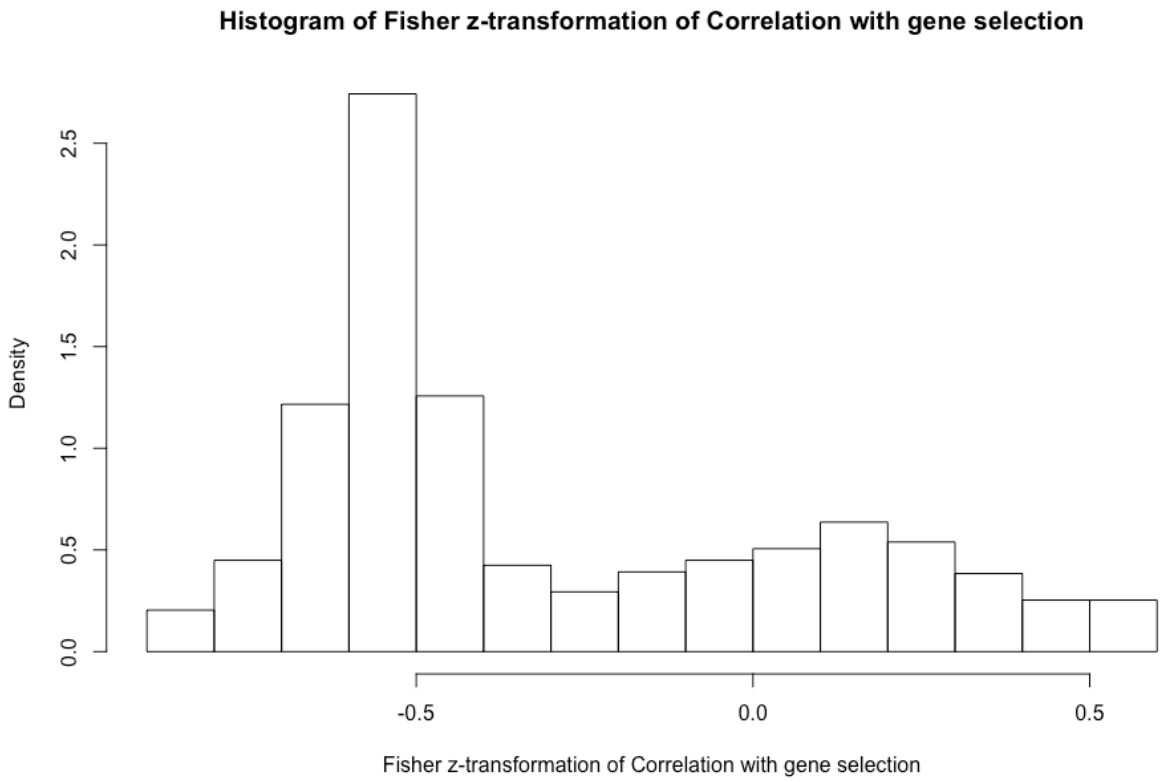

Figure 4.2.2.1 The figure shows the histogram of Fisher z-transformation of Correlation with gene selection, and it provides a distribution from at least two combining normal distributions.

From *Figure 4.2.2.1*, we can tell most information are from a normal distribution whose mean is around -0.55, and the rest information might from a normal distribution whose mean is around 0.15. It can also prove that, during the screening, we lost a lot of information.

Table 4.2.2.6

Top 10 Interactions with Smallest p-values & Fisher z-transformation of Partial Correlation with Gene Selection

	X_m	X_n	PC	p-value
1	32	690	0.49	< 0.001
2	33	690	0.46	< 0.001
3	19	690	0.45	< 0.001
4	56	690	0.45	< 0.001
5	690	57	0.45	< 0.001
6	427	690	0.45	< 0.001
7	399	426	0.45	< 0.001
8	207	180	-0.44	< 0.01
9	85	690	0.43	< 0.01
10	399	427	0.43	< 0.01

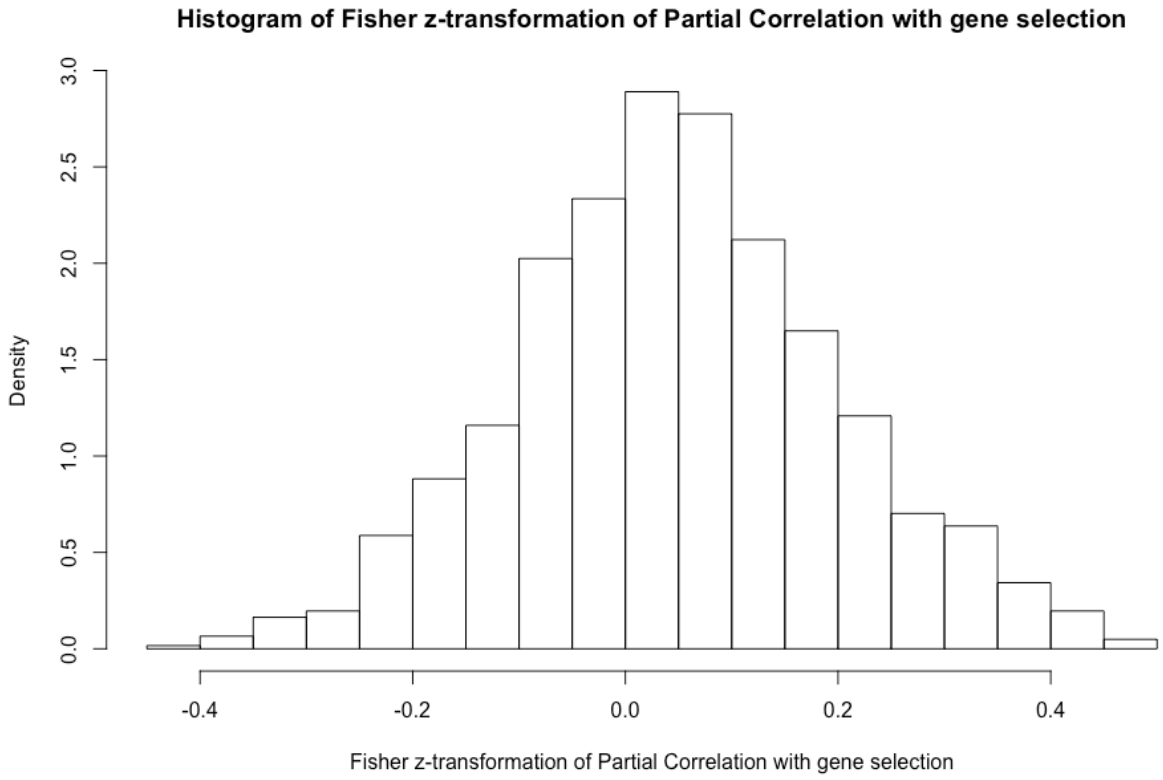

Figure 4.2.2.2 The figure shows the histogram of Fisher z-transformation of Partial Correlation with gene selection, and it provides a normal distribution whose mean is slightly shifting from 0 to the right.

From *Figure 4.2.2.2*, we can tell Fisher z-transformations of PC are generally from a normal distribution whose mean is around 0.05, even though it is not a perfect normal distribution, since we might lose some information by screening.

Table 4.2.2.7

Top 10 Interactions with Smallest p-values & Fisher z-transformation of Conditional Mutual Information with Gene Selection

	Xm	Xn	CMI	p-value
1	52	165	0.14	0.31
2	53	165	0.13	0.35
3	18	165	0.12	0.37
4	92	165	0.12	0.37
5	87	585	0.12	0.37
6	87	165	0.12	0.38
7	188	165	0.12	0.38
8	72	165	0.12	0.38
9	399	165	0.12	0.38
10	426	427	0.12	0.38

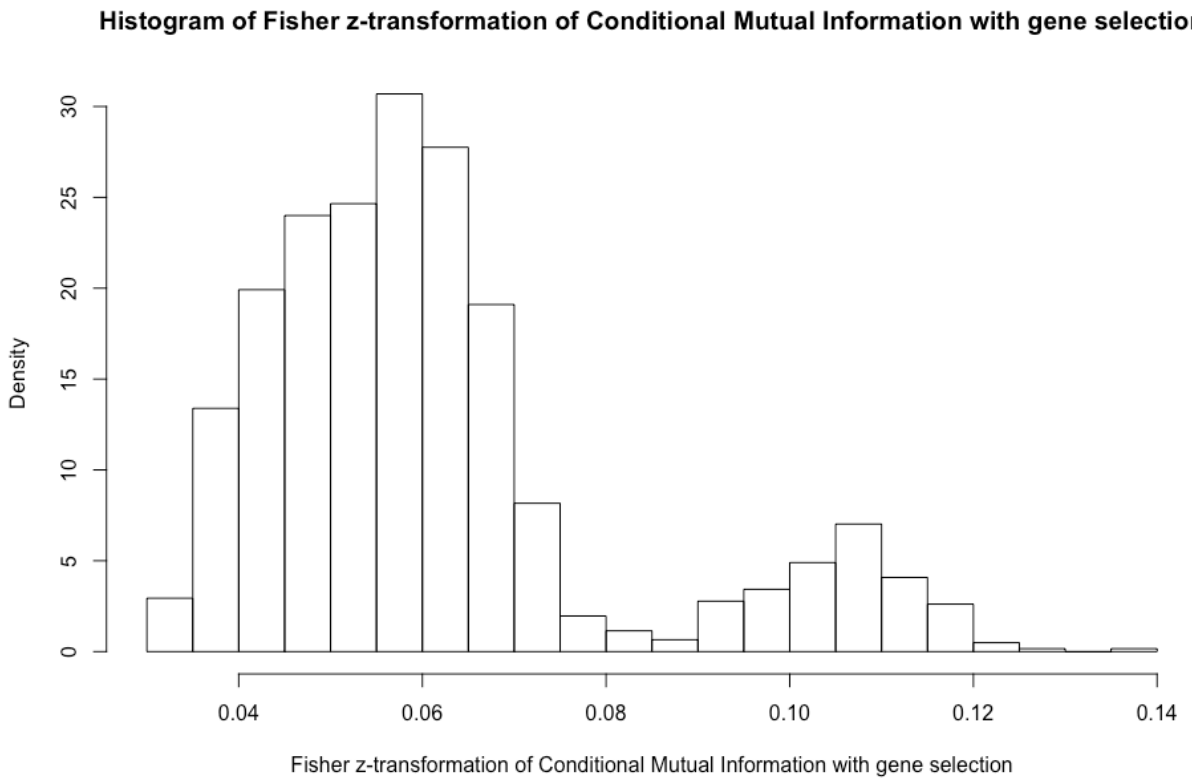

Figure 4.2.2.3 This figure shows the histogram of Fisher z-transformation of Conditional Mutual Information with gene selection, and it provides that this distribution comes from at least two

The histogram of Fisher z-transformation of CMI (*Figure 4.2.2.3*) shows us a distribution contained at least two peaks with three means that were far enough apart, and I believed it was a mixture distribution which means it was a mixture of at least two normal distributions with different means. The first pick means there were a lot interaction terms that contain similar amount information, and those interaction terms formed a bigger normal distribution compared with the second pick. It can also prove that, during the screening, we lost a lot of information.

4.2.3 Result with Square Root Transformation

Table 4.2.3.1

Interaction Effects Ranked According to Correlation in Descending Order with sqrt Transformation

	Xm	Xn	Corr	PC	CMI	|Corr-PC|
1	55	139	-0.65	-0.30	0.36	0.34
2	56	184	-0.64	-0.22	0.42	0.42
3	56	139	-0.63	-0.30	0.43	0.33
4	56	338	-0.63	-0.30	0.34	0.33
5	56	238	-0.63	-0.36	0.41	0.27
6	56	67	-0.62	-0.14	0.40	0.48
7	39	56	-0.61	-0.11	0.57	0.51
8	56	464	-0.61	-0.10	0.41	0.51
9	56	273	-0.61	-0.17	0.37	0.44
10	55	273	-0.61	-0.15	0.35	0.46

From the Table 4.2.3.1, some genes show high frequencies in the first 10 interaction terms with top Correlation magnitudes, and these genes might be implicative to cervical cancer and deserve future research. Gene 56 appears 8 times, and gene 55, 139, and 273 appears 2 times. The first 10 interaction terms show a substantial discrepancy between Correlation and PC in magnitude, which are caused by ignoring the main effects in DIS. Considering the log transformation case, gene 56 appears most time in both case in top 10 interaction terms.

Table 4.2.3.2

Interaction Effects Ranked According to PC in Descending Order with sqrt Transformation

	X_m	X_n	Corr	PC	CMI	|Corr-PC|
1	332	477	-0.29	0.52	0.42	0.81
2	34	432	0.09	0.52	0.45	0.42
3	101	108	0.01	0.51	0.38	0.51
4	66	337	-0.05	0.50	0.42	0.56
5	124	304	-0.01	0.50	0.45	0.51
6	101	119	0.08	0.50	0.32	0.42
7	179	227	0.17	-0.50	0.44	0.67
8	137	153	0.22	-0.50	0.42	0.71
9	138	477	-0.17	0.49	0.38	0.66
10	138	517	-0.04	0.49	0.39	0.53

From the Table 4.2.3.2, some genes appear frequently in the first 10 interaction terms with top PC magnitudes. Gene 101 and 38 appear twice. These 10 interaction terms have very

high PC, however, their Correlations are much lower. Comparing with log transformation case, there are no genes appear in top 10 interaction terms in two cases at same time.

Table 4.2.3.3

Interaction Effects Ranked According to CMI in Descending Order with sqrt Transformation

	X_m	X_n	Corr	PC	CMI	|Corr-PC|
1	263	496	0.23	0.04	0.90	0.19
2	221	496	0.24	-0.10	0.90	0.34
3	237	352	0.23	0.09	0.90	0.14
4	474	496	0.31	-0.06	0.89	0.37
5	351	352	0.17	-0.17	0.89	0.34
6	311	496	0.35	0.27	0.89	0.09
7	69	438	-0.19	0.02	0.89	0.21
8	352	354	0.15	-0.16	0.89	0.30
9	84	496	0.35	0.19	0.89	0.16
10	121	352	0.26	0.13	0.88	0.13

From the Table 4.2.3.3, some genes also show high frequencies in the first 10 interaction terms with top CMI magnitudes. Gene 496 appears 5 times, and gene 352 appears 4 times. When CMI of these interaction terms are high, their PC are relatively low. Comparing with log transformation case, genes appear in top 10 interaction terms are exactly same.

Table 4.2.3.4

Interaction Effects Ranked According to |Corr-PC| in Descending Order with sqrt Transformation

	X_m	X_n	Corr	PC	CMI	|Corr-PC|
1	32	332	-0.45	0.45	0.41	0.90
2	32	134	-0.41	0.45	0.69	0.86
3	32	72	-0.47	0.39	0.43	0.86
4	18	32	-0.43	0.43	0.45	0.86
5	152	332	-0.40	0.44	0.45	0.85
6	72	348	-0.39	0.46	0.64	0.85
7	18	331	-0.42	0.42	0.46	0.84
8	72	331	-0.44	0.40	0.48	0.84
9	72	194	-0.40	0.44	0.53	0.83
10	72	133	-0.41	0.42	0.45	0.83

From the Table 4.2.3.4, some genes appear frequently in the first 10 interaction terms with top absolute values of difference of correlation and partial correlation. Gene 72 appears 5 times, gene 32 appears 4 times, and gene 331 and 332 appears 2 times. The result shows that the difference between Correlation and PC could be as large as 0.90. Considering the log transformation case, gene 32 would be the one appears in both case.

We applied Fisher z-transformation ($z = 0.5 * \ln \frac{1+r}{1-r}$) to Correlation, PC, CMI, and *p*-values for the three methods were also provided in the tables. The results are showed in figures and tables as below.

Table 4.2.3.5

Top 10 Interactions with Smallest p-values & Fisher z-transformation of Correlation with sqrt Transformation

	X_m	X_n	Corr	p-value
1	55	139	-0.77	< 0.001
2	56	184	-0.75	< 0.001
3	56	139	-0.75	< 0.001
4	56	338	-0.74	< 0.001
5	56	238	-0.74	< 0.001
6	56	67	-0.72	< 0.001
7	39	56	-0.72	< 0.001
8	56	464	-0.71	< 0.001
9	56	273	-0.71	< 0.001
10	55	273	-0.71	< 0.001

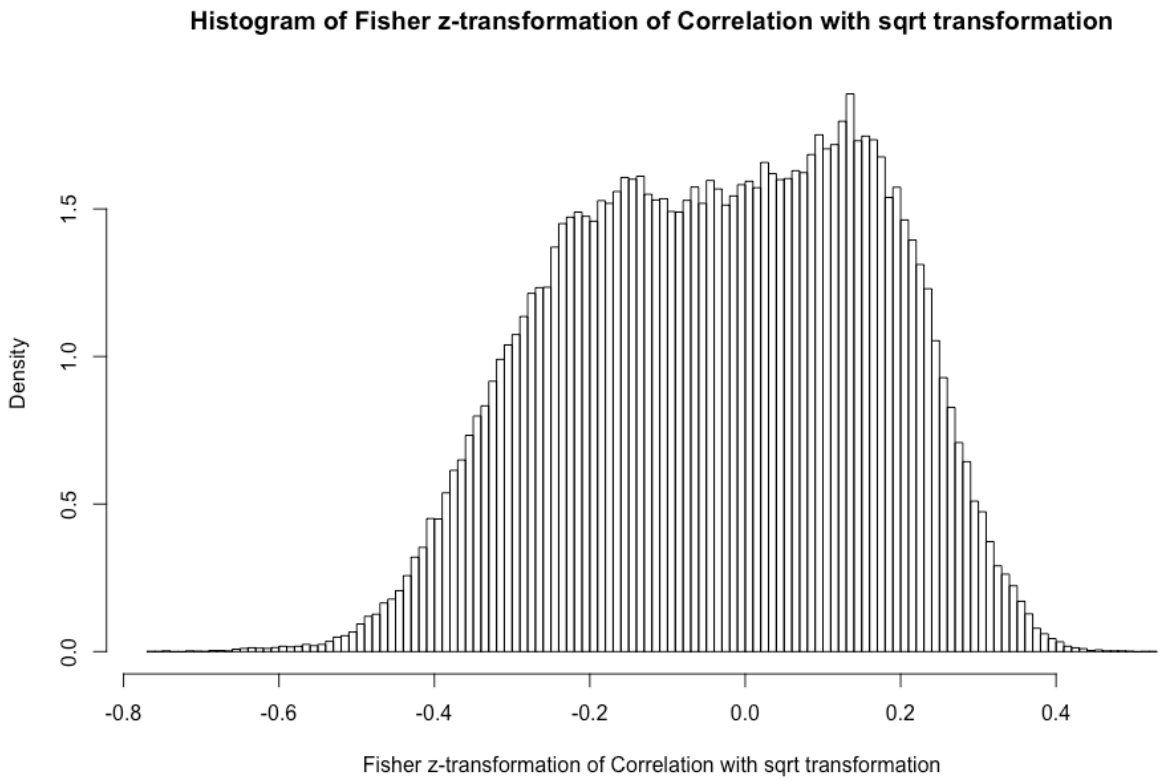

Figure 4.2.3.1 The figure shows the histogram of Fisher z-transformation of Correlation with sqrt transformation, and it provides a distribution from at least two combining normal distributions.

From *Figure 4.2.3.1*, we can tell most information are from a normal distribution whose mean is around -0.15, and the rest information might from a normal distribution whose mean is around 0.15. However, the distributions are not separated very well. It can also prove that, during the screening, we lost a lot of information.

Table 4.2.3.6

Top 10 Interactions with Smallest p-values & Fisher z-transformation of Partial Correlation with sqrt Transformation

	X_m	X_n	PC	p-value
1	332	477	0.58	< 0.001
2	34	432	0.57	< 0.001
3	101	108	0.57	< 0.001
4	66	337	0.56	< 0.001
5	124	304	0.55	< 0.001
6	101	119	0.55	< 0.001
7	179	227	0.54	< 0.001
8	137	153	0.54	< 0.001
9	138	477	0.54	< 0.001
10	138	517	0.54	< 0.001

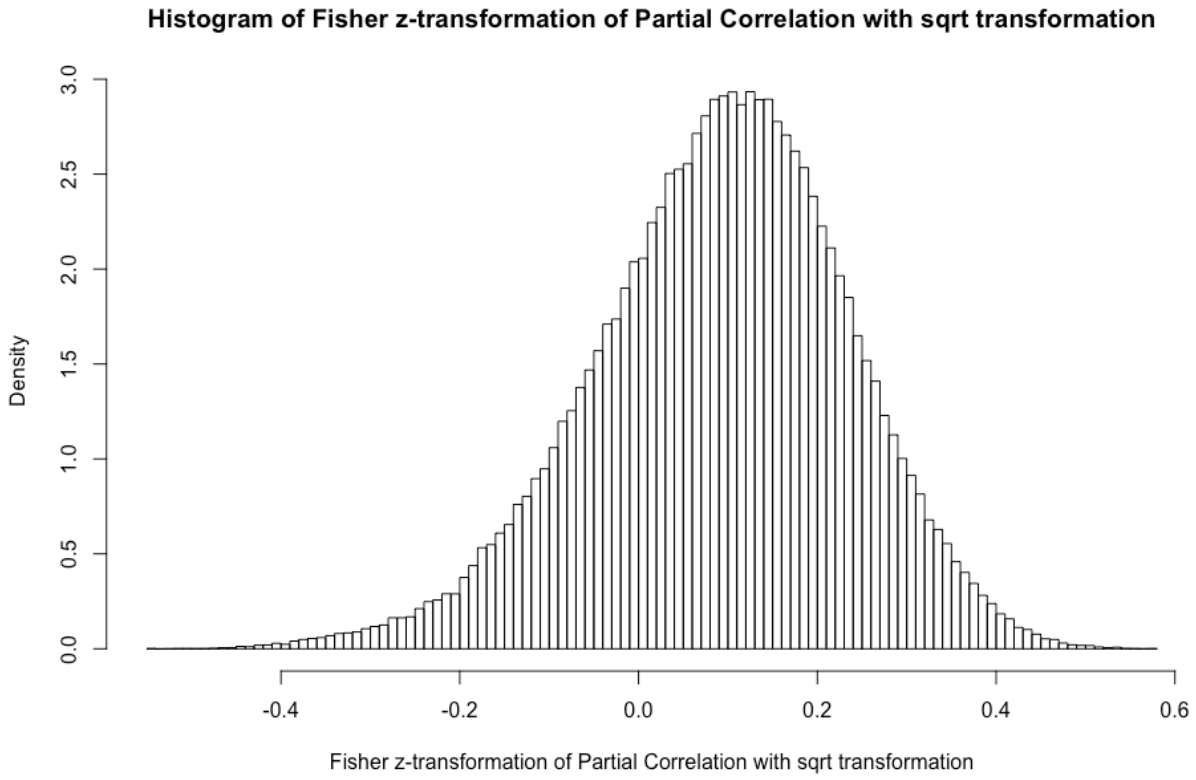

Figure 4.2.3.2 The figure shows the histogram of Fisher z-transformation of Partial Correlation with sqrt transformation, and it provides a normal distribution whose mean is shifting from 0 to the right.

From *Figure 4.2.3.2*, we can tell Fisher z-transformations of PC are generally from a normal distribution whose mean is around 0.1, even though it is not a perfect normal distribution, since we might lose some information by screening.

Table 4.2.3.7

Top 10 Interactions with Smallest p-values & Fisher z-transformation of Conditional Mutual Information with sqrt Transformation

	Xm	Xn	CMI	p-value
1	69	438	0.15	0.26
2	226	438	0.15	0.27
3	32	168	0.15	0.27
4	237	352	0.15	0.28
5	351	352	0.15	0.28
6	352	354	0.15	0.28
7	56	168	0.15	0.28
8	121	352	0.15	0.28
9	55	168	0.14	0.28
10	263	496	0.14	0.28

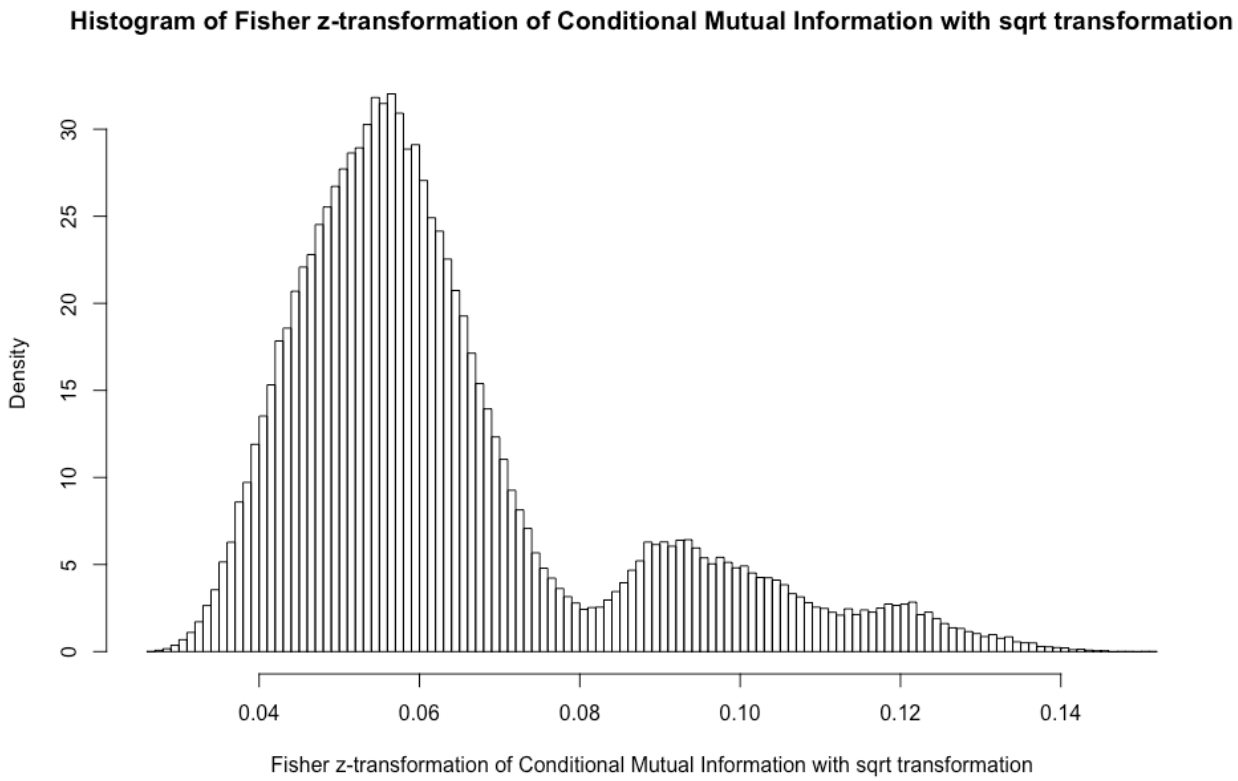

Figure 4.2.3.3 This figure shows the histogram of Fisher z-transformation of Conditional Mutual Information with sqrt transformation, and it provides that this distribution comes from at least three normal distributions combined.

The histogram of Fisher z-transformation of CMI (*Figure 4.2.3.3*) shows us a distribution contained at least three peaks with three means that were far enough apart, and I believed it was a mixture distribution which means it was a mixture of at least three normal distributions with different means. The first pick means there were a lot interaction terms that contain similar amount information, and those interaction terms formed a bigger normal distribution compared with the second pick. It can also prove that, during the screening, we lost a lot of information.

5. CONCLUSIONS

In this study, we compared 3 interaction screening methods: DIS, PCIS, and CMIIS. From the simulation examples and real data analysis, we found although DIS has certain drawbacks discussed in Chapter 2, it is still applicable in those models that are not very complex and especially with discrete independent variables. PCIS is comparable in complex models with continuous independent variables, even though it is not very stable in models with discrete independent variables. CMIIS performs quite competitive in both simple and complex models with discrete independent variables, however, its accuracy rate in both models with continuous variables is quite low. Compared with discrete cases, we found the 3 methods works much better in continuous cases. For real data, we found screening through different methods may get different results.

REFERENCES

Bien, J., Simon, N., & Tibshirani, R. (2015). Convex hierarchical testing of interactions. *The Annals of Applied Statistics*, *9*(1), 27-42.

Bien, J., Taylor, J., & Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of statistics*, *41*(3), 1111.

Bühlmann, P., & Kalisch, M. (2007). *Variable selection for high-dimensional models: partial faithful distributions, strong associations and the PC-algorithm*. Technical report, ETH Zürich.

Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, *24*(1), 17-36.

Choi, N. H., Li, W., & Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, *105*(489), 354-364.

Cordell, H. J. (2009). Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, *10*(6), 392.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. NewYork, NY, USA: Wiley.

Dobrushin, R. L. (1963). General formulation of Shannon's main theorem in information theory. *Amer. Math. Soc. Trans*, *33*(2), 323-438.

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(5), 849-911.

Hao, N., & Zhang, H. H. (2014). A Note on High Dimensional Linear Regression with Interactions. *arXiv preprint arXiv:1412.7138*.

Li, G., Peng, H., Zhang, J., & Zhu, L. (2012). Robust rank correlation based screening. *The Annals of Statistics*, *40*(3), 1846-1877.

Li, R., Zhong, W., & Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, *107*(499), 1129-1139.

McCullagh, P. (2002). What is a statistical model?. *Annals of statistics*, 1225-1267.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (Vol. 37). CRC Press.

Moore, J. H., Asselbergs, F. W., & Williams, S. M. (2010). Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, *26*(4), 445-455.

Nelder, J. A. (1977). A reformulation of linear models. *Journal of the Royal Statistical Society. Series A (General)*, 48-77.

Nelder, J. A. (1994). The statistics of linear models: back to basics. *Statistics and Computing*, *4*(4), 221-234.

Niu, Y. S., Hao, N., & Zhang, H. H. (2018). Interaction screening by partial correlation. *Statistics and its Interface, 11*(2), 317-325.

Park, M. Y., & Hastie, T. (2007). Penalized logistic regression for detecting gene interactions. *Biostatistics*, *9*(1), 30-50.

Schneider, T. D. (2007). Information theory primer with an appendix on logarithms. In *National Cancer Institute*.

Sotoca, J. M., & Pla, F. (2010). Supervised feature selection by clustering using conditional mutual information-based distances. *Pattern Recognition*, *43*(6), 2068-2081.

Wikipedia contributors. (2018, March 13). Partial correlation. In *Wikipedia, The Free Encyclopedia*. Retrieved 04:35, March 16, 2018, from https://en.wikipedia.org/w/index.php?title=Partial_correlation&oldid=830195526

Wikipedia contributors. (2018, March 21). Fisher transformation. In *Wikipedia, The Free Encyclopedia*. Retrieved 19:46, March 28, 2018, from https://en.wikipedia.org/w/index.php?title=Fisher_transformation&oldid=831684282

Wu, J., Devlin, B., Ringquist, S., Trucco, M., & Roeder, K. (2010). Screen and clean: a tool for identifying interactions in genome-wide association studies. *Genetic epidemiology*, *34*(3), 275-285.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., & Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, *25*(6), 714-721.

Wyner, A. D. (1978). A definition of conditional mutual information for arbitrary ensembles. *Information and Control*, *38*(1), 51-59.

Zhao, P., Rocha, G., & Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 3468-3497.

Zhu, L. P., Li, L., Li, R., & Zhu, L. X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, *106*(496), 1464-1475.