Statistical analysis for genomic studies involving measurement error, multiple populations, and limited sample size

by

Juan Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2013

Doctoral Committee:

Professor Kerby Shedden
Professor Matthias Kretzler
Associate Professor Ben Hansen
Assistant Professor Hui Jiang

For all the people

TABLE OF CONTENTS

LIST OF FIGURES

Figure

ABSTRACT

Statistical analysis for genomic studies involving measurement error, multiple populations, and limited sample size

by

Juan Zhang

Advisor: Kerby Shedden

Genomic studies involve various types of high-dimensional data. Study designs are often complex, and data are difficult to collect. For example, the subjects may belong to distinct populations, the number of subjects is often small, and substantial measurement error is usually present. In this thesis, we consider three important issues that arise in this research setting. The impact of measurement error on parameter estimation has been extensively studied, but its effects on predictive performance have not been. In part 1 of the thesis, we partially characterize the data generating models that are most adversely impacted by measurement error. These results may help researchers judge whether improving data collection procedures, or identifying more informative markers would have a greater impact on predictive performance. In part 2 of the thesis, we present a new approach for identifying the common and unique marker/outcome associations that are present in a genomic dataset consisting of several subpopulations. We show that the natural plug-in style estimates of overlap are biased, and we demonstrate a copula-based approach to reducing the bias. Part 3 of the thesis considers situations in which power for attributing effects to specific

markers is low, but meaningful relationships between marker/outcome associations and other statistical properties of the markers can be identified.

CHAPTER I

Introduction

1.1 Overview

This thesis considers several challenging issues that arise when analyzing genomic data. Difficulties that arise in this area commonly result from the effects of covariate measurement error, complicated dependence structure, and data sets with high dimension and small sample size. In this thesis, we are interested in how covariate measurement error affects predictive accuracy for outcome prediction under different data-generating models (chapter 2), the identification of common and unique effects in multiple subpopulations (chapter 3) and the relationship between effect sizes and properties of the marginal distributions of the markers (chapter 4).

1.1.1 Impact of covariate measurement error on prediction

Many genomic quantities are not measured with high accuracy. There exist many sources of measurement errors. In terms of the laboratory measurements, genomic assays such as microarrays attempt to quantify the abundances of many molecular types that are present in small amounts in a complex mixture. Such assays are known to exhibit only partial concordance, even between technical replicates. Moreover, in research involving human subjects, there may be transient variation within individuals that is irrelevant for many research goals. Transient qualities of the individuals include

mood, motivation, degree of alertness, boredom, and fatigue, and situational factors involving the physical setting such as noise level, lighting and time.

There are many researchers focusing on estimating the reproducibility of microarray data (*Larkin et al.* (2005), *Draghici et al.* (2005)) and the signal to noise ratio for microarray analysis (*He and Zhou* (2008)). In chapter 2 we use a triplicated expression array experiment on a panel of 59 cell lines to estimate the signal-to-noise ratio (SNR) ranges from 3:1 to 8:1. The main statistical focus in this area has been the impact of measurement error on estimation and inference for unknown parameters such as means and regression coefficients (*Fuller* (1987), *Carroll et al.* (2006)). However, issues resulting from measurement error also arise in predictive analysis. The effect of measurement error on predictive accuracy has received much less attention.

Chapter 2 considers the effects of covariate measurement error on predictive accuracy. Predictivity declines with increasing measurement error magnitude. But at a more detailed level, it is unclear whether the absolute or relative amount of decline in predictivity will differ according to the structure of the outcome generating distribution $P(Y|X)$. Our main focus in this chapter is to consider what attributes of the distribution P(Y,X) affect the degree to which covariate measurement error adversely impacts predictive accuracy for binary outcomes.

As an application, we will focus on gene expression used as a quantitative predictor of disease outcomes. Gene expression measurements are made with substantial measurement error, so it is important to know how this measurement error affects predictive performance, and whether or not measurement error plays a major role in limiting prediction accuracy. Doing this would allow researchers to focus on either improving measurement technology, or alternatively, on discovering new types of markers, based on whichever of these two strategies is likely to give the greatest improvement in predictive performance.

1.1.2 Common and unique associations in multiple subpopulations

Graphical displays of effect sizes across many tested variables are often included in scientific reports, for example, in genetic association studies (*Ioannidis et al.* (2005), *So and Sham* (2010)). But methods for formal analysis of effect size distributions have only recently been considered. For example, *Efron* (2007) used the empirical distribution of effect sizes to calculate false discovery rates. There are many opportunities to more deeply explore effect sizes in large, complex data sets. For example, clinical genomic studies often involve populations that can be subdivided into several distinct subpopulations. Associations between gene expression markers and patient outcome can be common or unique across such subpopulations.

In chapter 3, we consider the proportion of markers having large marker/outcome associations in two subpopulations as a measure of the overlap of effect sizes. However, the simple empirical measure of this overlap can be quite biased. We propose a new copula-based method to estimate this quantity, and show that it substantially reduces the bias.

1.1.3 Relationships between marginal properties of variables and their external correlations

In chapter 4, we consider another aspect of effect size distributions in complex data sets. Our goal is to consider whether markers that are correlated with an outcome have different marginal statistical properties than those that are not correlated with the outcome. We call this a property/marker/outcome association. We present a method for identifying distributions of genomic markers that are statistically related to the strengths of the marker/outcome associations.

This leads to a type of integrated correlation measure, for which it is difficult to assess the statistical properties. Therefore we propose a simulation-based approach to assess the bias and variability of the estimated property/marker/outcome association.

Another issue is that there may be not only monotone associations between genomic markers and the outcome. The non-monotone associations like u-shape association are not detected by the Pearson correlation coefficient. We develop a way that decomposes an association into a monotone component and a symmetric concave/convex component (plus a residual function) to see which association is dominant.

1.2 The Chronic Kidney Disease dataset

Much of the work described in this thesis was motivated by a genomic dataset that we call the "Chronic Kidney Disease" (CKD) dataset. Here we give a brief overview of this dataset. Chronic kidney disease, also known as chronic renal disease, is a progressive loss of renal function that takes place over a period of months or years. Chronic kidney disease is identified by a blood test for serum creatinine, with higher levels of creatinine indicating a falling glomerular filtration rate (GFR) and as a result a decreased capability of the kidneys to excrete waste products. The CKD dataset is a collection of clinical and genomic data for subjects with one of several diseases that give rise to CKD. The diseases in the CKD dataset include Focal segmental glomerulosclerosis (FSGS), Systemic lupus erythematosus (SLE), and Minimal Change Disease (MCD).

The genomic data in the CKD dataset consist of microarray measurements of gene expression on specific cell types obtained from kidney tissue biopsy specimens taken early in the disease course. The main clinical parameter of interest is the GFR taken at the biopsy time. GFR is a widely used overall index of kidney function. Specifically, it estimates how much blood passes through the tiny filters in the kidneys, called glomeruli, each minute. Normal GFR results range from 90-120 mL/min, GFR below 60 mL/min implies moderate loss of renal function, and GFR below 30 mL/min is considered to be severe. The dataset includes genomic and clinical data for 195 subjects, and the gene expression data quantify gene expression for 12,023 distinct

genes or transcripts. While the relatively small number of genes whose function is specific to the kidneys are of particular interest, CKD is associated with many physiological processes such as inflammation. Therefore exploratory analyses continue to play a major role in this area.

A major long-term goal for research in this area is to identify genomic markers that predict a rapidly declining GFR trend, and to clarify the molecular processes involved in CKD progression. Much of this work involves correlative analyses in data sets such as the CKD dataset. The issues discussed in this thesis all address significant challenges to progress in this field.

CHAPTER II

Differential effects of covariate measurement error in outcome prediction

2.1 Introduction

Regression models are often defined in terms of independent variables (covariates) that are not measured perfectly, or for some reason are not directly observable. In such situations, error-prone measurements or surrogate covariates, X_{obs}, are used instead of the true covariates X. The substitution of X_{obs} for X usually biases the coefficient estimates, and much research has been done on methods to correct and adjust for this bias (*Fuller* (1987), *Carroll et al.* (2006)). A related but distinct question is to consider how prediction methods are impacted by the presence of covariates that are measured with error. Predictivity must decline with increasing measurement error magnitude. But at a more detailed level, is the amount of decline strongly dependent on the structure of the outcome generating distribution $P(Y|X)$? Our main focus in this chapter is to consider what attributes of the distribution P(Y,X) might affect the rate at which covariate measurement error adversely impacts predictive accuracy for binary outcomes.

As an application, we will focus on gene expression used as a quantitative predictor of disease outcomes. Gene expression measurements are made with substantial

measurement error, so it is important to know how this measurement error affects predictive performance, and whether or not measurement error plays a major role in limiting predictive accuracy. Doing this would allow researchers to focus on either improving measurement technology, or alternatively, on discovering new types of markers, based on whichever of these two strategies is likely to give the greatest improvement in predictive performance.

We next consider examples of potentially important factors of data generating models and how they affect the amount of decline in predictive accuracy due to measurement error. The potential factors are the structure of the covariance matrix of measurement error Σ_η, the structure of the covariance matrix of covariates Σ_X and the true regression coefficients β.

First, we generate $(X_1, X_2) \in \mathbb{R}^2$ following a centered bivariate normal distribution with standard deviations 0.5 and correlation r. The binary outcome is generated by $P(Y = 1) = 1/(1+\exp(-c(X_1 + X_2)))$, where c is a constant which is chosen to make the Bayes' rule predictive accuracy with no measurement error equal to 0.9. Then measurement errors $(\eta_1, \eta_2) \in \mathbb{R}^2$ which follow independent centered bivariate normal distribution with standard deviations s are added to the covariates. The Bayes' rule predictive accuracy with measurement error magnitude s is calculated and plotted on figure 2.1 with different values of correlation r (r=0.8, -0.8, and 0). Figure 2.1 focuses on the impact of Σ_X on the predictive accuracy while holding other factors fixed. We see that highly positively correlated covariates exhibit a greater decline of predictive accuracy than negatively correlated covariates.

Second, we repeat the example above using $r = 0$ and the binary outcome is generated by $P(Y = 1) = 1/(1 + \exp(-c(\beta_1 X_1 + \beta_2 X_2)))$. Define $\beta = (\beta_1, \beta_2)$. With the same procedure, the Bayes' rule predictive accuracy with measurement error magnitude s is calculated and plotted on figure 2.2 with different structure of β ($\beta = (1, 1), (1, -1)$, and $(0, 1)$). Figure 2.2 focuses on the impact of regression

coefficients β on the predictive accuracy while holding other factors fixed. We see that different structure of β will not change the amount of decline of predictive accuracy very much.

Third, we also repeat the example above using $r = 0.8$ and the binary outcome is generated by $P(Y = 1) = 1/(1 + \exp(-c(X_1 + X_2)))$. Here the measurement errors (η_1, η_2) are not always independent, they have correlation $\tilde{r}$. With the same procedure, the Bayes' rule predictive accuracy with measurement error magnitude s is calculated and plotted on figure 2.3 with different value of $\tilde{r}$ ($\tilde{r} = 0.8, -0.8$, and 0). Figure 2.3 focuses on the impact of Σ_η on the predictive accuracy while holding other factors fixed. We see that highly negatively correlated measurement error exhibits a greater decline of the predictive accuracy than positively correlated measurement error.

Our overall strategy is to first identify factors that may impact the drop of performance in predictive modeling due to measurement error. These factors are identified from the theoretical derivation of predictive accuracy and from analogies to the linear case where the issues are much more straightforward. Following our analytic studies, we then use simulation to assess how the factors we have identified impact the drop of performance in predictive modeling for binary outcomes, and to assess the relationships between these factors. Results are given for both linear and binary cases to see if there is any interpretable difference. To put this to practical use, we will consider how to estimate the key attributes from the data and to estimate the amount of decline of predictivity in a real data analysis. Another important goal is to estimate the predictive accuracy that would be obtained if no measurement error were present from data observed with measurement error.

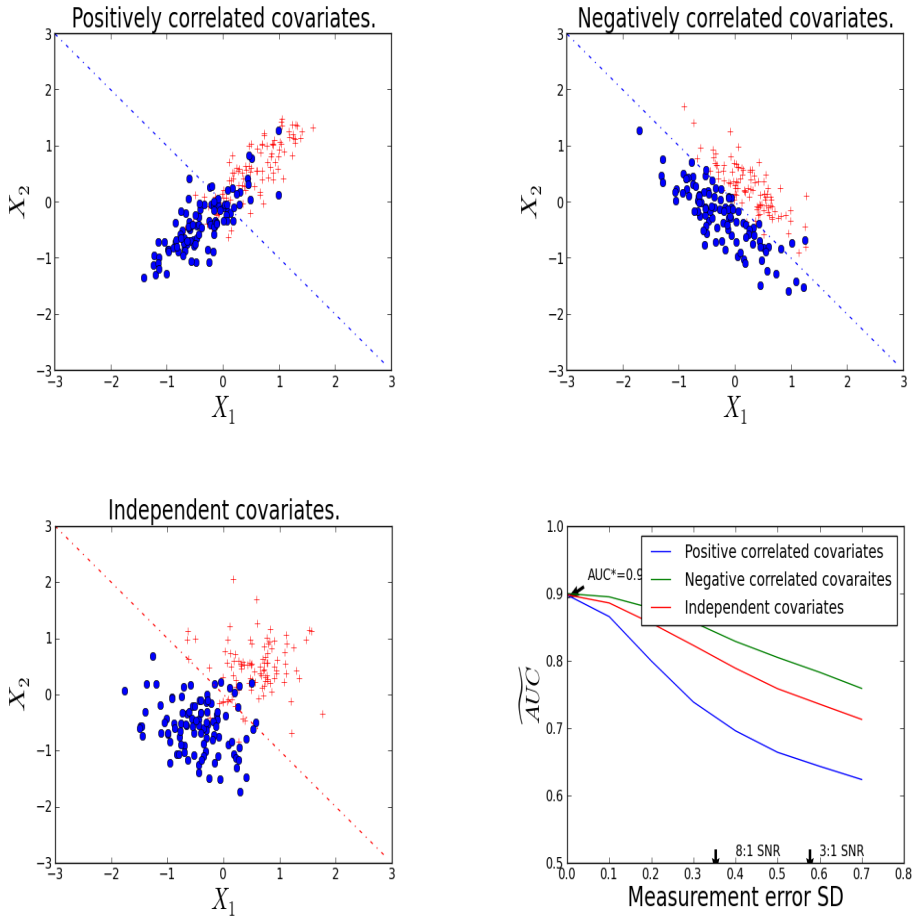

Figure 2.1: Example of the impact of the covariance structure of the covariates, Σ_x on the decline of predictive accuracy due to measurement error. Upper Left plot shows the data when the true covariates are highly positively correlated with $r = 0.8$; upper right plot shows the data when the true covariates are highly negatively correlated with $r = -0.8$; lower left shows the data when the two true covariates are independent; lower right plot shows the relationship between Bayes' rule predictive accuracy $\widehat{AUC}$ and the magnitude of measurement error for three structures of Σ_x.

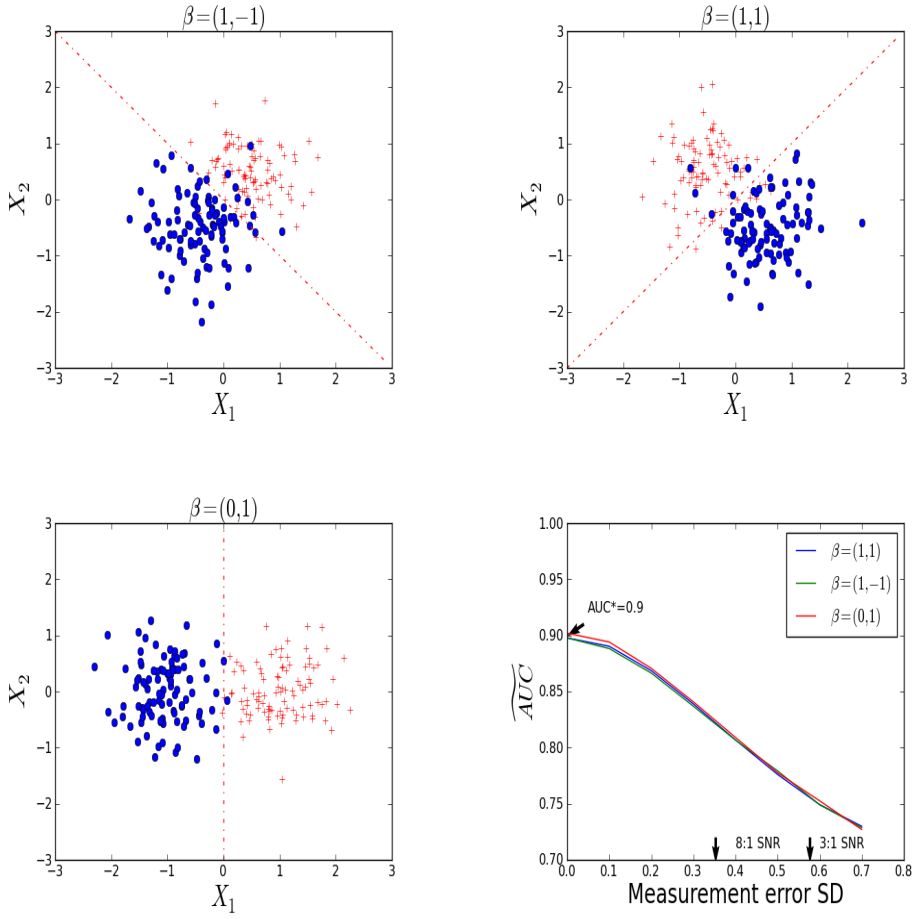

Figure 2.2: Example of the impact of the true regression coefficients β on the decline of predictive accuracy due to measurement error. Upper Left plot shows the data when $\beta = c_1(1, -1)$; upper right plot shows the data when $\beta = c_2(1, 1)$; lower left shows the data when $\beta = c_3(0, 1)$; c_1, c_2 and c_3 are constants that make the Bayes' rule predictive accuracy with no measurement error euqal to 0.9. Lower right plot shows the relationships between the Bayes' rule predictive accuracy $\widetilde{AUC}$ and the magnitude of measurement error for three structures of β.

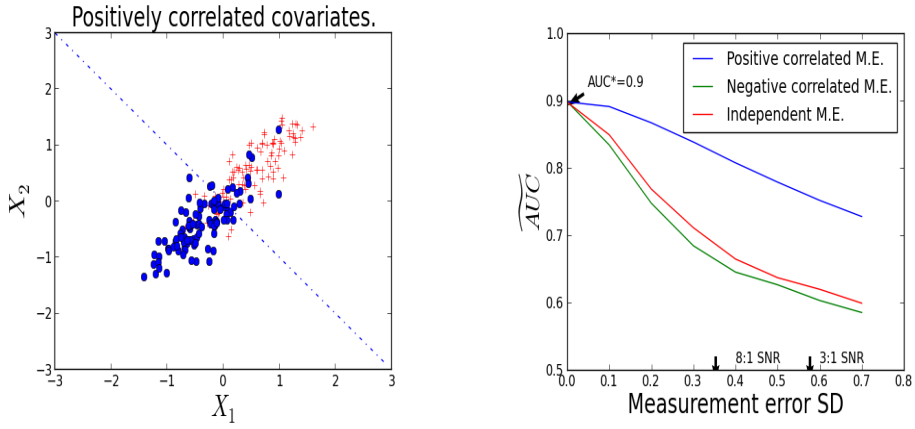

Figure 2.3: Example of the impact of the covariance structure of measurement error, Σ_η on the decline of predictive accuracy due to measurement error. Left plot shows the data without measurement error and right plot shows the relationships between the Bayes' rule predictive accuracy $\widetilde{AUC}$ and the magnitude of measurement error for three structures of Σ_η.

2.2 Impact of covariate measurement error on AUC for linear risk scores

2.2.1 Errors in predictive models

In predictive analysis, the decline of predictive accuracy may be caused by a misspecified working model, inadequate covariates, sampling error, and/or covariate measurement error, among other possible factors. Error caused by a misspecified working model could be reduced by using modern flexible regression and model selection methods. Errors caused by inadequate covariates could be reduced by adding more covariates. Sampling error could be reduced by increasing the sample size. Our goal here is to focus on the effects of covariate measurement error. Therefore for comparison reasons, we want the other factors remain the same throughout our study. Thus we use a fixed training set size ($p = 10, n = 400$) and a fixed outcome-generating model (e.g. logistic regression). We also hold fixed the Bayes' rule predictive accuracy for the model with no measurement error.

When every configuration has the same Bayes' rule predictive performance, the same outcome-generating model and the same regression technique, the decline of predictive accuracy is solely due to sampling error and measurement error. Then we can look at the relationship between measurement error and the decline of predictive accuracy without considering other factors.

2.2.2 Predictive model

We focus on logistic regression model as a predictive model, and we use AUC (area under the ROC) to quantify the predictive accuracy. The logistic regression model is given by:

$$\log(\frac{p}{1-p}) = \beta' X; \quad p = E[Y|X]$$

where $y \in \{0, 1\}$ is a binary response variable, X is a vector of observed covariates, β is a vector of regression coefficients. We use $\hat{\beta}' X$ as a risk score.

2.2.3 Predictive accuracy *AUC*

When a predictive model for a dichotomous outcomes produces a continuous "risk score", a standard procedure for measuring its predictive accuracy is to apply the model to a test set of subjects whose true responses are known, allowing us to obtain unbiased estimates of sensitivity and specificity for various risk score thresholds. Sensitivity is defined as the probability that a truly "positive" subject is predicted as positive, and specificity is the probability that a truly negative subject is predicted as negative. As the risk score threshold is varied, a non-decreasing relationship between sensitivity and 1-specificity results. This curve is called the receiver operating characteristic (ROC) curve. The area under ROC curve (AUC) (*Dodd and Pepe* (2003), *Hanley and McNeil* (1982)) is a standard index for diagnostic accuracy, ranging from 0 to 1, with greater values indicating greater accuracy, and $AUC = 0.5$ indicates predictive equivalent to random assignment.

An alternative interpretation of the AUC is that it represents the probability that a randomly chosen positive example is correctly rated (ranked) with greater suspicion than a randomly chosen negative example. In logistic regression, we use R_1 to represent the risk score for subjects in group $Y = 1$, and R_2 to represent the risk score for subjects in group $Y = 0$. Then, we can write

$$AUC = P(R_1 > R_2)$$

Since AUC is an average measure of predictive performance, so is not dependent on a decision threshold. In addition it is invariant to the marginal class probabilities.

2.2.4 Predictive performance under no measurement error

Define $AUC^\star$ as the population AUC for a given linear risk score $\beta' X$ with no measurement or sampling error. We then define $\widetilde{AUC}$ as the realized AUC under measurement error, and when estimating coefficients β from training data as $\hat{\beta}$. The difference between $\widetilde{AUC}$ and $AUC^\star$ is the decline of predictive accuracy is our main interest.

In order to find the the attibutes of the data generating model that have influence on the decline of predictive accuracy, more insight into AUC is needed.

2.2.5 Analysis of AUC under measurement error

In binary classification, we have two groups of subjects. Suppose in the group where $Y = 1$, we have the true covariate X_1 and the risk score $R_1 = \beta' X_1$; in the group where $Y = 0$, we have the true covariate X_2 and the risk score $R_2 = \beta' X_2$. The risk scores are related to a monotone single-index model (*Xia* (2006)):

$$P(Y = 1|X) = F(\beta' X) = p, \tag{2.1}$$

where $0 \leq F(\cdot) \leq 1$, so that $0 \leq p \leq 1$, and F is a monotone increasing function on $\beta' X$. For example, $F(z) = \frac{e^z}{1+e^z}$ for logit model. The predictive performance for the limiting risk score $\beta' X$ under no measurement error is given by

$$AUC^{\star} = P(R_1 > R_2) = P(\beta' X_1 > \beta' X_2).$$

If X is observed with measurement error η, then in the groups where $Y = 1$ and $Y = 0$, the observed covariate has the form

$$X_1^{obs} = X_1 + \eta_1, \quad X_2^{obs} = X_2 + \eta_2, \tag{2.2}$$

where $\eta_j | X_j$ are random measurement error with mean 0 and covariance matrix Σ_{η_j}, $j = 1, 2$. Let $\hat{\beta}$ be the estimated regression coefficient of Y on X_{obs} and $\hat{\beta} \sim N(\tilde{\beta}, \Sigma_{\tilde{\beta}})$, where $\tilde{\beta}$ is the limiting estimated regression coefficient of Y on X_{obs}. Our focus in this chapter is $\widetilde{AUC}$ with no sampling error or the limiting $\hat{\beta}$, $\tilde{\beta}$. When measurement error exists, the estimated risk score is defined as

$$\begin{aligned}
\tilde{R}_1 &= \tilde{\beta}' X_1^{obs} = (\tilde{\beta} - \beta)' X_1 + \tilde{\beta}' \eta_1 + R_1, \\
\tilde{R}_2 &= \tilde{\beta}' X_2^{obs} = (\tilde{\beta} - \beta)' X_2 + \tilde{\beta}' \eta_2 + R_2.
\end{aligned}$$

The realized predictive performance

$$\begin{aligned}
\widetilde{AUC} &= P(\tilde{R}_1 > \tilde{R}_2) \\
&= P((\tilde{\beta} - \beta)' X_1 + \tilde{\beta}' \eta_1 + R_1 > (\tilde{\beta} - \beta)' X_2 + \tilde{\beta}' \eta_2 + R_2) \\
&= P(R_1 - R_2 + (\tilde{\beta} - \beta)'(X_1 - X_2) + \tilde{\beta}'(\eta_1 - \eta_2) > 0) \\
&= P(D + S_1 + S_2 > 0),
\end{aligned} \tag{2.3}$$

where $D = R_1 - R_2$, $S_1 = (\tilde{\beta} - \beta)'(X_1 - X_2)$, $S_2 = \tilde{\beta}'(\eta_1 - \eta_2)$ are constructed by random vectors $(X_1, X_2, \eta_1, \eta_2)$.

To learn how $\widetilde{AUC}$ is influenced by the factors D, S_1 and S_2 from the above equation, the moment structures of D, S_1, S_2 might be related from our initial intuition and we could see that the first and second moments of D, S_1 and S_2 will influence $\widetilde{AUC}$ if the following two assumptions hold.

(A1) $D + S_1 + S_2$ is from a location scale family that

$$D + S_1 + S_2 = \sigma u + \theta,$$

where $\sigma^2 = \mathrm{Var}(D + S_1 + S_2)$, $\theta = E(D + S_1 + S_2)$, u is also from location scale family with mean 0 and variance 1.

(A2) $E(\tilde{R}_1) > E(\tilde{R}_2)$ or $\theta = E(D + S_1 + S_2) > 0$.

From $(A1)$,

$$\widetilde{AUC} = P(D + S_1 + S_2 > 0) = P(\sigma u + \theta > 0) = 1 - F_u(-\frac{\theta}{\sigma}),$$

where F_u is the cdf of u and a monotone increasing function. Therefore $\widetilde{AUC}$ will increase when θ increases. If $\theta < 0$, $\widetilde{AUC}$ will increase when σ increases; If $\theta > 0$, $\widetilde{AUC}$ will increase when σ decreases. From $(A2)$, we know that θ is always greater than 0, then $\widetilde{AUC}$ will increase when σ decreases. $(A2)$ is always true since in our model, $P(Y = 1)$ is a monotone increasing function of the risk score $\beta' X$ from equation 2.1, then the group where $Y = 1$ always has a higher risk score than the group where $Y = 0$ on average. $AUC > 0.5$ is reasonable since the performance of the classification will always be better than random assignment where $AUC = 0.5$. Therefore $E(\tilde{R}_1) > E(\tilde{R}_2)$ or $\theta = E(D+S_1+S_2) > 0$. As a conclusion, we imply $\widetilde{AUC}$ will increase when $E(D + S_1 + S_2)$ increases, and will decrease when $\mathrm{Var}(D + S_1 + S_2)$

increases.

Expand the mean and variance of $D + S_1 + S_2$ to nine terms, which are the mean of D, S_1, S_2, the variance of D, S_1, S_2, and the covariance between D, S_1, S_2. They are listed below:

(1) $E(D) = E(R_1 - R_2) = E(\beta'(X_1 - X_2))$,

(2) $E(S_1) = E((\tilde{\beta} - \beta)'(X_1 - X_2))$,

(3) $E(S_2) = E(\tilde{\beta}'(\eta_1 - \eta_2))$,

(4) $\text{Var}(D) = \beta'\Sigma_{X_1}\beta + \beta'\Sigma_{X_2}\beta$,

(5) $\text{Var}(S_1) = (\tilde{\beta} - \beta)'\Sigma_{X_1}(\tilde{\beta} - \beta) + (\tilde{\beta} - \beta)'\Sigma_{X_2}(\tilde{\beta} - \beta)$,

(6) $\text{Var}(S_2) = \tilde{\beta}'\Sigma_{\eta_1}\tilde{\beta} + \tilde{\beta}'\Sigma_{\eta_2}\tilde{\beta}$,

(7) $\text{Cov}(S_1, D) = (\tilde{\beta} - \beta)'\Sigma_{X_1}\beta + (\tilde{\beta} - \beta)'\Sigma_{X_2}\beta$,

(8) $\text{Cov}(S_2, D) = \tilde{\beta}'\text{Cov}(\eta_1 - \eta_2, X_1 - X_2)\beta$,

(9) $\text{Cov}(S_1, S_2) = (\tilde{\beta} - \beta)\text{Cov}(\eta_1 - \eta_2, X_1 - X_2)\tilde{\beta}$.

Here we only focus on the effect of measurement error on the decline of $\widetilde{AUC}$ from $AUC^\star$ while holding the model fixed. Then factors (1) and (4) which are not functions of measurement error will not impact the decline of $\widetilde{AUC}$ from $AUC^\star$, so will be removed from the list. While S_2 is a direct function of measurement error and S_1 is also impacted by measurement error through $\tilde{\beta}$ which is the estimated regression coefficient of Y on X_{obs} with measurement error. With the assumptions from Equation 2.2, the measurement error η_1, η_2 have zero means and are independent with our true covariates X_1, X_2, then $\text{Cov}(\eta_1 - \eta_2, X_1 - X_2) = 0$, so S_2 is independent with S_1 and D. Factors (8) and (9) are approximately zero and will be removed from the list. Also $E(S_2) = 0$ since the measurement error has zero mean, then factor (3) will be

removed from the list. At last, there are four factors remaining here that influence the decline of $\widetilde{AUC}$ from $AUC^{\star}$ due to measurement error, they are

(1) $E(S_1) = E((\tilde{\beta} - \beta)'(X_1 - X_2))$,

(2) $\mathrm{Var}(S_1) = (\tilde{\beta} - \beta)'\Sigma_{X_1}(\tilde{\beta} - \beta) + (\tilde{\beta} - \beta)'\Sigma_{X_2}(\tilde{\beta} - \beta)$,

(3) $\mathrm{Var}(S_2) = \tilde{\beta}'\Sigma_{\eta_1}\tilde{\beta} + \tilde{\beta}'\Sigma_{\eta_2}\tilde{\beta}$,

(4) $\mathrm{Cov}(S_1, D) = (\tilde{\beta} - \beta)'\Sigma_{X_1}\beta + (\tilde{\beta} - \beta)'\Sigma_{X_2}\beta$.

The relationships of these four factors and $\widetilde{AUC}$ are that $\widetilde{AUC}$ will increase when $E(S_1)$ increase, $\widetilde{AUC}$ will increase when $\mathrm{Var}(S_1)$, $\mathrm{Var}(S_2)$ and $\mathrm{Cov}(S_1, D)$ decrease. Also these relationships could be seen from the lower bound of $\widetilde{AUC}$ using Chebyshev inequality and Vysochanskii-Petunin inequality.

Chebyshev inequality: Let M be a random variable with expected value μ and finite variance σ^2. Then for any real number $k > 0$,

$$P(|M - \mu| \geq K\sigma) \leq \frac{1}{K^2}$$

Equality holds when:

$$M = \begin{cases} \text{-1} & \frac{1}{2K^2} \\ 0 & 1 - \frac{1}{K^2} \\ 1 & \frac{1}{2K^2} \end{cases}$$

Apply Chebyshev inequality to the equation of $\widetilde{AUC}$ (2.3):

$$\begin{aligned}
\widetilde{AUC} &= P(\tilde{R}_1 > \tilde{R}_2) \\
&= P(\tilde{R}_1 - \tilde{R}_2 - E(\tilde{R}_1 - \tilde{R}_2) > -E(\tilde{R}_1 - \tilde{R}_2)) \\
&= 1 - P(\tilde{R}_2 - \tilde{R}_1 - E(\tilde{R}_2 - \tilde{R}_1) > E(\tilde{R}_1 - \tilde{R}_2)) \\
&= 1 - P(\tilde{R}_2 - \tilde{R}_1 - E(\tilde{R}_2 - \tilde{R}_1) > \frac{E(\tilde{R}_1 - \tilde{R}_2)}{SD(\tilde{R}_1 - \tilde{R}_2)} SD(\tilde{R}_1 - \tilde{R}_2)) \\
&= 1 - P(M > K\sigma) \\
&\geq 1 - \frac{1}{K^2}.
\end{aligned}$$

Where

$$\begin{gathered}
M = \tilde{R}_2 - \tilde{R}_1 - E(\tilde{R}_2 - \tilde{R}_1), \\
\sigma = SD(\tilde{R}_1 - \tilde{R}_2), \\
K = \tfrac{E(\tilde{R}_1 - \tilde{R}_2)}{SD(\tilde{R}_1 - \tilde{R}_2)}, \\
E(\tilde{R}_1 - \tilde{R}_2) = E(R_1 - R_2) + E(S_1), \\
SD(\tilde{R}_1 - \tilde{R}_2) = \sqrt{\mathrm{Var}(S_1) + \mathrm{Var}(S_2) + \mathrm{Cov}(S_1, R_1 - R_2) + \mathrm{Var}(R_1 - R_2)}.
\end{gathered}$$

The lower bound of $\widetilde{AUC}$ is an increasing function of K, and then is an increasing function of $E(S_1)$ and a decreasing function of $\mathrm{Var}(S_2)$, $\mathrm{Var}(S_1)$, $\mathrm{Cov}(S_1, R_1 - R_2)$. Since the lower bound of $\widetilde{AUC}$ using chebyshev's inequality is not tight enough, we could use Vysochanskii-Petunin inequality which assume unimodality of random variable X instead. But the relationships of our four properties and $\widetilde{AUC}$ do not change.

Vysochanskii-Petunin inequality: Let M be a random variable with expected value

μ and finite variance σ^2. Then for any real number $k > 0$,

$$P(|M - \mu| \geq K\sigma) \leq \frac{4}{9K^2}.$$

From the theoretical derivation of AUC, we know that there are four properties of the data generating model influence the decline of AUC caused by measurement error. We imply the relationship of these four properties and $\widetilde{AUC}$ under the assumption of the location-scale family and from the lower bound of $\widetilde{AUC}$. Next, we will confirm these relationships by simulation studies and explore more about the relationships between these four properties themselves, then decide the miminum number of properties which influence the amount of decline of $\widetilde{AUC}$.

2.3 Simulation approach

2.3.1 Gene expression data

Since we want the data generating model in our simulation study to be close enough to the real data, many parameters or attributes in my simulation study is chosen by calibrating them from the real data.

We use the data set for Microarray Innovations in Leukemia (MILE) study program with n=2096 sample sizes devided into 18 classes and over thousands of genes. We focus on the two classes "CLL" and "AML with normal karyotype + other abnormalities" which have the largest sample sizes which are 448 and 351. We split the data evenly into two subgroups, one for identifying the significant genes used as covariates from some criterions, (e.g. who have their gene expression values most different comparing these two classes using t statistics), the other subgroup is used to estimate the distribution of $\hat{\beta}$ which is the estimated coefficient of the logistic regression of the observed two classes on the gene covariates we just selected.

Here we use three methods to select the genes to be used in our simulation study.

One is choose 10 genes having the highest t values comparing the two classes, and we could also choose the next 10 genes having the highest t values except the first 10 genes and so on. The limitation of this method is the genes we choose are highly positive or negative correlated with each other (e.g. $\bar{\rho} = 0.75$ which is the average pairwise correlation of the first 10 genes having the highest t values in MILE data), the high multicollinearty will lead to opposite sign of estimated coefficients $\hat{\beta}$ and larger covariance matrix of $\hat{\beta}$, which will lead to more unstable result.

The second method is that we choose 10 genes whose pairwise correlation less than some fixed value (e.g. $\rho < 0.7$) and have higher t values also. In more details, we order our genes with their t values from high to low and we could choose the first gene having the highest t value. Then search the genes by the order until its correlation with the first gene is less than 0.7. Then search the third gene until its correlation with the first and the second are both less than 0.7 and go on. This method controls the multicollearity of the genes and the 10^{th} gene still has t value greater than 2 in MILE data.

The third method is that we choose 10 genes randomly from all the genes and they are expected to be independent with each other. From these three methods, we could generate three structures of Σ_X in which the distribution of the pairwise correlations between covariates are approximately normal with mean μ and standard deviation σ. In MILE data, $\mu = 0.7$, $\sigma = 0.1$ for the first method; $\mu = 0.5$, $\sigma = 0.1$ for the second method; $\mu = 0.2$, $\sigma = 0.1$ for the third method.

2.3.2 Simulation steps

1. We construct the true covariates X without measurement error, having covariance matrix Σ_x. Figure 2.4 are the examples of pattern of elements in covariance matrix of X using gene expression data from MILE study by three different methods illustrated above. From these examples, we could see that

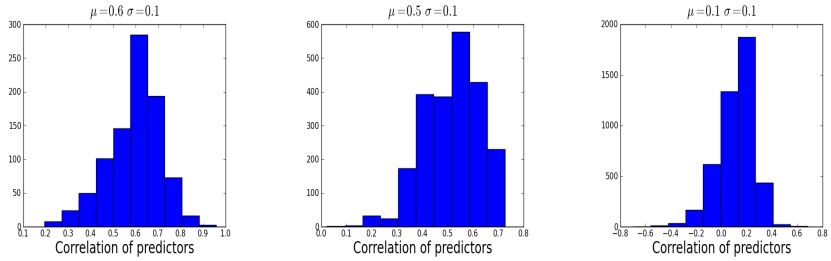

Figure 2.4: Left plot is the histogram of the correlations of pairs of genes selected from method 1; middle plot is the histogram of the correlations of pairs of genes selected from method 2; third plot is the histogram of the correlations of pairs of genes selected from method 3.

the elements in covariance matrix of predictors are approximately normally distributed with different mean and standard deviation. We need the elements in Σ_x in my simulation study has the same distribution with that in MILE study. Let $\Sigma_x = I + FF'$, where $F = (f_1, \cdots, f_p)^T$. The elements of corresponding Σ_x are $\rho_{ij} = \frac{f_i f_j}{\sqrt{(1+f_i^2)(1+f_j^2)}}, i \neq j$. We used a numerical optimization scheme to optimize the fit of elements in $\Sigma_x = I + FF'$ to a normal distribution over F. The Kolmogorov-Smirnov distance was used to assess the fit. The normal distribution of the elements in Σ_X has three different sets of mean and standard deviation: ($\mu = 0.7$, $\sigma = 0.1$); ($\mu = 0.5$, $\sigma = 0.1$); ($\mu = 0.2$, $\sigma = 0.1$).

2. Generate true coefficients β with known structure.

   We have two ways to generate β:

   (1) Use $\beta_j = \pm c \times 2^{(1-j)d}, j = 0, 1, \cdots$ to define a family of true coefficient vectors with different patterns. Figure 2.5 is an example of this kind of β.

   (2) Generate a set of population β vectors by sampling from the distribution $N(\beta_0, \Sigma_0)$, where β_0 is the estimated logistic regression coefficient for the observed two classes and selected gene expression data set in MILE study.

3. Control $AUC^\star = 0.9$ and set balance of outcome to be 0.6 with known struc-

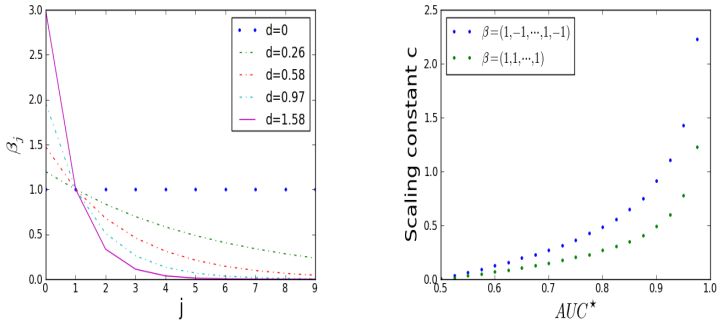

Figure 2.5: left plot is examples of different structure of true coefficient β has structure $\beta_j = \pm c \times 2^{(1-j)d}, j = 0, 1, \cdots$ with $c = 1$ and different value of parameter d; Right plot is examples of scaling constant c with different $AUC^\star$ and β.

ture of X and β by setting the scaling constant c of β using bisection numerical method. $AUC^\star$ is the predictive performance with no measurement and sampling error. Since we are only interested in the effect of measurement error on predictive performance, for comparison reason, we need every data generating model in my simulation study have the same $AUC^\star$. From figure 2.5, we know the scaling constant c increases with the increase of $AUC^\star$. Different pattern of β has different scaling constant c.

4. Calibrating the magnitude of measurement error.

   In measurement error analysis, it is almost always necessary to have an internal or external measure of the level of measurement error. For gene expression analysis, internal replication is uncommon. Therefore, to estimate the level of measurement error among gene expression predictors, we use a triplicated expression array experiment on a panel of 59 cell lines. Specifically, given three replicates X_1, X_2, X_3, $\frac{1}{2}\sum_{i=1}^{3}(x_i - \bar{x})^2$ unbiasedly estimates the measurement error for a particular gene in a particular cell line. We then average over the cell lines and consider the relationship between measurement error standard deviation and overall standard deviation in Figure 2.6. Focusing on the trend

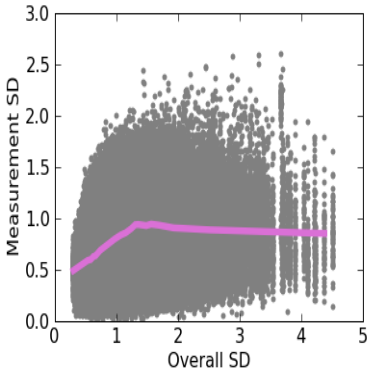

Figure 2.6: Plot of measurement error standard deviation and overall standard deviation in a triplicated expression array data.

line in the left panel, we see that measurement error standard deviation increases linearly at first, then becomes independent of the overall standard deviation. This suggests a fixed additive variance due to measurement error, except for the genes that are nearly constant. Many of these nearly constant genes are non-responsive probes, where measurement error would not be expected to be detectable. Taking the middle of the range as a nominal value, we arrive at signal-to-noise ratio (SNR) estimates from 3:1 to 8:1.

5. Simulate binary response Y with linear predictor $\beta' X$ by

$$P(Y = 1) = \frac{\exp(\beta' X)}{1 + \exp(\beta' X)},$$

then add independent measurement error with different magnitude probably between 0 to 1 to standardized X to create X_{obs}, then calculate $\widetilde{AUC}$ by using estimated risk score $\hat{\beta}' X$, where $\hat{\beta}$ is the estimated coefficient of logistic regression of Y on X_{obs}. Repeat step 5 100 times to estimate $\tilde{\beta}$ by averaging the values of $\hat{\beta}$ in each repetition and then since the covariance matrix of measurement error Σ_η, covariance matrix of the true covariates Σ_X, true coefficient β are all known, we could estimate the value of these four factors of the data generating model

(1) $E(S_1) = E((\tilde{\beta} - \beta)'(X_1 - X_2))$,

(2) $\text{Var}(S_1) = (\tilde{\beta} - \beta)'\Sigma_{X_1}(\tilde{\beta} - \beta) + (\tilde{\beta} - \beta)'\Sigma_{X_2}(\tilde{\beta} - \beta)$,

(3) $\text{Var}(S_2) = \tilde{\beta}'\Sigma_{\eta_1}\tilde{\beta} + \tilde{\beta}'\Sigma_{\eta_2}\tilde{\beta}$,

(4) $\text{Cov}(S_1, D) = (\tilde{\beta} - \beta)'\Sigma_{X_1}\beta + (\tilde{\beta} - \beta)'\Sigma_{X_2}\beta$.

and make graphs of each of the four factors and $\widetilde{AUC}$.

In simulation step 5, we add independent measurement errors to the true covariate X. But in real data, the measurement errors could be dependent with each other and the covariance matrix of measurement error does not equal to identity matrix, $\Sigma_\eta \neq I$. But we could guarantee that with some matrix transformations, we could assume that $\Sigma_\eta = I$ without lose of generality of the data generating models. If we have the logistic regression model

$$\log(\frac{p}{1-p}) = \beta' X; \quad p = E[Y|X],$$

with true regression coefficient β, true covariate X, and the measurement error added to the covariate η, with $E(\eta|X) = 0$, and covariance matrix Σ_η. Then the risk score with measurement error is $\beta'(X + \eta)$. By Cholesky decomposition, $\Sigma_\eta = RR'$, where R is a lower triangular matrix. Let

$$\tilde{\eta} = R'^{-1}\eta, \quad \tilde{X} = R'^{-1}X, \quad \tilde{\beta} = R\beta,$$

Then the risk score with measurement error is $\tilde{\beta}'(\tilde{X} + \tilde{\eta}) = \beta'(X + \eta)$ will not change with $\Sigma_{\tilde{\eta}} = I$.

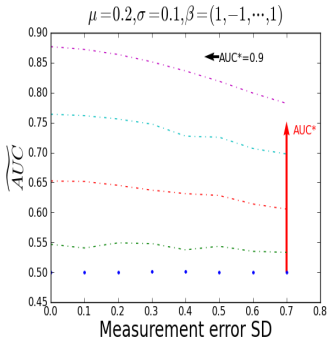

Figure 2.7: Plot of $\widetilde{AUC}$ and measurement error SD magnitude with different level of $AUC^{\star}$.

2.4 Simulation results for logistic regression

2.4.1 Question 1: Will the effect of covariate measurement error on predictive performance be different with different level of $AUC^{\star}$?

Since we only interest in the decline of predictive accuracy, $AUC^{\star} - \widetilde{AUC}$ due to measurement error, whether or not to fix $AUC^{\star}$ is a question. With different level of $AUC^{\star}$, if $AUC^{\star} - \widetilde{AUC}$ is not effected by $AUC^{\star}$, then we do not need to fix $AUC^{\star}$.

From intuition, $0 \leq \widetilde{AUC} \leq AUC^{\star}$ due to measurement error. If $AUC^{\star} \approx 0.5$, then $\widetilde{AUC}$ also ≈ 0.5. We expect $AUC^{\star} - \widetilde{AUC}$ to be greater when $AUC^{\star}$ is greater. We could also use simple simulation to check this.

Choose Σ_x with $\mu = 0.2, \sigma = 0.1$, $\beta = (1, -1, \cdots, 1, -1)$, $\Sigma_\eta = s^2 I$ with s ranging from 0.0 to 0.7. Vary $AUC^{\star}$ from 0.5 to 0.9, from figure 2.7, we could see that the decline of predictive accuracy is different with different level of $AUC^{\star}$ due to the same amount of measurement error. In our following simulation study, we always use $AUC^{\star} = 0.9$.

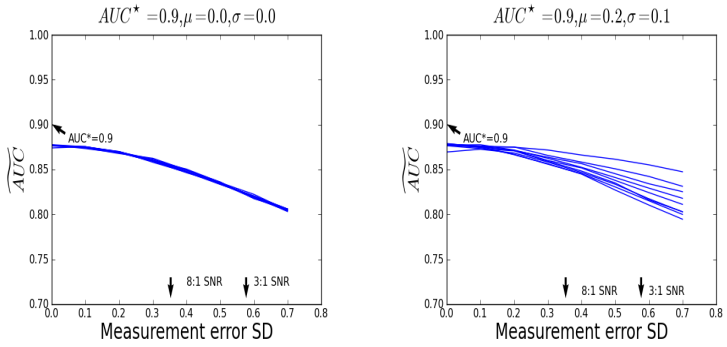

Figure 2.8: Left Plot is the plot of $\widetilde{AUC}$ and measurement error SD magnitude with different β when $\Sigma_x = \Sigma_\eta = I$; right Plot is the plot of $\widetilde{AUC}$ and measurement error SD magnitude with different β when the elements in Σ_x has normal distribution with $\mu = 0.2, \sigma = 0.1$ and $\Sigma_\eta = I$.

2.4.2 Question 2: For fixed $AUC^\star$, does the relationship between $\widetilde{AUC}$ and measurement error variance depend on $\beta, \Sigma_x, \Sigma_\eta$?

From the example we show at the beginning of this paper, we know that some attributes of data generating model, like $\beta, \Sigma_x, \Sigma_\eta$ will influence the decline of predictive accuracy due to measurement error and in some situations, the influence is small, in other situations, the influence is dramatic.

From figure 2.8, when $\Sigma_x = \Sigma_\eta = I$, no matter how we choose β, the decline of $\widetilde{AUC}$ is same. But when $\Sigma_x \neq \Sigma_\eta$, different structure of β has different $\widetilde{AUC}$ with the same amount of measurement error.

From the simulation study we know that some attributes of data generating model, like $\beta, \Sigma_x, \Sigma_\eta$ will influence the decline of predictive accuracy due to measurement error in most case, but we do not know exactly how these attributes influence predictive accuracy while question 3 does.

2.4.3 Question 3: What is the relationship of the four factors we find through theoretical derivation of AUC and how them effect the decline of $\widetilde{AUC}$?

From the theoretical derivation of AUC, we know there are four factors that will influence the decline of $\widetilde{AUC}$ when fixing $AUC^{\star}$. They are

(1) $E(S_1) = E((\tilde{\beta} - \beta)'(X_1 - X_2))$,

(2) $\text{Var}(S_1) = (\tilde{\beta} - \beta)'\Sigma_{X_1}(\tilde{\beta} - \beta) + (\tilde{\beta} - \beta)'\Sigma_{X_2}(\tilde{\beta} - \beta)$,

(3) $\text{Var}(S_2) = \tilde{\beta}'\Sigma_{\eta_1}\tilde{\beta} + \tilde{\beta}'\Sigma_{\eta_2}\tilde{\beta}$,

(4) $\text{Cov}(S_1, D) = (\tilde{\beta} - \beta)'\Sigma_{X_1}\beta + (\tilde{\beta} - \beta)'\Sigma_{X_2}\beta$.

We also imply how they effect the decline of $\widetilde{AUC}$ that greater $E(S_1)$ lead to higher $\widetilde{AUC}$, greater $\text{Var}(S_2)$, $\text{Var}(S_1)$, $\text{Cov}(S_1, R_1 - R_2)$ will lead to lower $\widetilde{AUC}$ from the theoretical derivation of the definition of AUC. Then we will check if we have the similar relationships in our simulation study.

In our simulation study, we generate covariate X with covariance matrix Σ_x, whose off-diagonal element has approximately normal distribution with three sets of mean and standard deviation, (1) $\mu = 0.7$, $\sigma = 0.1$; (2) $\mu = 0.5$, $\sigma = 0.1$; (3) $\mu = 0.2$, $\sigma = 0.1$. True coefficient β is generate with distribution $N(\beta_0, \Sigma_0)$, where β_0 is the estimated logistic regression coefficient for the observed two classes and selected gene expression data set in MILE study from three different methods. Control $AUC^{\star} = 0.9$ by multiplying a constant c to β and then the true binary outcome Y is generated with the linear predictor $c\beta'X$. Add independent measurement error with magnitude from 0 to 1 to the covariates X, then the estimated predictive accuracy $\widetilde{AUC}$ and the estimates of the four factors are calculated.

Figure 2.9 is the plots of the relationships between $\widetilde{AUC}$ and the magnitude of measurement error with the three sets of mean and standard deviation of the

distribution of the elements in Σ_x. The plot shows that overall $\widetilde{AUC}$ decline with the increasing magnitude of measurement error and the variation of $\widetilde{AUC}$ due to different structure of true coefficient β increase with increasing magnitude of measurement error. When the distribution of the elements in Σ_x has lower mean, the decline of $\widetilde{AUC}$ due to measurement error is smaller.

Figure 2.10 shows the relationship between $\widetilde{AUC}$ and the four factors of the data generating model with the three sets of mean and standard deviation of the distribution of the elements in Σ_x. We could see that $E(S_1)$ has a positive relationship with $\widetilde{AUC}$, while $\text{Var}(S_2)$, $\text{Var}(S_1)$, $\text{Cov}(S_1, X_1 - X_2)$ have negative relationships with $\widetilde{AUC}$.

Figure 2.11 shows the relationships of these four factors themselves and we could see that $E(S_1)$, $\text{Var}(S_1)$ and $\text{Cov}(S_1, X_1 - X_2)$ are highly dependent with each other, while $\text{Var}(S_2)$ is independent of them. We choose $\text{Var}(S_1)$ and $\text{Var}(S_2)$ as the main factors that influence the decline of $\widetilde{AUC}$.

As a conclusion, in binary outcome predictive, predictive performance is negatively affected by the increase of magnitude of measurement error. Moreover, the effect is influenced by other attributes of data generating model. From the theoretical derivation of predictive accuracy AUC, we find that there are four factors might influence the decline of predictive accuracy $\widetilde{AUC}$ when controlling $AUC^\star$ and find similar results in the simulation study that $E(S_1)$ has a negative relationship with the decline of $\widetilde{AUC}$, while $\text{Var}(S_2)$, $\text{Var}(S_1)$, $\text{Cov}(S_1, X_1 - X_2)$ have positive relationships with the decline of $\widetilde{AUC}$. From these four factors, we find two independent factors

$$\text{Var}(S_2) = \tilde{\beta}'\Sigma_{\eta_1}\tilde{\beta} + \tilde{\beta}'\Sigma_{\eta_2}\tilde{\beta},$$

$$\text{Var}(S_1) = (\tilde{\beta} - \beta)'\Sigma_{X_1}(\tilde{\beta} - \beta) + (\tilde{\beta} - \beta)'\Sigma_{X_2}(\tilde{\beta} - \beta),$$

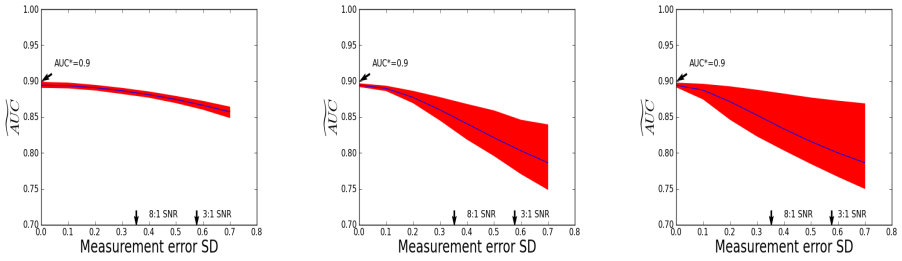

Figure 2.9: Plots of $\widetilde{AUC}$ and measurement error SD magnitude with different β generated from $N(\beta_0, \Sigma_0)$ with three situations of mean and standard deviation of the distribution of the elements in Σ_x. Left plot is the situation when $\mu = 0.2$, $\sigma = 0.1$; middle plot is the situation when $\mu = 0.5$, $\sigma = 0.1$; right plot is the situation when $\mu = 0.7$, $\sigma = 0.1$.

could be representative of these four factors and both of them have linearly positive relationships with the decline of AUC. But there is not enough theoretical proof for these relationships since there is no close form of the estimated regression coefficient $\hat{\beta}$, then we could not get a close form of the predictive accuracy AUC. However, this could be done in linear regression. We are interested in whether we could find similar properties and relationships between these properties and predictive accuracy in linear case?

2.5 Similar finding in linear case

2.5.1 linear model and predictive accuracy

Linear Model:

$$Y = \beta' X + \epsilon$$

$$E(\epsilon|X) = 0, \mathrm{Var}(\epsilon|\mathrm{X}) = \sigma_\epsilon^2$$

Define $\tilde{\beta}$ as the estimated linear regression coefficient when regressing continuous

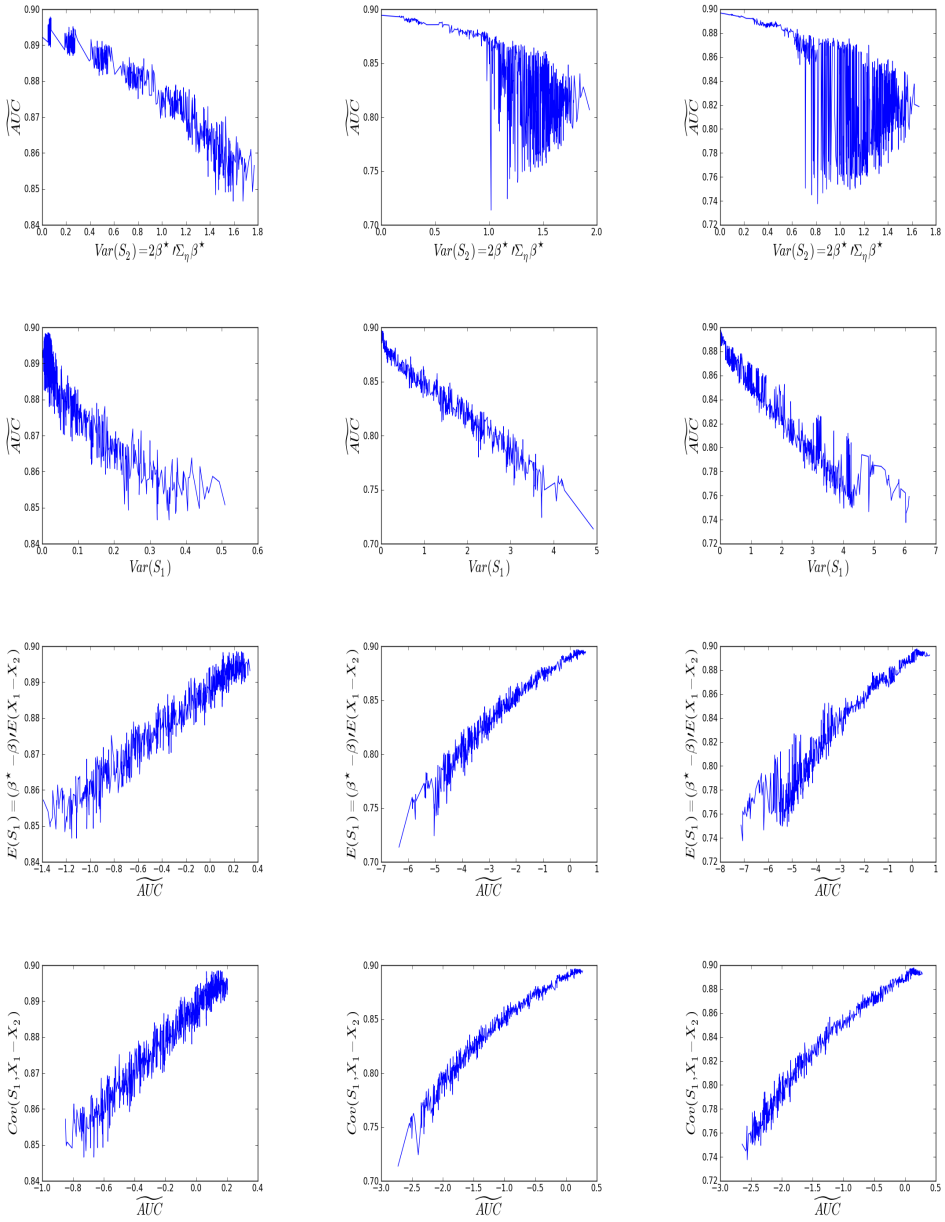

Figure 2.10: Plots of $\widetilde{AUC}$ and four factors of the data generating model with different β and measurement error in three situations of mean and standard deviation of the distribution of the elements in Σ_x. Left plot is the situation when $\mu = 0.2$, $\sigma = 0.1$; middle plot is the situation when $\mu = 0.5$, $\sigma = 0.1$; right plot is the situation when $\mu = 0.7$, $\sigma = 0.1$. First row is for factor $\mathrm{Var}(S_2)$, second row is for factor $\mathrm{Var}(S_1)$, third row is for factor $E(S_1)$, fourth row is for factor $\mathrm{Cov}(S_1, D)$.

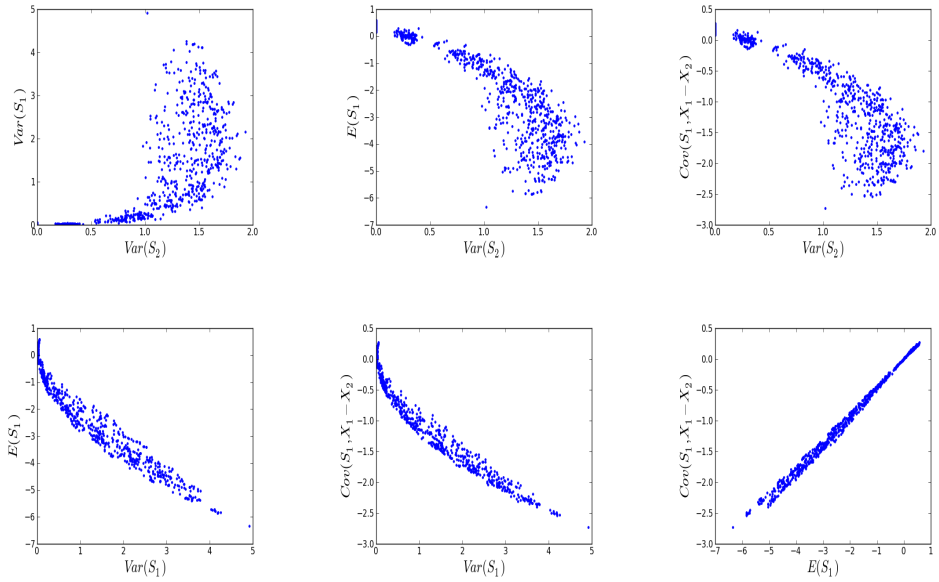

Figure 2.11: Scatterplots of these four factors in the simulation study when $\mu = 0.5$, $\sigma = 0.1$.

outcome Y on covariates X, and the sample size goes to infinity, then

$$\tilde{\beta} = E(\hat{\beta}) = (\Sigma_X + \Sigma_\eta)^{-1}\Sigma_X\beta,$$

where $\hat{\beta}$ is the estimated linear regression coefficient when regressing Y on X and $\tilde{\beta} = \beta$ when no measurement error exists. Predictive accuracy in linear model without sampling error:

$$R^2 = 1 - \frac{\|\hat{Y} - Y\|^2}{||Y - \bar{Y}||} = 1 - \frac{\|\tilde{\beta}' X_{obs} - \beta' X - \epsilon\|^2}{\beta'\Sigma_X\beta + \sigma_\epsilon^2}$$

Predictive accuracy without sampling and measurement error:

$$R^2_{ideal} = 1 - \frac{\|\tilde{\beta}' X - \beta' X - \epsilon\|^2}{\beta'\Sigma_X\beta + \sigma_\epsilon^2} = 1 - \frac{\|\sigma\|^2}{\beta'\Sigma_X\beta + \sigma_\epsilon^2} = \frac{\beta'\Sigma_x\beta}{\beta'\Sigma_x\beta + \sigma_\epsilon^2}.$$

Predictive accuracy with measurement error:

$$\widetilde{R^2} = 1 - \frac{\|(\tilde{\beta} - \beta)'X + \tilde{\beta}'\eta - \epsilon\|^2}{\beta'\Sigma_X\beta + \sigma_\epsilon^2}$$

Since ϵ, η, X are independent with each other, then $(\tilde{\beta}-\beta)'X$, $\tilde{\beta}'\eta$, ϵ are independent. Therefore,

$$\widetilde{R^2} = 1 - \frac{(\tilde{\beta} - \beta)'\Sigma_X(\tilde{\beta} - \beta) + \tilde{\beta}'\Sigma_\eta\tilde{\beta} + \sigma_\epsilon^2}{\beta'\Sigma_X\beta + \sigma_\epsilon^2}.$$

Let

$$S_1 = (\tilde{\beta} - \beta)'X, \quad S_2 = \tilde{\beta}'\eta,$$

$$\widetilde{R^2} = 1 - \frac{\mathrm{Var}(S_1) + \mathrm{Var}(S_2) + \sigma_\epsilon^2}{\beta'\Sigma_X\beta + \sigma_\epsilon^2}$$

To make each outcome-generating model has the same best predictive performance, we need to fix R^2_{ideal}, and σ_ϵ^2 is known, then $\beta'\Sigma_X\beta$ is fixed. $\widetilde{R^2}$ is affected by only two properties:

(1) $\mathrm{Var}(S_2) = \tilde{\beta}'\Sigma_\eta\tilde{\beta}$,

(2) $\mathrm{Var}(S_1) = (\tilde{\beta} - \beta)'\Sigma_X(\tilde{\beta} - \beta)$.

These two factors have negative relationships with $\widetilde{R}^2$. Then we use the simulation study to show that we could find similar relationships between $\widetilde{R}^2$ and the two factors $\mathrm{Var}(S_1)$, $\mathrm{Var}(S_2)$ of the linear model. The simulation steps are quite similar with those for binary case. We first generate covariate X with covariance matrix Σ_x, whose off-diagonal element has approximately normal distribution with three sets of mean and standard deviation, (1) $\mu = 0.7$, $\sigma = 0.1$; (2) $\mu = 0.5$, $\sigma = 0.1$; (3) $\mu = 0.2$, $\sigma = 0.1$. True coefficient β is generate with distribution $N(\beta_0, \Sigma_0)$, where β_0 is the estimated logistic regression coefficient for the observed two classes and selected gene expression data set in MILE study from three different methods. Control $R^2_{ideal} = 0.9$ by multiplying a constant c to β, and then the true continuous outcome Y is generated

with the linear predictor $c\beta' X$ by

$$Y = c\beta' X + \epsilon,$$

where $E(\epsilon|X) = 0$,$\text{Var}(\epsilon|X) = 1$. Then add independent measurement error with magnitude from 0 to 1 to the covariates X, then the estimated predictive accuracy $\widetilde{R}^2$ and the estimate of the two factors are calculated. The relationship between each of the two factors and $\widetilde{R}^2$ and the relationship between the two factors themselves are shown in figure 2.12. It indicates that larger amount of the values of the two factors will lead to larger amount of the decline of the predictive accuracy $R^2_{ideal} - \tilde{R}^2$ and this finding is consistent with the finding in binary case.

2.6 Estimation of these attributes from real data

To put our finding to practical use, we need to estimate these two factors $\text{Var}(S_2) = \tilde{\beta}'\Sigma_\eta\tilde{\beta}$ and $\text{Var}(S_1) = (\tilde{\beta} - \beta)'\Sigma_X(\tilde{\beta} - \beta)$ correctly from real data. In real data, we observe outcome Y and covariates X_{obs} with measurement error and covariance matrix of measurement error Σ_η, then we could estimate the true covariance matrix of X using the equation $\hat{\Sigma}_X = \Sigma_{X_{obs}} - \Sigma_\eta$ and calculate $\hat{\beta}$ which is the mle of β. In linear regression, we have

$$\tilde{\beta} = E(\hat{\beta}) = (\Sigma_X + \Sigma_\eta)^{-1}\Sigma_X\beta,$$

then

$$\beta = (\Sigma_X + \Sigma_\eta)\Sigma_X^{-1}\tilde{\beta}.$$

Thus the two key factors can be wrote as:

(1) $\text{Var}(S_2) = \tilde{\beta}'\Sigma_\eta\tilde{\beta}$,

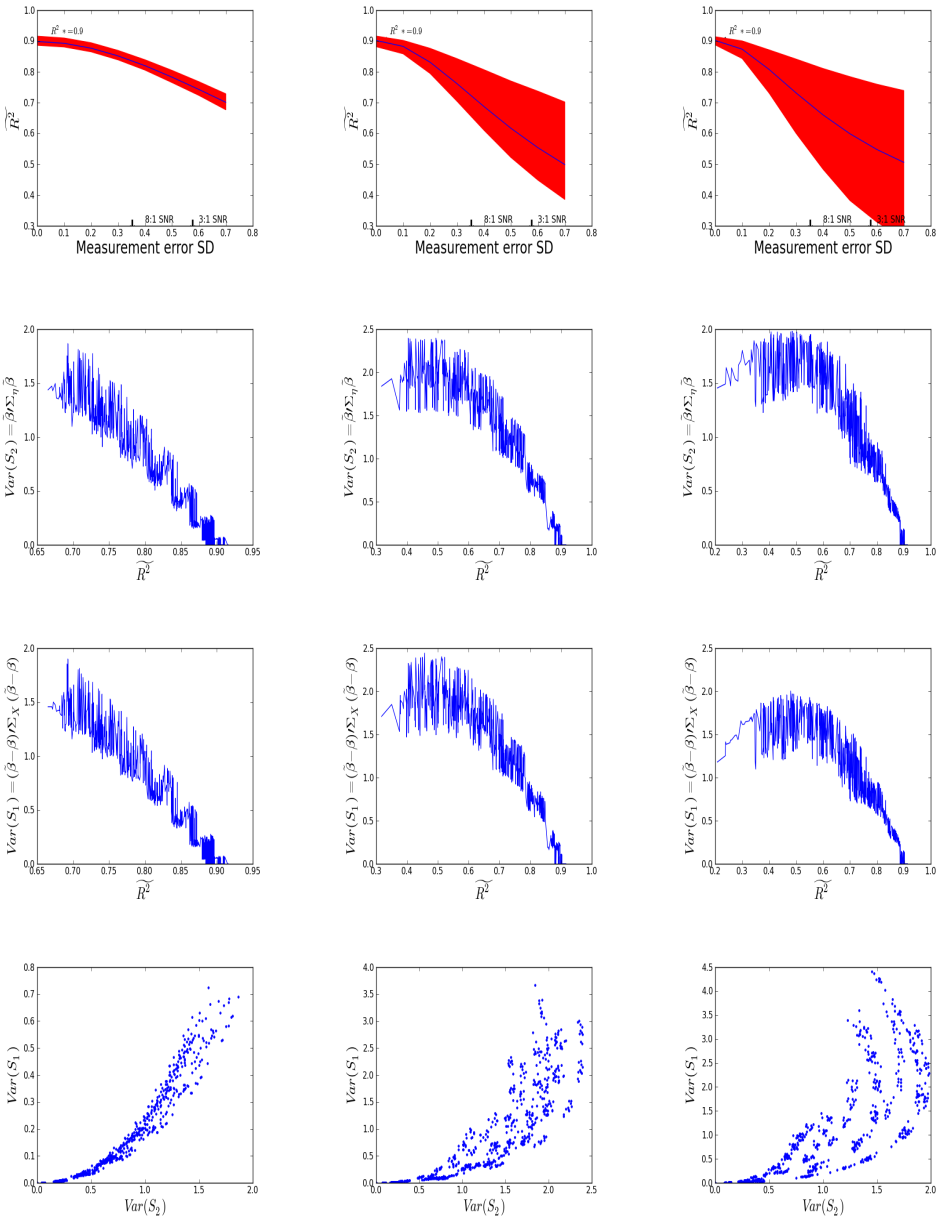

Figure 2.12: First row is the plots of $\widetilde{R^2}$ and measurement error SD magnitude with different β; second row is the plots of $\widetilde{R^2}$ and factor $\mathrm{Var}(S_2)$; third row is the plots of Plots of $\widetilde{R^2}$ and factor $\mathrm{Var}(S_1)$; fourth row is the scatterplots of $\mathrm{Var}(S_2)$ and $\mathrm{Var}(S_1)$. There are three situations of mean and standard deviation of the distribution of the elements in Σ_x. Left plot is the situation when $\mu = 0.2$, $\sigma = 0.1$; middle plot is the situation when $\mu = 0.5$, $\sigma = 0.1$; right plot is the situation when $\mu = 0.7$, $\sigma = 0.1$.

(2) $\mathrm{Var}(S_1) = \tilde{\beta}' M \tilde{\beta}$,

$M = (I - (\Sigma_X + \Sigma_\eta)\Sigma_X^{-1})'\Sigma_X(I - (\Sigma_X + \Sigma_\eta)\Sigma_X^{-1})$.

By using the equation,

$$\begin{aligned}\hat{\beta}'\Sigma_\eta\hat{\beta} &= tr(\Sigma_\eta E(\hat{\beta}\hat{\beta}')) \\ &= tr(\Sigma_\eta\Sigma_{\hat{\beta}}) + \tilde{\beta}'\Sigma_\eta\tilde{\beta} \\ &= \hat{\sigma^2} tr(\Sigma_\eta(X'_{obs}X_{obs})^{-1}) + \tilde{\beta}'\Sigma_\eta\tilde{\beta}.\end{aligned}$$

the first factor $\mathrm{Var}(S_2)$ could be estimated by

$$\mathrm{Var}(S_2) = \hat{\beta}'\Sigma_\eta\hat{\beta} - \hat{\sigma^2} tr(\Sigma_\eta(X'_{obs}X_{obs})^{-1}),$$

where $\hat{\sigma^2} = \frac{\|Y-\hat{Y}\|^2}{n-p-1}$. With the same procedure, $\mathrm{Var}(S_1)$ could be estimated by

$$\mathrm{Var}(S_1) = \hat{\beta}' M \hat{\beta} - \hat{\sigma^2} tr(M(X'_{obs}X_{obs})^{-1}).$$

In binary case, the estimation for property $\mathrm{Var}(S_2) = \tilde{\beta}'\Sigma_\eta\tilde{\beta}$ is the same as in linear case. But since there is no exact equation of $\tilde{\beta}$ and β, so we could not transform $\mathrm{Var}(S_1)$ to the format $\tilde{\beta}' M \tilde{\beta}$. Alternatively, we use a simex procedure to estimate $\tilde{\beta}-\beta$. We calculate $\hat{\beta}_1$ by regressing Y on $X+\eta$, and $\hat{\beta}_2$ by regressing Y on $X+2\eta$, then $\tilde{\beta}-\beta \approx \hat{\beta}_2-\hat{\beta}_1$. Though we could calculate Σ_x by using equation $\hat{\Sigma_x} = \Sigma_{x_{obs}} - \Sigma_\eta$, we are not able to calculate Σ_{X_1} or Σ_{X_2} in binary case, since the distribution of covariates in each group is unknown. Then in binary case, we use $\mathrm{Var}(S_1) = (\hat{\beta}_2-\hat{\beta}_1)'\hat{\Sigma}_x(\hat{\beta}_2-\hat{\beta}_1)$ instead.

Figure 2.13 is the plot of estimated and true factors in simulation study and it shows that the estimation is very precise in linear case, but the estimation of $\mathrm{Var}(S_1)$ by simex procedure is not very precise in binary case. The simulation steps are the

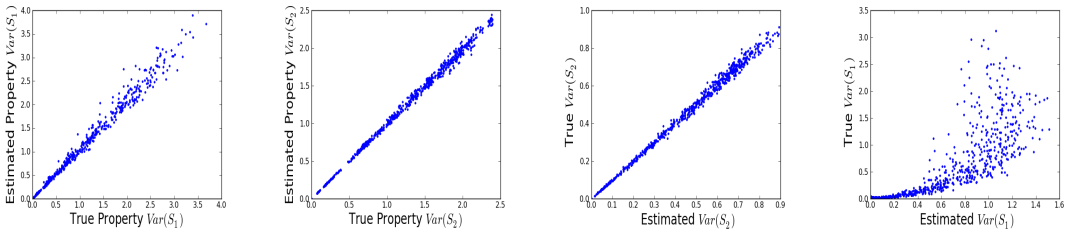

Figure 2.13: First two Plots are the scatterplots of True and estimated properties in linear case; last two Plots are the scatterplots of True and estimated properties in binary case when $\mu = 0.5$, $\sigma = 0.1$.

same with the simulation we used to get the relationships between $\widetilde{AUC}$ and the four factors in section 2.3.2.

In linear case, since we could estimate $\mathrm{Var}(S_1)$, $\mathrm{Var}(S_2)$, $\beta'\Sigma_X\beta$, σ^2 from observed data, we could calculate R^2_{ideal} and $\widetilde{R^2}$ and the decline of predictive accuracy $R^2_{ideal} - \widetilde{R^2}$. We could use a ratio of predictive accuracy due to measurement error and overall decline of predictive accuracy from 1, $\frac{R^2_{ideal} - \widetilde{R^2}}{1 - \widetilde{R^2}}$ to see how much decline of predictive accuracy is due to measurement error. If the ratio is large, which means the effect of measurement error is dominate, we should pay more attention on improving measurement technique. If the ratio is small, then we could focus on use more advanced regression techniques or find more variables or collect more samples to reduce other errors causing the decline of predictive accuracy.

2.7 conclusion and future direction

This chapter focuses on the statistical assessment of predictive performance due to covariate measurement error. Overall, the predictive performance is negatively affected by the increase of magnitude of measurement error. The effect is also influenced by other attributes of data generating model related to the true regression coefficient β, the covariance matrix of true covariates X, Σ_x, and the covariance matrix of the measurement error, Σ_η. From the theoretical derivation of predictive accuracy AUC,

we find that there are four factors might influence the decline of predictive accuracy $\widetilde{AUC}$ when controlling $AUC^{\star}$ and similar findings is shown for the linear case. Then in the simulation study, we find that $E(S_1)$ has a negative relationship with the decline of $\widetilde{AUC}$, while $\mathrm{Var}(S_2)$, $\mathrm{Var}(S_1)$, $\mathrm{Cov}(S_1, X_1 - X_2)$ have positive relationships with the decline of $\widetilde{AUC}$. From these four factors, we find two independent factors, $\mathrm{Var}(S_2)$, $\mathrm{Var}(S_1)$ could be representative of these four factors and both of them have linearly positive relationship with the decline of AUC.

To apply this to practical use, we propose a SIMEX procedure to estimate these two factors from real data, though the estimate is not very accurate. Then we define a ratio of the decline of predictive accuracy due to measurement error compare to the overall decline of predictive accuracy. If the ratio is large, the effect of measurement error dominate the decline of predictive accuracy, otherwise, we do not need to worry much about the measurement error. This could help researchers to decide whether to improve technologies to measure the data more accurately or to use more advanced regression techniques, find more relevant covariates or collect more samples to reduce other errors causing the decline of predictive accuracy.

CHAPTER III

Common and unique associations in screening analyses with multiple subpopulations

3.1 Introduction

Many genomic studies involve the analysis of large numbers of association parameters. One such example would be a biomarker screening study in which a large number of candidate markers are assessed for potential use as predictors of an outcome of interest. The association parameter may be calculated between a single outcome and each of thousands of potential molecular markers. Such studies often involve populations that can be subdivided into several distinct subpopulations. Then it is of interest to ask whether the marker/outcome associations are similar or different among the subpopulations and to estimate the proportion of markers having large effect in both subpopulations.

To set notations, let X_{ij}, where $i = 1, \cdots, m$ denote a set of markers, and let $j = 1, \cdots, n$ denote independent research subjects. The X_{ij} may represent gene expression, genotype, DNA copy number, protein expression, DNA methylation, or any of a number of other molecular assays. The molecular marker data are typically then compared to a phenotype or outcome Y_j $(j = 1, \cdots, n)$ to identify markers that may be used to predict the outcome, or that may mechanistically influence

the outcome. For univariate analysis, common association statistics are the Pearson correlation coefficient $\hat{\rho}_i$ calculated between the i^{th} marker and the outcomes, or the standardized two-group difference of mean marker levels (if the Y_j indicate group membership). To be concrete, we will use correlation coefficients in our presentation here, but our results would apply to many other statistics.

The ultimate aim underlying most "screening studies" is to identify the largest effects, and attribute them to specific markers or sets of markers. However screening studies tend to be modestly powered, and strict control for multiple testing may result in most of the dataset "uninteresting". For example, the familywise error rate (FWER) procedures (such as the Bonferroni correction) controls the probability of making even one false positives in the multiple testings at level α. Then only a small list of markers having the strongest effects are identified though the probability of false positives is really low. An alternative approach, False Discovery Rate controlling procedures are designed to control the expected proportion of false positives in a set of findings (i.e. markers for which the null hypothesis could be rejected). Then a larger list of markers are identified than the the familywise error rate procedures at the cost of a given proportion of the markers in the list are false positives. If we increase the number of markers in the list to be identified as interesting, then the proportion of the markers in the list to be false positives is also increasing. Finally if we choose the whole set of markers, then the proportion of the markers to be false positives equals one minus the proportion of markers having large effects (true positive). Therefore estimating the proportion of markers having large effects in one or two subpopulations without attributing them to specific markers is our main interest. In this paper, we use correlation coefficients ρ_i as the effect sizes, we will estimate the distribution of the effect sizes F and the proportion of the magnitude of the effect sizes ρ_i greater than some threshold t.

An effect size is a measure of the strength of the relationship between two vari-

ables in a statistical population, or a sample-based estimate of that quantity. it is commonly used in genome-wide association studies, and there are different types of effect sizes used by researchers, like Pearson r correlation, effect sizes based on means (Cohen's d, Glass's Δ) and odds ratios (*Hedges and Olkin* (1985)). Effect sizes often refer to a statistic calculated from a sample of data, and are usually estimated with error and may be biased. If many researchers are carrying out studies under low statistical power, the reported effect sizes are biased to be stronger than the true effects (*A et al.* (2008)). Many researchers report the estimates of effect sizes in genome-wide association studies and the empirical distributions of the significant effect sizes (*Nakagawa and Cuthill* (2007), *Park et al.* (2011)), but fewer of them focus on reducing the bias of the estimate and the true effect sizes and the distribution of the true effect sizes and even fewer of them focus on the overlap of the true effect sizes in multiple subpopulations.

Here we focus on a situation that commonly arises in practice, where the units of analysis are not homogeneous, and are structured into groups derived from different subpopulations. Specifically, we have a sample of subjects with chronic kidney disease, each subjects has one of nine underlying diseases that resulted in the kidney disease. In this situation, it is often of interest to consider the fraction of markers having large effect sizes in any given subpopulation and the fraction of markers having large effect sizes in specific pairs of two subpopulations. That is, we ask whether the number and strength of marker/outcome relationships are similar in the different subpopulatons, and whether the strongest predictors are common or unique across different subpopulations.

This task is made more challenging by the fact that statistical power is uneven among the subpopulatons. Thus, even if the number of relationships of a given effect size in two subpopulations are similar under FWER or FDR procedures, the better powered subpopulation will show a greater number of associations. However effect

size ρ_i is invariant to statistical power, then it is used in this chapter to represent the marker/outcome relationship.

3.2 Measures of strength and overlap of effects

Here we view the effect sizes $\rho_i = \mathrm{Cor}(Y, X_i)$, the Pearson correlation coefficients between the outcome and i^{th} marker, to be a random variable having a univariate distribution function F. Based on F, we can define a measure for a given threshold $t > 0$ as the fraction of markers with effect size magnitude (e.g. true correlation coefficient) equal to or exceeding t in one population:

$$N_1(t) = P(|\rho| > t) = 1 - F(t) + F(-t).$$

This measure $N_1(t)$ represents the strength of marker/outcome relationships, as larger values of $N_1(t)$ represent stronger relationships.

If we have two subpopulations A and B, let ρ_i^A and ρ_i^B denote the population associations for the i^{th} marker and the outcome in subpopulations A and B. Then the set of paired values (ρ_i^A, ρ_i^B) can be described with a bivariate distribution function F_{AB}. Based on F_{AB}, we can define an overlap measure for a given threshold $t > 0$ as the fraction of markers with effect size magnitude equal to or exceeding t in both subpopulations:

$$\begin{aligned} N_2(t) &= P(|\rho^A| > t, |\rho^B| > t) \\ &= F_{AB}(-t, -t) + 1 - (F_{AB}(t, \infty) + F_{AB}(\infty, t) - F_{AB}(t, t)). \end{aligned} \tag{3.1}$$

Similar with $N_1(t)$, $N_2(t)$ represents the strength of marker/outcome relationships in both subpopulations.

Since we cannot observe the true effect sizes ρ_i, we work with the observed ef-

fect sizes $\hat{\rho}_i = \widehat{\text{Cor}}(Y, X_i)$, where $X_i = (X_{i1}, \cdots, X_{in})$. There is a estimation error ϵ between ρ and $\hat{\rho}$, where ϵ approximately has a normal distribution with standard deviation $\frac{1}{n}$. This is motivated by the fact that over a large class of data generating distributions for the underlying independent paired data, $\sqrt{n}\hat{\rho}$ is asymptotically standard normal. Here is the sampling error model:

$$\hat{\rho}_i = \rho_i + \epsilon_i, \tag{3.2}$$

where $\epsilon_i \sim N(0, \frac{1}{n})$.

Usually we will transform the correlation coefficient ρ to a variance stablized standardized statistic Z , by using the Fisher transformation

$$Z_i = \frac{\sqrt{n-3}}{2} \log \frac{1+\hat{\rho}_i}{1-\hat{\rho}_i},$$

and vary approximately with a normal distribution around their cental value. We can write

$$Z_i = \theta_i + \eta_i,$$

where η_i is approximately normal and

$$\theta_i \approx \frac{\sqrt{n-3}}{2} \log \frac{1+\rho_i}{1-\rho_i},$$

is an approximate relationship between θ_i and our true effect sizes ρ_i. Now we focus on the standardized parameter θ_i and the standardized statistic Z_i instead of the true effect size ρ_i and the estimated effect size $\hat{\rho}_i$, since θ_i and Z_i are standardized and have invariant variances with different n_i, which is the number of subjects in subgroup i. Then it is more convenient to estimate the distribution of the standardized parameter θ_i than the distribution of the true effect sizes ρ_i. Also it is sufficient to measure $R_1(t) = P(|\theta| > t)$ instead of $N_1(\tilde{t}) = P(|\rho| > \tilde{t})$, where $t = \frac{\sqrt{n-3}}{2} \log \frac{1+\tilde{t}}{1-\tilde{t}}$.

For two subpopulations A and B, (ρ^A, ρ^B) are the population correlation coefficients between markers and outcome in subpopulaton A and B. The variance stablized standardized fisher transformed statistics

$$Z_i^A = \frac{\sqrt{n_A - 3}}{2} \log \frac{1 + \hat{\rho}_i^A}{1 - \hat{\rho}_i^A},$$
$$Z_i^B = \frac{\sqrt{n_B - 3}}{2} \log \frac{1 + \hat{\rho}_i^B}{1 - \hat{\rho}_i^B}.$$

We can write

$$Z_i^A = \theta_i^A + \eta_i^A, \quad Z_i^B = \theta_i^B + \eta_i^B,$$

where (η_i^A, η_i^B) follow independent standard normal distributions. Then

$$\theta_i^A \approx \frac{\sqrt{n_A - 3}}{2} \log \frac{1 + \rho_i^A}{1 - \rho_i^A},$$
$$\theta_i^B \approx \frac{\sqrt{n_B - 3}}{2} \log \frac{1 + \rho_i^B}{1 - \rho_i^B},$$

are the approximate relationships between θ_i^A and ρ_i^A, θ_i^B and ρ_i^B. (θ_i^A, θ_i^B) are the variance stablized standardized parameters in two subpopulations A and B, and it is more convenient to estimate the bivariate distribution of the pairs of parameters (θ_i^A, θ_i^B). Then it is sufficient to estimate $R_2(t_1, t_2) = P(|\theta^A| > t_1, |\theta^B| > t_2)$ instead of $N_2(t) = P(|\rho^A| > t, |\rho^B| > t)$, where $t_1 = \frac{\sqrt{n_A - 3}}{2} \log \frac{1+t}{1-t}$, $t_2 = \frac{\sqrt{n_B - 3}}{2} \log \frac{1+t}{1-t}$.

3.2.1 Plug-in estimation of effect size summaries

The most common and direct way to estimate $R_1(t)$ is just to use the standardized statistic Z_i. We can get the empirical distribution function $\hat{F}$:

$$\hat{F}(t) = \sum_i I(Z_i < t)/m$$

to substitute F when plugging into $R_1(t)$. This is $\hat{R}_1(t)$ called the plug-in estimate of $R_1(t)$, which can be substantially biased.

From the sampling error model, we know the distribution of Z_i is the convolution of the distribution of θ_i and a standard normal distribution. The bias of $\hat{R}_1(t)$ is easily seen to be related to the shape of F. The bias is small when F is diffuse. For example, if F is a uniform distribution, there is no bias since $f \star \phi \approx f$ when $f \sim \text{Unif}(-a, a)$ for large a. The bias is large when F is concentrated near zero. For example, if f is a point mass at zero, then $f \star \phi$ is a normal distribution whose tail probabilities differ strongly from those of f.

For two subpopulations A and B, the plug-in estimator $\hat{R}_2(t_1, t_2)$ is calculated by plugging in the empirical distribution of the pairs of standardized statistics (Z_A, Z_B),

$$\widehat{F_{AB}}(t_1, t_2) = \sum_{ij} I(Z_i^A < t_1 \ \& \ Z_i^B < t_2)/m$$

to $R_2(t_1, t_2)$. The direction of bias in $\hat{R}_2(t_1, t_2)$ is difficult to anticipate. The standardized statistics (Z_A, Z_B) are more dispersed than their true standardized parameters (θ_A, θ_B). This will bias $R_2(t_1, t_2)$ upward. But (Z_A, Z_B) will be less dependent than (θ_A, θ_B). This will bias $R_2(t_1, t_2)$ downward.

3.2.2 Illustration of bias in plug-in estimates

We present some examples to highlight how the bias in the plug-in estimate of $R_2(t_1, t_2)$ occur. Figure 3.1a depicts one extreme situation, where the true (ρ^A, ρ^B) values cluster just below the threshold value t, as depicted in the darker grey color. The lighter grey color depicts the observed distribution of $(\hat{\rho}^A, \hat{\rho}^B)$ values. In this extreme case, all the true parameters fall just outside the region of interest (the set of points x,y such that $\min(x, y) \geq t$), while up to half of these estimated points are extended to fall inside the region of interest. Figure 3.1b shows a contrasting extreme

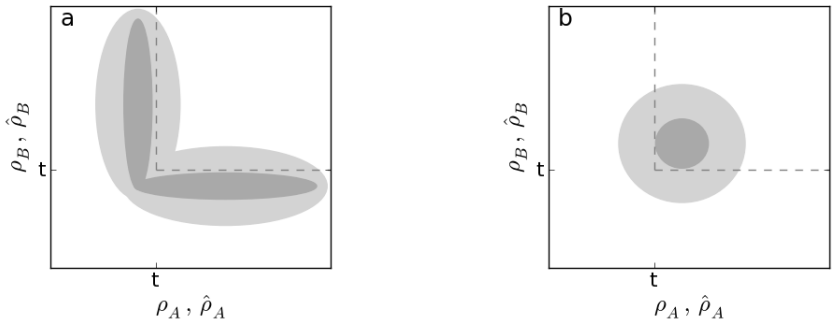

Figure 3.1: Schematic example showing positive bias (a) and negative bias (b) between the observed and true proportions of association statistics in a region of interest. The distribution of true statistics is shown in the darker color, and the distribution of observed statistics is shown in the lighter color. The region of interest is $(x, y : min(x, y) \geq t)$.

situation, where all the true parameters lie inside the region of interest, but up to three quarters of the estimated parameters are expected to lie outside it.

Figure 3.2 and 3.3 are two realistic examples from the CKD data described in section 1.2 of the thesis. The standardized statistic Z_i is the fisher transformation of the sample correlation coefficient $\hat{\rho}_i$ between each marker and the outcome GFR for different disease subgroups. The right plots of figure 3.2 and 3.3 are the scatterplots of the standardized statistics (Z_i^A, Z_i^B) for disease subgroups (MCD, LD) and (IgA, Pima). Though we do not know the true standardized parameter θ_i for disease subgroups, we could estimate them using the methods I will discuss later. Assume the estimate of the true standardized parameter $\hat{\theta}_i$ is known, the left plots of figure 3.2 and 3.3 are the scatterplots of the estimate of the true standardized parameters $(\hat{\theta}_i^A, \hat{\theta}_i^B)$ for disease subgroups (MCD, LD) and (IgA, Pima).

We could estimate $R_2(t_1, t_2)$ using the standardize statistics Z_i^A, Z_i^B by

$$\hat{R}_2(t_1, t_2) = \sum_{i=1}^{m} I_{\{|Z_i^A|>t_1 \ \& \ |Z_i^B|>t_2\}}/m.$$

Also we could estimate $R_2(t_1, t_2)$ using the estimate of the true standardized param-

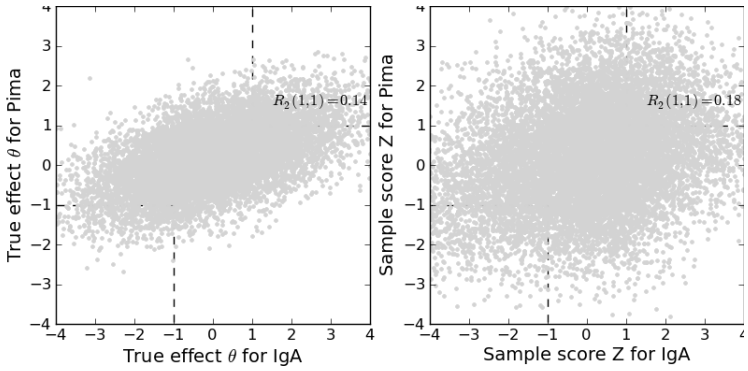

Figure 3.2: Left is the scatterplot of the estimated standardized parameters θ_i^A, θ_i^B for disease subgroups MCD, LD; right is the scatterplot of the standardized statistics Z_i^A, Z_i^B for disease subgroups MCD, LD.

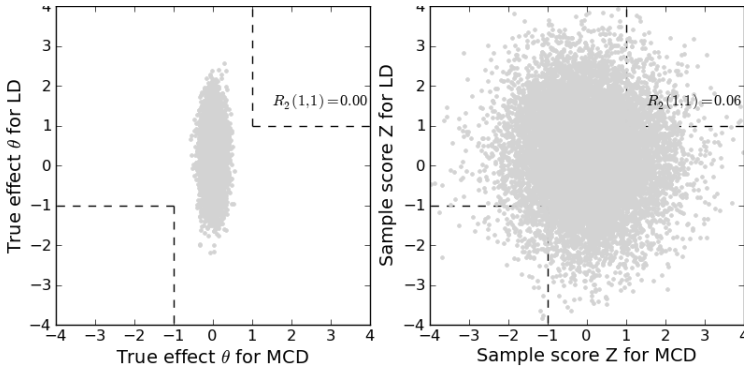

Figure 3.3: Left is the scatterplot of the estimated standardized parameters θ_i^A, θ_i^B for disease subgroups IgA, Pima; right is the scatterplot of the standardized statistics Z_i^A, Z_i^B for disease subgroups IgA, Pima.

eters $\hat{\theta}_i^A$, $\hat{\theta}_i^B$ by

$$\tilde{R}_2(t_1, t_2) = \sum_{i=1}^{m} I_{\{|\hat{\theta}_i^A|>t_1 \ \& \ |\hat{\theta}_i^B|>t_2\}}/m.$$

Figure 3.2 shows $\tilde{R}_2(1,1) = 0.15$ and $\hat{R}_2(1,1) = 0.17$. There is not much difference between $\tilde{R}_2(1,1)$ and $\hat{R}_2(1,1)$ in this situation, because the estimate of the true standardized parameters (θ_i^A, θ_i^B) in disease subgroups IgA and Pima are highly correlated with $r = 0.7$, then the standardized statistics (Z_i^A, Z_i^B) are less dependent than (θ_i^A, θ_i^B) which will bias the measure down and at the same time (Z_i^A, Z_i^B) are more dispersed than (θ_i^A, θ_i^B) which will bias the measure up.

Figure 3.3 shows $\tilde{R}_2(1,1) = 0.0$ and $\hat{R}_2(1,1) = 0.06$. It appears that there is little

correlation of the estimate of the true standardized parameters (θ_i^A, θ_i^B) between MCD and LD disease subgroups, since LD is the control group of people who do not have kidney disease which is irrelevant to other disease subgroups in CKD data. The measure of $R_2(1,1)$ will always bias upwards since (Z_i^A, Z_i^B) will be more dispersed than (θ_i^A, θ_i^B) but the dependencies will remain the same.

3.3 Approaches to bias reduction of the estimation of the effect size summaries

In order to estimate $R_1(t)$ and $R_2(t_1, t_2)$ without much bias, we need to first estimate univariate distribution function F (or F_{AB} for two subgroups) by considering the effects of the sampling error η_i on standardized sample statistics Z_i. Estimating F based on noisy observations with known error distribution is the well-studied "density deconvolution" problem.

We have two general ways to estimate the distribution of the true standardized parameters (θ_i^A, θ_i^B). One is parametric way if we know the statistical model of the distribution, like normal distribution, t distribution and here we will also introduce two ways to estimate the unknown parameters of the distribution, which are moment estimates and maximum likelihood estimates. The other way to estimate the distribution is nonparametric way if we have no idea about how the distribution looks like. Here we also introduce two ways, rescaling method and copula method. In the following paper, we will discuss the advantages and disadvantages of all these methods provided with different distributions of the true standardized parameters θ both in univariate and multivariate cases and simulation results are given. Then we will apply these methods to real data set (CKD).

3.3.1 Parametric approaches

3.3.1.1 Moment estimate

If the framework of the distribution F is known with parameter α_j, $j = 1, \cdots, k$, k is the number of parameters need to be estimated. Then the first k moments of θ would be:

$$\begin{aligned}
\mu_1 &= E[\theta^1] = g_1(\alpha_1, \cdots, \alpha_k), \\
\mu_2 &= E[\theta^2] = g_2(\alpha_1, \cdots, \alpha_k), \\
&\vdots \\
\mu_k &= E[\theta^k] = g_k(\alpha_1, \cdots, \alpha_k),
\end{aligned}$$

where $g_1, \cdots, g_k$ is known for the known distribution F. Let $\hat{\mu_j} = (\sum_{i=1}^{n} \theta_i^j)/m$ be the j^{th} sample moment corresponding to the population moment μ_j, the method of moments estimator for $\alpha_1, \cdots, \alpha_k$ denoted by $\hat{\alpha_1}, \cdots, \hat{\alpha_k}$ is defined by the solution to the equations:

$$\begin{aligned}
\hat{\mu_1} &= g_1(\hat{\alpha_1}, \cdots, \hat{\alpha_k}), \\
\hat{\mu_2} &= g_2(\hat{\alpha_1}, \cdots, \hat{\alpha_k}), \\
&\vdots \\
\hat{\mu_k} &= g_k(\hat{\alpha_1}, \cdots, \hat{\alpha_k}).
\end{aligned} \tag{3.3}$$

Since $g_1, \cdots, g_k$ is known, to get the moment estimates $\hat{\alpha_1}, \cdots, \hat{\alpha_k}$ of $\alpha_1, \cdots, \alpha_k$, we need to estimate the sample moments $\hat{\mu_1}, \cdots, \hat{\mu_k}$. From the sampling error model, we know $\theta_i = Z_i - \eta_i$ and $\bar{\eta_i} = 0, \widehat{Var}(\eta_i) = 1$, η_i is independent with θ_i. Then the sample moments $\hat{\mu_1}, \cdots, \hat{\mu_k}$ could be easily calculated from the sample standardized

statistic $\{Z_i\}$ by

$$\begin{aligned}
\hat{\mu_1} &= \textstyle\sum_{i=1}^{m} \theta_i/m = \sum_{i=1}^{m} Z_i/m, \\
\hat{\mu_2} &= \textstyle\sum_{i=1}^{m} \theta_i^2/m = \sum_{i=1}^{m} Z_i^2/m - 1, \\
\hat{\mu_3} &= \textstyle\sum_{i=1}^{m} \theta_i^3/m = \sum_{i=1}^{m} Z_i^3/m - 3\sum_{i=1}^{m} Z_i/m, \\
&\cdots .
\end{aligned} \tag{3.4}$$

Then by plugging into the value of $\hat{\mu_1}, \cdots, \hat{\mu_k}$ to equations 3.3, $\hat{\alpha_1}, \cdots, \hat{\alpha_k}$ could be solved.

Here we will introduce generalized normal distribution as an example of the distribution F. Generalized normal distribution is a family of continuous probability distribution in which the shape parameter can be used to introduce skew. When the shape parameter is zero, the normal distribution results. Positive values of the shape paramter yield left-skewed distribution bounded to the right, and negative values of the shape parameter yield right-skewed distributions bounded to the left. Its probability density function is

$$f(x) = \frac{\phi(y)}{\alpha - \kappa(x - \xi)},$$

where

$$y = -\frac{1}{\kappa}\log[1 - \frac{\kappa(x-\xi)}{\alpha}] \text{ if } \kappa \neq 0.$$

ϕ is the standard normal pdf, $x \in (-\infty, \xi + \alpha/\kappa)$ if $\kappa > 0$; $x \in (-\infty, \infty)$ if $\kappa = 0$; $x \in (\xi + \alpha/\kappa, \infty)$ if $\kappa < 0$. The functions $g_1, \cdots, g_3$ are

$$g_1(\alpha, \kappa, \xi) = \xi - \frac{\alpha}{\kappa}(e^{\kappa^2/2} - 1),$$

$$g_2(\alpha, \kappa, \xi) = \frac{\alpha^2}{\kappa^2} e^{\kappa^2}(e^{\kappa^2} - 1) + g_1^2,$$

$$g_3(\alpha, \kappa, \xi) = \frac{\alpha^3 e^{\kappa^3}(3e^{\kappa^2} - e^{3\kappa^2} - 2)}{\kappa^3}\text{sign}(\kappa) + 3g_1g_2 - 2g_1^3.$$

Then if we could calculate the values of $\hat{\mu_1}, \cdots, \hat{\mu_3}$ using the sample standardized statistics Z_i by equations 3.4 and use the functions $g_1, \cdots, g_3$ above, $\hat{\alpha}, \hat{\kappa}, \hat{\xi}$ could be estimated using equations 3.3.

3.3.1.2 Maximum Likelihood estimate

As we all know maximum-likelihood estimation (MLE) is a method of estimating the parameters of a given statistical model. In general, for a fixed set of data and underlying statistical model, the method of maximum likelihood selects values of the model parameters that produce a distribution that gives the observed data the greatest probability (i.e. parameters that maximum the likelihood function). Here we have observed statistic $\{Z_i\}$, and if we know the density function of true parameter $\{\theta_i\}$ is $f(\theta, \alpha)$, $\alpha = (\alpha_1, \cdots, \alpha_k)$, using sampling error model, we know the density function for $\{Z_i\}$ is $g(Z, \alpha) = f(\theta, \alpha) * \phi$, the convolution of f and a standard normal density. The likelihood function is

$$L(\alpha; Z_1, \cdots, Z_m) = \prod_{i=1}^{m} g(Z_i, \alpha).$$

In normal and many other cases, if the statistical model is known, the method of moments and MLE method would be the most simple and quick way to estimate the parameters of the distribution function F of our true parameters θ and then estimate the measure of the fraction of the markers with effect size magnitude greater than some threshold in one population $R_1(t)$.

For two subpopulations A and B, if the bivariate distribution F_{AB} of the true parameters θ_A, θ_B is known with parameter (α, β, r), $\alpha = (\alpha_1, \cdots, \alpha_k)$, $\beta = (\beta_1, \cdots, \beta_k)$. We could still use moments method and MLE method to estimate the parameters of the bivariate distribution F_{AB} and then estimate the overlap measure of the fraction

of the markers with effect size magnitude both greater than some threshold t in two subpopulations $R_2(t_1, t_2)$. Here we use bivariate normal distribution as an example of F_{AB} in two subpopulations A, B. Then the probability density function for true standardized parameters (θ_i^A, θ_i^B) is:

$$f_{AB}^i(\theta_i^A, \theta_i^B; \mu_A, \mu_B, \sigma_A, \sigma_B, r)$$
$$= \frac{1}{2\pi\sigma_A\sigma_B\sqrt{1-r^2}} \exp(-\frac{1}{2(1-r^2)}[\frac{(\theta_i^A-\mu_A)^2}{\sigma_A^2} + \frac{(\theta_i^B-\mu_B)^2}{\sigma_B^2} - \frac{2r(\theta_i^A-\mu_A)(\theta_i^B-\mu_B)}{\sigma_A\sigma_B}]), \quad (3.5)$$

where r is the correlation between θ_A and θ_B and $\sigma_A > 0$, $\sigma_B > 0$. Then the probability density function for the sample standardized statistics (Z_i^A, Z_i^B) is:

$$f_{AB}^i(Z_i^A, Z_i^B; \mu_A, \mu_B, \sigma_A, \sigma_B, r)$$
$$= \frac{1}{2\pi\tilde{\sigma}_A\tilde{\sigma}_B\sqrt{1-r^2}} exp(-\frac{1}{2(1-r^2)}[\frac{(Z_i^A-\mu_A)^2}{\tilde{\sigma}_A^2} + \frac{(Z_i^B-\mu_B)^2}{\tilde{\sigma}_B^2} - \frac{2r(Z_i^A-\mu_A)(Z_i^B-\mu_B)}{\tilde{\sigma}_A\tilde{\sigma}_B}]). \quad (3.6)$$

where $\tilde{\sigma_A} = \sqrt{\sigma_A^2 + 1}$, $\tilde{\sigma_B} = \sqrt{\sigma_B^2 + 1}$. Then the likelihood function for the observed standardized statistics (Z_A, Z_B) is

$$L(Z_A, Z_B; \mu_A, \mu_B, \sigma_A, \sigma_B, r) = \prod_{i=1}^{m} f_{AB}^i(Z_i^A, Z_i^B; \mu_A, \mu_B, \sigma_A, \sigma_B, r). \quad (3.7)$$

The parameters μ_A, μ_B, σ_A, σ_B and r of the bivariate normal distribution F_{AB} could be estimated by maximizing the likelihood function $L(Z_A, Z_B; \mu_A, \mu_B, \sigma_A, \sigma_B, r)$.

For moment estimators of μ_A, μ_B, σ_A, σ_B and r of the bivariate normal distribution F_{AB} based on the observed standardized statistics (Z_i^A, Z_i^B). First we know the relationships between the parameters μ_A, μ_B, σ_A, σ_B, r and the first and second

moments of F_{AB}:

$$\begin{aligned}
\mu_A = E[\theta_A], \quad \mu_B = E[\theta_B], \\
\mu_A^2 + \sigma_A^2 = E[\theta_A^2], \quad \mu_B^2 + \sigma_B^2 = E[\theta_B^2], \\
r = cor(\theta_A, \theta_B) = cov(\theta_A, \theta_B)/(sd(\theta_A) * sd(\theta_B)).
\end{aligned} \tag{3.8}$$

Second we could estimate the sample moments of F_{AB} using the observed standardized statistics Z_i^A, Z_i^B. Since we have the equations $\eta_i^A = Z_i^A - \theta_i^A$, $\eta_i^B = Z_i^B - \theta_i^B$, $i = 1, \cdots, m$. As noted above, basic asymptotic theory suggests treating $\eta^A|\theta_A$ as following a standard normal distribution. If we view θ_A as random, η_A is unconditionally standard normal. We also assume that η_A and θ_A are independent. A parallel set of statements holds for η_B and θ_B. Under these assumptions, we get the identity that

$$\begin{aligned}
\bar{\eta}_A = \textstyle\sum_{i=1}^m \eta_i^A/m = 0, \quad \bar{\eta}_B = \textstyle\sum_{i=1}^m \eta_i^B/m = 0. \\
\mathrm{var}(\eta_A) = \textstyle\sum_{i=1}^m (\eta_i^A)^2/m = 1, \quad \mathrm{var}(\eta_B) = \textstyle\sum_{i=1}^m (\eta_i^B)^2/m = 1. \\
\mathrm{Cov}(\theta_A, \eta_A) = \textstyle\sum_{i=1}^m \theta_i^A \eta_i^A/m = 0, \quad \mathrm{Cov}(\theta_B, \eta_B) \textstyle\sum_{i=1}^m \theta_i^B \eta_i^B/m = 0.
\end{aligned} \tag{3.9}$$

Using these identities, we have

$$\begin{aligned}
\textstyle\sum_{i=1}^m \theta_i^A/m = \textstyle\sum_{i=1}^m Z_i^A/m - \textstyle\sum_{i=1}^m \eta_i^A/m = \textstyle\sum_{i=1}^m Z_i^A/m, \\
\textstyle\sum_{i=1}^m \theta_i^B/m = \textstyle\sum_{i=1}^m Z_i^B/m - \textstyle\sum_{i=1}^m \eta_i^B/m = \textstyle\sum_{i=1}^m Z_i^B/m, \\
\textstyle\sum_{i=1}^m (Z_i^A)^2/m = \textstyle\sum_{i=1}^m (\theta_i^A)^2/m + 2\textstyle\sum_{i=1}^m \theta_i^A \eta_i^A/m + \textstyle\sum_{i=1}^m (\eta_i^A)^2/m = \textstyle\sum_{i=1}^m (\theta_i^A)^2/m + 1. \\
\textstyle\sum_{i=1}^m (Z_i^B)^2/m = \textstyle\sum_{i=1}^m (\theta_i^B)^2/m + 2\textstyle\sum_{i=1}^m \theta_i^B \eta_i^B/m + \textstyle\sum_{i=1}^m (\eta_i^B)^2/m = \textstyle\sum_{i=1}^m (\theta_i^B)^2/m + 1. \\
\hat{Cov}(Z_A, Z_B) = \hat{Cov}(\theta_A + \eta_A, \theta_B + \eta_B) = \hat{Cov}(\theta_A, \theta_B).
\end{aligned} \tag{3.10}$$

Then we could estimate the parameters μ_A, μ_B, σ_A, σ_B and r by solving the following

equations:

$$\hat{\mu_A} = \sum_{i=1}^{m} \theta_i^A/m = \sum_{i=1}^{m} Z_i^A/m,$$
$$\hat{\mu_B} = \sum_{i=1}^{m} \theta_i^B/m = \sum_{i=1}^{m} Z_i^B/m,$$
$$\hat{\mu_A^2} + \hat{\sigma}_A^2 = \sum_{i=1}^{m} (\theta_i^A)^2/m = \sum_{i=1}^{m} (Z_i^A)^2/m - 1,$$
$$\hat{\mu_B^2} + \hat{\sigma}_B^2 = \sum_{i=1}^{m} (\theta_i^B)^2/m = \sum_{i=1}^{m} (Z_i^B)^2/m - 1,$$
$$\hat{r} = \hat{cor}(\theta_i^A, \theta_i^B) = \hat{Cov}(Z_i^A, Z_i^B)/(\hat{\sigma_A}\hat{\sigma_B}), \tag{3.11}$$

3.3.2 Nonparametric approaches

3.3.2.1 Rescaling estimator

If we do not know the statistical model of the true standardized parameters (θ_A, θ_B), there is a method called "Rescaling method" may help estimate $R_1(t)$ and $R_2(t_1, t_2)$. The basic idea of rescaling method is to produce two sets of points whose sample variance, and sample mean are the same with true parameters (θ_A, θ_B) and the correlation of these two sets of points is the same with the correlation between true parameters (θ_A, θ_B). This will give us two sets of points whose dispersion and degree of association are comparable to the true standardized parameters θ_A and θ_B values. The empirical distribution function of these points, denoted as $\tilde{F}_{AB}$ will be plugged into $R_1(t)$ and $R_2(t_1, t_2)$.

We could calculated the sample variance $\hat{\sigma}_A^2$, $\hat{\sigma}_B^2$, sample mean $\hat{\mu}_A, \hat{\mu}_B$ and sample

correlation $\hat{r}$ from equations 3.11:

$$\hat{\mu_A} = \sum_{i=1}^{m} \theta_i^A/m = \sum_{i=1}^{m} Z_i^A/m,$$
$$\hat{\mu_B} = \sum_{i=1}^{m} \theta_i^B/m = \sum_{i=1}^{m} Z_i^B/m,$$
$$\hat{\mu_A^2} + \hat{\sigma}_A^2 = \sum_{i=1}^{m} (\theta_i^A)^2/m = \sum_{i=1}^{m} (Z_i^A)^2/m - 1,$$
$$\hat{\mu_B^2} + \hat{\sigma}_B^2 = \sum_{i=1}^{m} (\theta_i^B)^2/m = \sum_{i=1}^{m} (Z_i^B)^2/m - 1,$$
$$\hat{r} = \hat{cor}(\theta_i^A, \theta_i^B) = \hat{Cov}(Z_i^A, Z_i^B)/(\hat{\sigma_A}\hat{\sigma_B}),$$

For univariate case, we have the transformed statistics

$$\tilde{Z} = \lambda_1 Z + \lambda_2, \tag{3.12}$$

where λ_1 ranges from -1 to 1 is set to meet the desired variance $\hat{\sigma}^2$ and then choose λ_2 which will not effect the variance of $\tilde{Z}$ to meet the desired mean value $\hat{\mu}$. Once $\tilde{Z}$ is generated, the rescaling estimate of $R_1(t) = P(|\tilde{Z}| > t)$.

For bivariate case, we have the transformed statistics

$$\tilde{Z_A} = \lambda_{A_1}(Z_A + \lambda_r Z_B) + \lambda_{A_2},$$
$$\tilde{Z_B} = \lambda_{B_1}(Z_A + \lambda_r Z_B) + \lambda_{B_2}. \tag{3.13}$$

Note that as λ_r ranges from -1 to 1, the correlation coefficient between $\tilde{Z_A}$ and $\tilde{Z_B}$ ranges from -1 to 1 monotonically. Thus there is always a unique value of λ_r such that the correlation between $\tilde{Z_A}$ and $\tilde{Z_B}$ equals $\hat{r}$. This value can easily be found numerically using bisection computing method. Once this value, $\lambda_{\hat{r}}$, is found, the correlation will not change when λ_{A_1}, λ_{B_1} are set to give the desired variance $\hat{\sigma}_A^2$, $\hat{\sigma}_B^2$ and λ_{A_2}, λ_{B_2} are set to give the desired mean $\hat{\mu_A}, \hat{\mu_B}$. Once $\tilde{Z_A}, \tilde{Z_B}$ are generated, the rescaling estimate of $R_2(t_1, t_2) = P(|\tilde{Z_A}| > t_1, |\tilde{Z_B}| > t_2)$.

We note that this approach is exact for large samples if F_{AB} is approximately

Gaussian. If the exact values of r, σ_A, σ_B are used rather than the estimates, the (Z_i^A, Z_i^B) pairs can be linearly transformed to the exact joint distribution of (θ^A, θ^B). If F_{AB} is not Gaussian, the linearly transformed $(\tilde{Z_i}^A, \tilde{Z_i}^B)$ values will in general not be exactly distributed according to F_{AB}, even if r, σ_A, σ_B are estimated exactly.

3.3.2.2 Copula method

If we specify a parametric statistical model F_{AB} for the joint distribution of the true standardized parameters (θ_i^A, θ_i^B), we could use the MLE or method of moments to estimate the parameters of F_{AB}. As a more general approach, we can follow the idea used in a Gaussian Copula to describe the F_{AB}. Specifically, we model (θ_i^A, θ_i^B) as $(t_A(X_i^A), t_B(X_i^B))$, where (X_i^A, X_i^B) are centered bivariate normal random variables, with $SD(X_i^A) = SD(X_i^B) = 1$, and $\text{cor}(X_i^A, X_i^B) = r$, and t_A, t_B are non-decreasing real-valued functions of a real variable.

To review, the basic idea of a copula is that we consider a random vector $(Y_1, \cdots, Y_d)$. Suppose its marginal CDFs $F_1, \cdots, F_d$ are continuous functions. By applying the probability integral transform to each component, the random vector

$$(U_1, \cdots, U_d) = (F_1(Y_1), \cdots, F_d(Y_d))$$

has uniform margins. The copula of $(Y_1, \cdots, Y_d)$ is defined as the joint cumulative distribution function of $(U_1, \cdots, U_d)$,

$$C(u_1, \cdots, u_d) = P[U_1 \leq u_1, \cdots, U_d \leq u_d].$$

The copula C contains all information on the dependence structure between the components of $(Y_1, Y_2, \cdots, Y_d)$ whereas the marginal cumulative distribution functions F_i contain all information on the marginal distributions. The importance of the above is that the reverse of these steps can be used to generate random samples from gen-

eral classes of multivariate probability distributions. That is, given a procedure to generate a sample $(U_1, U_2, \cdots, U_d)$ from the copula distribution, the required sample can be constructed as

$$(Y_1, Y_2, \cdots, Y_d) = (F_1^{-1}(U_1), F_2^{-1}(U_2), \cdots, F_d^{-1}(U_d)).$$

The inverses F_i^{-1} are unproblematic as the F_i were assumed to be continuous. The above formula for the copula function can be rewritten to correspond to this as:

$$C(u_1, \cdots, u_d) = P[Y_1 < F_1^{-1}(u_1), \cdots, Y_d < F_d^{-1}(u_d)]$$

Here we use the Gaussian copula by projecting a multivariate normal distribution on R^d by means of the probability integral transform to the unit cube $[0, 1]^d$. For a given correlation matrix Σ, the Gaussian copula is

$$C_\Sigma^{Gauss}(u) = \Phi_\Sigma(\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_d)),$$

where Φ^{-1} is the inverse cumulative distribution function of a standard normal distribution and Φ_Σ is the joint cumulative distribution function of a multivariate normal distribution with mean vector zero and covariance matrix equal to the correlation matrix Σ. Let

$$X_1 = \Phi^{-1}(u_1) = \Phi^{-1}F(Y_1), \cdots, X_d = \Phi^{-1}(u_d) = \Phi^{-1}F(Y_d), \tag{3.14}$$

then $X_1, \cdots, X_d$ follows a joint multivariate normal distribution with mean vector zero and covariance matrix equal to the correlation matrix Σ.

In this paper, we focus on estimating the joint distribution of the true standardized parameters (θ_i^A, θ_i^B) for two subgroups A, B. Let

$$X_i^A = \Phi^{-1} F_A(\theta_i^A), \quad X_i^B = \Phi^{-1} F_B(\theta_i^B),$$

then (X_i^A, X_i^B) follows a bivariate normal distribution with $E(X_i^A) = E(X_i^B) = 0$, $SD(X_i^A) = SD(X_i^B) = 1$, and correlation matrix $\Sigma = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$, $-1 < r < 1$, having density function $f_\Sigma(X_i^A, X_i^B)$.

Also in our analysis, we cannot use the standard copula, because we are seeking to model the joint distribution of (θ_i^A, θ_i^B), which are not observed. Therefore, we extend the basic copula idea as follows. Since we do not observe θ_i^A and θ_i^B, we cannot simply compute their empirical distribution functions F_A, F_B and quantile functions. We therefore model $t_A = \Phi^{-1} F_A$ and $t_B = \Phi^{-1} F_B$ as continuous linear splines and t_A, t_B map θ_i^A to X_i^A, θ_i^B to X_i^B. Then the joint density function for (θ_i^A, θ_i^B),

$$f(\theta_i^A, \theta_i^B) = f_\Sigma(t_A(\theta_i^A), t_B(\theta_i^B)) | \frac{d(t_A(\theta_i^A))}{d\theta_i^A} | | \frac{d(t_B(\theta_i^B))}{d\theta_i^B} |, \tag{3.15}$$

and the joint density function for the observed standardized statistics (Z_i^A, Z_i^B),

$$f(Z_i^A, Z_i^B) = f(\theta_i^A, \theta_i^B) \star \phi, \tag{3.16}$$

where ϕ represent a standard bivariate normal distribution with correlation matrix I, can be easily computed numerically. We then optimize the joint log likelihood function of the standardized sample statistics Z_A, Z_B,

$$L(t_A, t_B, r; Z_A, Z_B) = \sum_{i=1}^{m} \log f(Z_i^A, Z_i^B) = \sum_{i=1}^{m} \log[f(\theta_i^A, \theta_i^B) \star \phi]. \tag{3.17}$$

over t_A, t_B and the correlation parameter r.

our approach for optimizing 3.17 is heuristic, and employs a greedy stochastic optimization. We first define a grid G_r on $[-1, 1]$, and for each $r \in G_r$, we optimize 3.17

over (t_A, t_B). The optimization over (t_A, t_B) are conducted by generating random non-decreasing sequences D, and first setting

$$t_A^{(i+1)} = (1-\lambda)t_A^{(i)} + \lambda D,$$

and fix $t_B^{(i)}$, where λ initially is set at $\lambda = 0.5$, and is successively halved until a higher log likelihood value of 3.17 is reached,

$$L(t_A^{(i+1)}, t_B^{(i)}, r; Z_A, Z_B) > L(t_A^{(i)}, t_B^{(i)}, r; Z_A, Z_B).$$

If no such value is reached, the function of t_A at the $(i+1)^{th}$ step will not change, $t_A^{(i+1)} = t_A^{(i)}$. Second set

$$t_B^{(i+1)} = (1-\lambda)t_B^{(i)} + \lambda D,$$

and fix $t_A^{(i+1)}$, where λ initially is set at $\lambda = 0.5$, and is successively halved until a higher log likelihood value of 3.17 is reached,

$$L(t_A^{(i+1)}, t_B^{(i+1)}, r; Z_A, Z_B) > L(t_A^{(i+1)}, t_B^{(i)}, r; Z_A, Z_B).$$

If no such value is reached, the function of t_B at the $(i+1)^{th}$ step will not change, $t_B^{(i+1)} = t_B^{(i)}$. A random non-decreasing sequences D is generated in each iteration and after k iterations, the final $t_A^{(k)}$ and $t_B^{(k)}$ is estimated from initial $t_A^{(0)} = t_B^{(0)} = I$. Then we compare the log likelihood value of 3.17, $L(t_A^{(k)}, t_B^{(k)}, r; Z_A, Z_B)$ for each correlation parameter $r \in G_r$ and report the value of r that optimize the log likelihood value of 3.17 and their corresponding $t_A^{(k)}$ and $t_B^{(k)}$.

The sequence D is generated by first simulate k i.i.d. values $\{U_{1i}\}$ uniformly on $[-\mathrm{mx}_1, \mathrm{mx}_2]$, where $\mathrm{mx}_1, \mathrm{mx}_2$ are values uniformly distributed on $[-10, 10]$ and k is the number of values in $\{U_{1i}\}$, which is a random integer from 3 to 20. Then simulate k i.i.d. values $\{U_{2i}\}$ uniformly on $[\mathrm{mx}_3, \mathrm{mx}_4]$, where $\mathrm{mx}_3, \mathrm{mx}_4$ are the minimum and

maximum values of observed standardized statistic Z_i and k is a random integer from 3 to 20. Then sequences $\{U_{1i}\}$ and $\{U_{2i}\}$ are sorted from low to high and there is a function F_u mapping $\{U_{2i}\}$ to $\{U_{1i}\}$ and any two adjacent values $(U_{1i}, U_{1,i+1})$ are connected linearly. We construct a grid G on $[\mathrm{mx}_3, \mathrm{mx}_4]$ with 200 knots, then the sequence $D = F_u(G)$.

3.4 Simulation study for univariate analysis

First, we focus on the univariate analysis. We will estimate the marginal distribution of the standardized parameter θ, and then estimate $R_1(t)$ when the true marginal distribution of θ has three situations. We will compare the results of the plug-in estimates, moment estimates, mle estimates, rescaling estimates and copula estimates of $R_1(t)$ with the true value of $R_1(t)$. The following are the simulation steps:

1. Generate true standardized parameter θ_i, $i = 1, \cdots, m$, $m = 10000$ from a given distribution F, F has three situations illustrated above, i): $N(0, 1)$, (ii): t distribution with $df = 3$ (iii): generalized normal distribution with $\xi = -0.5, \alpha = 2, \kappa = -0.5$. Then generate a sequence of theresholds T vary from 0 to 4 by 0.2 and the true value of $R_1(T)$ is calculated by $R_1(t) = \sum_{i=1}^{m} I_{|\theta_i|>t}/m$ for every t in T.

2. The observed standardized statistic Z_i is estimated by adding standard normal errors to θ_i, $Z_i = \theta_i + \eta_i$, η_i follows standard normal distribution. Then the plug-in estimator of $R_1(T)$ is calculated by $R_1(t) = \sum_{i=1}^{m} I_{|Z_i|>t}/m$ for every t in T.

3. If we assume that F is a normal distribution with mean μ and standard deviation σ, no matter what the true distribution F is. The moment estimator of $R_1(T)$ is calculated by estimating the parameters of the distribution F by matching

the first 2 moments of the distribution F to the first 2 sample moments of θ. From equation 3.11, we know that

$$\hat{\mu} = \sum_{i=1}^{m} Z_i/m,$$
$$\hat{\sigma} = \sqrt{\sum_{i=1}^{m} Z_i^2/m - (\sum_{i=1}^{m} Z_i/m)^2 - 1}.$$

Then the moment estimator of $R_1(T)$ is calculated by $R_1(t) = 1 - \Phi(\frac{t-\hat{\mu}}{\hat{\sigma}}) + \Phi(\frac{-t-\hat{\mu}}{\hat{\sigma}})$.

4. The mle estimator of $R_1(T)$ is calculated by estimating the parameters of the distribution F by maximizing the likelihood density function of the observed standardized statistic Z,

$$L(Z; \mu, \sigma) = \prod_{i=1}^{m} f(Z_i; \mu, \sigma),$$
$$f(Z_i; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{Z_i^2}{2*\sigma^2} \star \phi,$$

which is the convolution of the density function of θ and a standard normal density function. Then the mle estimator of $R_1(T)$ is calculated by $R_1(t) = 1 - \Phi(\frac{t-\hat{\mu}}{\hat{\sigma}}) + \Phi(\frac{-t-\hat{\mu}}{\hat{\sigma}})$.

5. The rescaling estimator of $R_1(T)$ is calculated by transforming the observed statistic Z to a new vector X which has the same mean and variance with the true parameter θ by using equation 3.12, then $R_1(T)$ is calculated by $R_1(t) = \sum_{i=1}^{n} I_{|X_i|} > t/n$ for every t in T.

6. The copula estimator of $R_1(T)$ for univariate analysis is just focusing on how to estimate the monotone increasing functions t that maps the true parameters θ to a random variable X which follows a standard normal distribution using the linear spline method we introduced in section 3.2.2. Once $\hat{t}$ is the estimated, we could construct a sample of values X with sample size $n = 10000$ from a

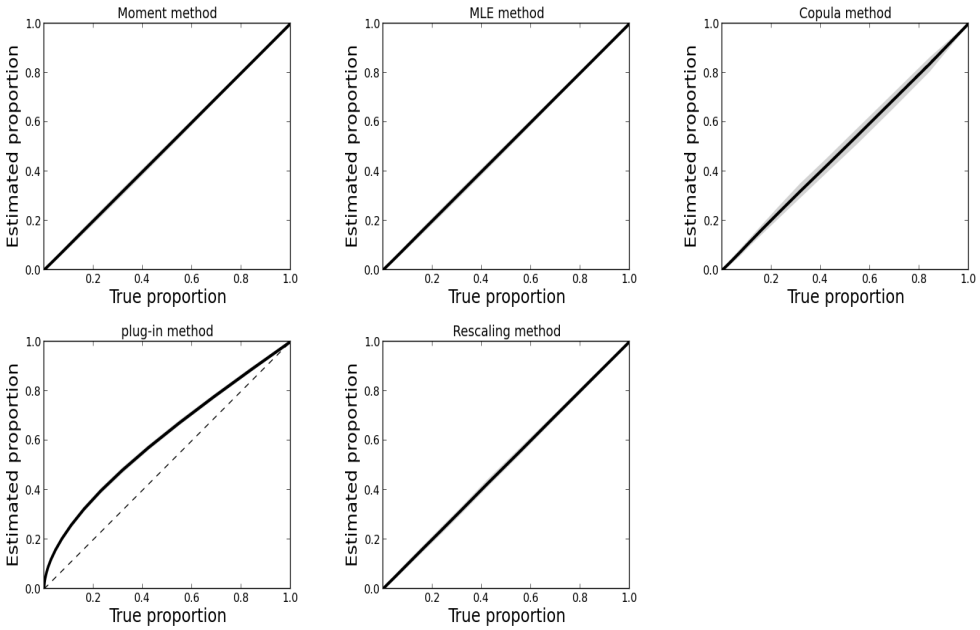

Figure 3.4: Plots of the true $R_1(t)$ and the average of the estimate of $R_1(t)$ for each parametric and nonparametric methods when the true marginal distribution is $N(0,1)$ and the grey area is the approximate 95% confidence intervals for the estimate of $R_1(t)$.

standard normal distribution, and then construct a sample of estimated parameters θ by $\tilde{\theta} = t^{-1}(X)$. Then the copula estimator of $R_1(T)$ is calculated by $R_1(t) = \sum_{i=1}^{m} I_{|\tilde{\theta}_i|>t}/n$ for every t in T.

7. The procedure was repeated 100 times to get the average value and the standard deviation of the estimate of $R_1(T)$ for each method. The plots comparing the true $R_1(t)$ and the average of the estimate of $R_1(T)$ and the approximate 95% confidence intervals for the estimate of $R_1(t)$ for each method is constructed.

Figure 3.4 shows the situation when the true parameters θ follows a standard normal distribution, all the methods except the plug-in method perform the same, all the estimates of the $R_1(T)$ are very close to the true value of $R_1(T)$ and have very small variation. Since the parametric model of the marginal distribution of θ for moments and mle method is the same with the true marginal distribution of θ and the variance of the true parameters θ is fairly large, then the moment and mle

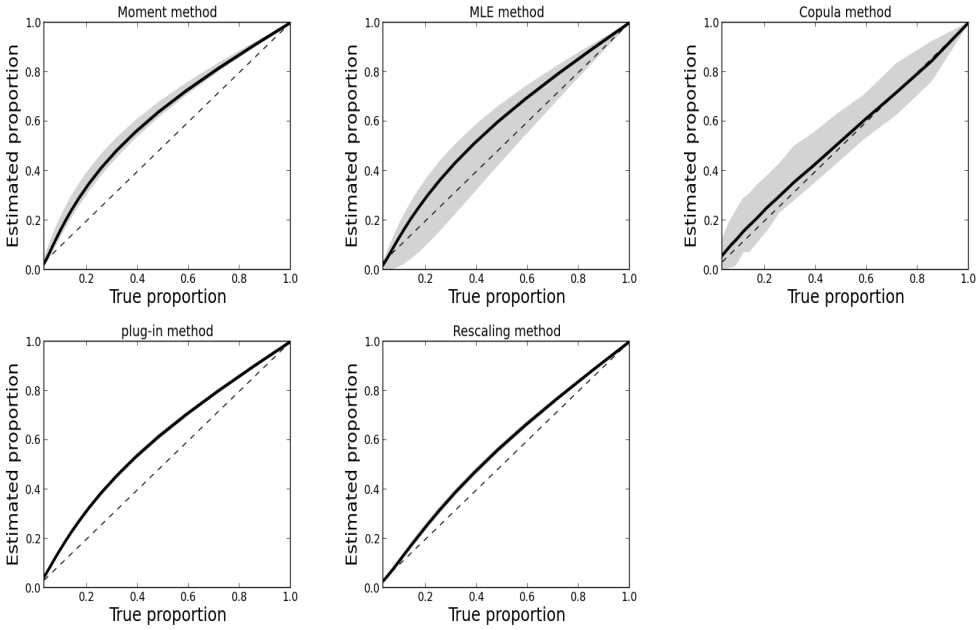

Figure 3.5: Plots of the true $R_1(t)$ and the average of the estimate of $R1(t)$ for each parametric and nonparametric methods when the true marginal distribution is $t(3)$ and the grey area is the approximate 95% confidence intervals for the estimate of $R_1(t)$.

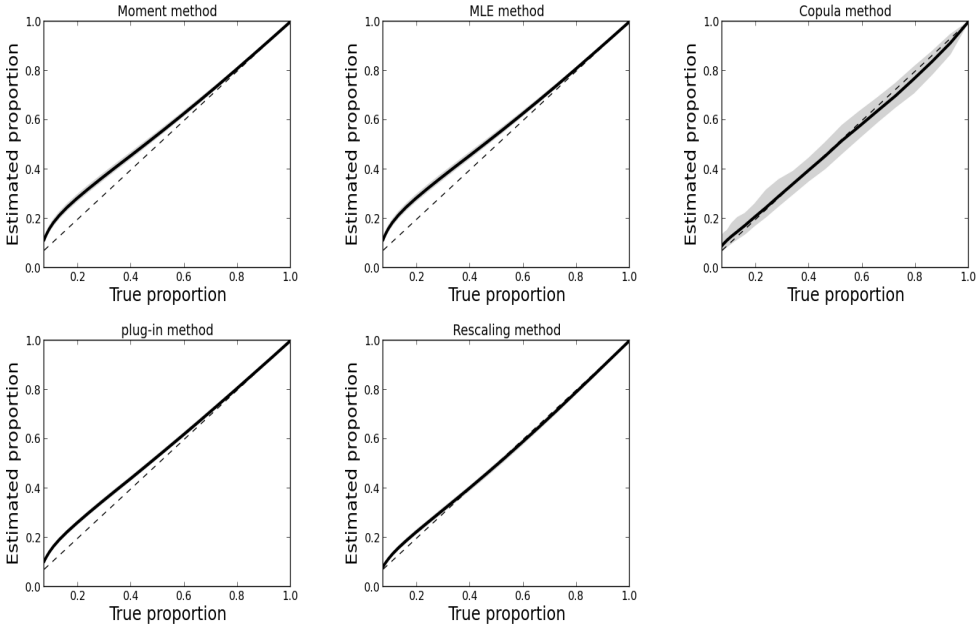

Figure 3.6: Plots of the true $R_1(t)$ and the average of the estimate of $R1(t)$ for each parametric and nonparametric methods when the true marginal distribution is generalized normal distribution with $\xi = -0.5, \alpha = 2, \kappa = -0.5$ and the grey area is the approximate 95% confidence intervals for the estimate of $R_1(t)$.

method perform well. So do the rescaling and copula method.

Figure 3.5 shows the situation when the true parameters θ follows t distribution with $df = 3$, then the parametric estimates of $R_1(T)$ which assume the normal model of the distribution of θ are biased. While the copula method give a more accurate estimate than the other methods but with more variabilities.

Figure 3.6 shows the situation when the true parameters θ follows a generalized normal distribution with $\xi = -0.5, \alpha = 2, \kappa = -0.5$, now the distribution of θ is not symmetric, then the parametric estimates of $R_1(T)$ which assume the normal model of the distribution of θ show much deviation from the true $R_1(T)$, and perform even worse than the plug-in estimates, while the copula method perform the best.

As a conclusion, the copula method for univariate analysis performs the better than the other method if the distributions of the true standardized parameter θ is not normal. Then we will look at whether this conclusion is also true for bivariate analysis.

3.5 Simulation study for bivariate analysis

Now, we focus on the bivariate analysis. We will estimate the bivariate distribution F_{AB} of the standardized parameters (θ_A, θ_B) for two subgroups A, B, and then estimate $R_2(t,t)$ when the true bivariate distribution is known. We will compare the results of the plug-in estimates, moment estimates, mle estimates, rescaling estimates and copula estimates of $R_2(t,t)$ with the true value of $R_2(t,t)$. The following are the simulation steps:

1. Generate scores X_i^A, X_i^B, $i = 1, \cdots, m$, $m = 10000$ from a bivariate normal distribution with mean vectors 0 and correlation matrix $\Sigma = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$, $r = 0$ or

0.5. Generate true standardized parameters (θ_i^A, θ_i^B) by

$$\theta_i^A = t_A^{-1}(X_i^A), \quad \theta_i^B = t_B^{-1}(X_i^B).$$

t_A, t_B has three situations, (i): $t_A(x) = t_B(x) = x$, then (θ_i^A, θ_i^B) follows standard bivariate normal distribution with correlation r. (ii): $t_A(x) = t_B(x) = \Phi^{-1}t_3(x)$, then θ_i^A, θ_i^B follows marginal t distribution with $df = 3$ and when they are transformed back to X_i^A, X_i^B using function t_A, t_B, they have correlation r. (iii): $t_A(x) = t_B(x) = 2\log(1 + (x + 0.5)/4)$, then θ_i^A, θ_i^B follows marginal generalized normal distribution with $\xi = -0.5, \alpha = 2, \kappa = -0.5$ and when they are transformed back to X_i^A, X_i^B using function t_A, t_B, they have correlation r. Then generate a sequence of theresholds T vary from 0 to 4 by 0.2 and the true value of $R_2(T, T)$ is calculated by $R_2(t, t) = \sum_{i=1}^m I_{\{|\theta_i^A|>t \ \& \ |\theta_i^B|>t\}}/m$ for every t in T.

2. The observed standardized statistics (Z_i^A, Z_i^B) are estimated by adding standard normal errors to (θ_i^A, θ_i^B).

$$Z_i^A = \theta_i^A + \eta_i^A, \quad Z_i^B = \theta_i^B + \eta_i^B,$$

where (η_i^A, η_i^B) follows independent standard normal distribution. Then the plug-in estimator of $R_2(T, T)$ is calculated by $R_2(t, t) = \sum_{i=1}^m I_{\{|Z_i^A|>t \ \& \ |Z_i^B|>t\}}/m$ for every t in T.

3. If we assume that the joint distribution F_{AB} of (θ_i^A, θ_i^B) is bivariate normal with parameters $\mu_A, \mu_B, \sigma_A, \sigma_B, r$, the moment estimator of $R_2(T, T)$ is calculated by estimating the parameters of the distribution F_{AB} by matching the first 2 moments of the distribution F_{AB} to the first 2 sample moments of θ_i^A, θ_i^B and

the sample correlation between θ_i^A and θ_i^B. From equation 3.11, we know that

$$\hat{\mu_A} = \sum_{i=1}^m Z_i^A/m,$$
$$\hat{\mu_B} = \sum_{i=1}^m Z_i^B/m,$$
$$\hat{\sigma}_A^2 = \sum_{i=1}^m (Z_i^A)^2/m - (\sum_{i=1}^m Z_i^A/m)^2 - 1,$$
$$\hat{\sigma}_B^2 = \sum_{i=1}^m (Z_i^B)^2/m - (\sum_{i=1}^m Z_i^B/m)^2 - 1,$$
$$\hat{r} = \hat{Cov}(Z_i^A, Z_i^B)/(\hat{\sigma_A}\hat{\sigma_B}).$$

Then we could generate samples $(\tilde{X}_A, \tilde{X}_B)$ from this bivariate normal distribution with parameters $\hat{\mu}_A, \hat{\mu}_B, \hat{\sigma}_A, \hat{\sigma}_B, \hat{r}$. Then the moment estimator of $R_2(T,T)$ is calculated by $R_2(t,t) = \sum_{i=1}^m I_{\{|\tilde{X}_i^A|>t \ \& \ |\tilde{X}_i^B|>t\}}/m$ for every t in T.

4. With the same assumption of step 3, The mle estimator of $R_2(T,T)$ is calculated by estimating the parameters of the distribution F_{AB} by maximizing the likelihood density function of the observed standardized statistic (Z_i^A, Z_i^B), which is the convolution of the density function of (θ_i^A, θ_i^B) and a standard normal density function from equations 3.7. Once the parameters $\hat{\mu}_A, \hat{\mu}_B, \hat{\sigma}_A, \hat{\sigma}_B, \hat{r}$ are estimated using the mle method, we could generate samples $(\tilde{X}_A, \tilde{X}_B)$ from this bivariate normal distribution with parameters $\hat{\mu}_A, \hat{\mu}_B, \hat{\sigma}_A, \hat{\sigma}_B, \hat{r}$. Then the mle estimator of $R_2(T,T)$ is calculated by $R_2(t,t) = \sum_{i=1}^m I_{\{|\tilde{X}_i^A|>t \ \& \ |\tilde{X}_i^B|>t\}}/m$ for every t in T.

5. The rescaling estimator of $R_2(T,T)$ is calculated by transforming the observed statistic (Z_A, Z_B) to a new vector (X_A, X_B) which has the same mean and variance with the true parameter (θ_A, θ_B) and the correlation between (X_A, X_B) should equal to the correlation between (θ_A, θ_B) by using equation 3.13, then $R_2(T,T)$ is calculated by $R_2(t,t) = \sum_{i=1}^n I_{\{|X_i^A|>t \ \& \ |X_i^B|>t\}}/n$ for every t in T.

6. The copula estimator of $R_2(T,T)$ is calculated by estimating the functions t_A,

t_B, which maps the true standardized parameters (θ_A, θ_B) to (X_A, X_B), and (X_A, X_B) follows a standard bivariate normal distribution with correlation r. The functions t_A, t_B could be constructed using the method in section 3.2.2 if we do not know the model of the joint distribution of (θ_A, θ_B). Once $\hat{t}_A$, $\hat{t}_B$, r are estimated, generate large samples (X_i^A, X_i^B), $i = 1, \cdots, n$, $n > m$ from a standard bivariate normal distribution with correlation $\hat{r}$. Then the samples of the true standardized parameters (θ_i^A, θ_i^B) are

$$\tilde{\theta}_i^A = \hat{t}_A^{-1}(X_A), \quad \tilde{\theta}_i^B = \hat{t}_B^{-1}(X_B).$$

The copula estimate of $R_2(T, T)$ is calculated by $R_2(t, t) = \sum_{i=1}^{n} I_{\{|\tilde{\theta}_i^A|>t \ \& \ |\tilde{\theta}_i^B|>t\}}/n$ for every t in T.

7. The procedure was repeated 100 times to get the average value and the standard deviation of the estimate of $R_2(T, T)$ for each method. The plots comparing the true $R_2(T, T)$ and the average of the estimate of $R_2(T, T)$ and the approximate 95% confidence intervals for the estimate of $R_2(T, T)$ for each method are constructed.

From figure 3.7 and 3.8, we know that when the true parameters (θ_i^A, θ_i^B) follow a bivariate normal distribution, the moment, mle and rescaling method perform the best with unbiased estimates of $R_2(T, T)$ and smaller standard deviations of the estimates. This is because the parametric model we use for the joint distribution of (θ_i^A, θ_i^B) for moments and mle method is the same with the true model. Also we know that rescaling method performs well for Guassion cases. Copula method which does not depend on the structure of the joint distribution performs a little worse with a little more bias of the estimate of $R_2(T, T)$ and more standard deviations of the estimates. The plug-in estimator performs the worst.

From figure 3.9-3.12, we know that when the true parameters θ_i^A, θ_i^B follows a marginal t distribution with $df = 3$ or the true parameters θ_i^A, θ_i^B follows a marginal generalized normal distribution, the copula method gives the smallest bias of the estimate of the $R_2(T,T)$ to the true values of $R_2(T,T)$ than the other estimators, especially on the tail (when the true proportion is less than 0.1). Since the true joint distribution of (θ_i^A, θ_i^B) is not bivariate normal anymore, then the moment estimator, mle estimator and rescaling estimator perform bad. As a conclusion, If the joint distribution of the true standardized parameter (θ_i^A, θ_i^B) is much deviated from bivariate normal distribution, the copula method performs much better than the other parametric or nonparametric methods we illustrated in this paper.

3.6 Real data analysis

Here we propose our new copula-based method to estimate the common and unique associations in CKD data set which was introduced in section 1.2. The genomic data in the CKD dataset consist of microarray measurements of gene expression on specific cell types obtained from kidney tissue biopsy specimens taken early in the disease course. The main clinical parameter of interest is the GFR taken at the biopsy time. GFR is a widely used overall index of kidney function. Specifically, it estimates how much blood passes through the tiny filters in the kidneys, called glomeruli, each minute. Normal GFR results range from 90-120 mL/min, GFR below 60 mL/min implies moderate loss of renal function, and GFR below 30 mL/min is considered to be severe. The dataset includes genomic and clinical data for 195 subjects, and the gene expression data quantify gene expression for 12,023 distinct genes or transcripts. The subjects have one of several diseases that give rise to CKD. The diseases in the CKD dataset include DN, LD, MCD, HT, RPGN, IgA, PIMA, SLE, FSGS, here LD is a control group of people who are healthy. Our interest is to identify marker/outcome associations both within and across disease subgroups.

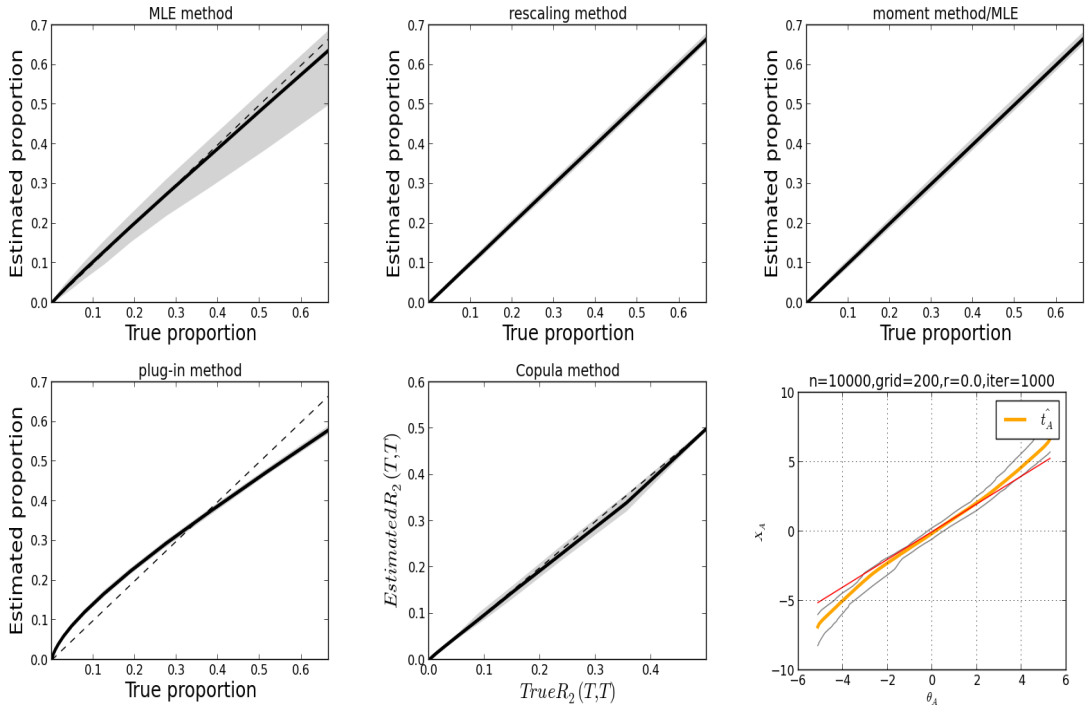

Figure 3.7: Plots compare the true standardized parameters θ_i^A, θ_i^B magnitude both greater than a sequence of thresholds T, $R_2(T,T)$ to the estimates of $R_2(T,T)$ for moments, mle, rescaling, copula and plug-in methods when the true standardized parameters θ_i^A, θ_i^B follow a bivariate normal distribution with mean 0 and std 1.0 and correlation 0. The lower right plot compare the true function t_A (red) with the estimated function $\hat{t}_A$ (orange) for copula method. Grey area is the approximate 95% confidence intervals for the estimators of $R_2(T,T)$, x axis is the true standardized parameter θ_A, y axis is the transformed standard normal vector $X_A = t_A(\theta_A)$.

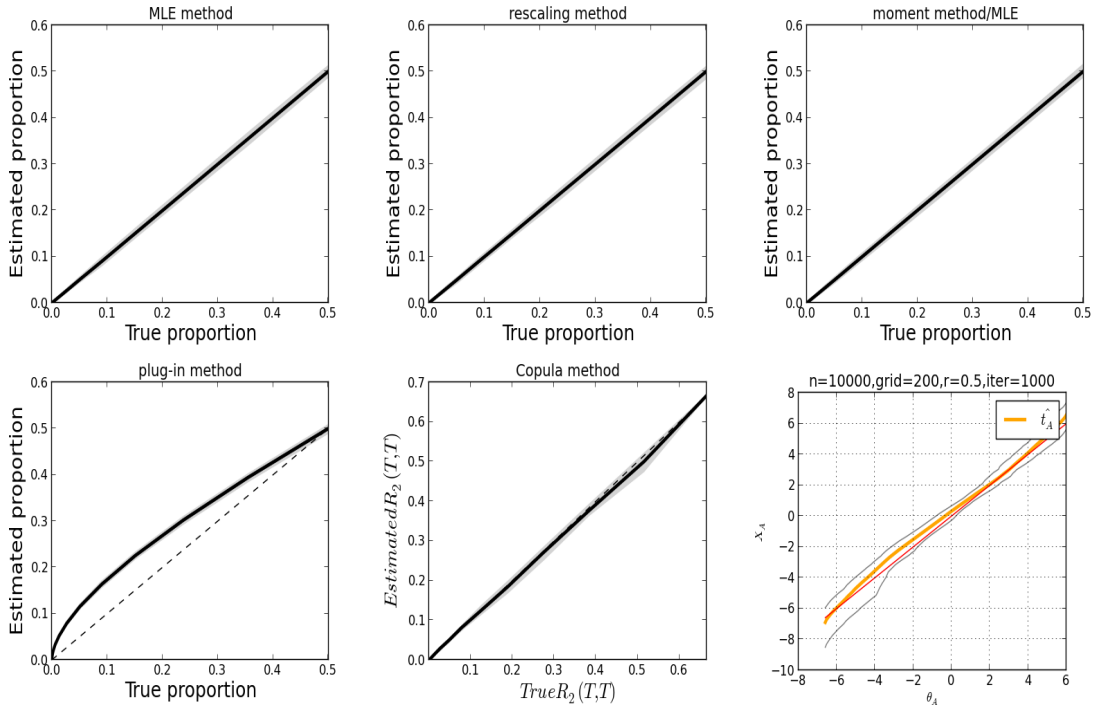

Figure 3.8: Plots compare the true standardized parameters θ_i^A, θ_i^B magnitude both greater than a sequence of thresholds T, $R_2(T,T)$ to the estimates of $R_2(T,T)$ for moments, mle, rescaling, copula and plug-in methods when the true standardized parameters θ_i^A, θ_i^B follow a bivariate normal distribution with mean 0 and std 1.0 and correlation 0.5. The lower right plot compare the true function t_A (red) with the estimated function $\hat{t}_A$ (orange) for copula method. Grey area is the approximate 95% confidence intervals for the estimators of $R_2(T,T)$, x axis is the true standardized parameter θ_A, y axis is the transformed standard normal vector $X_A = t_A(\theta_A)$.

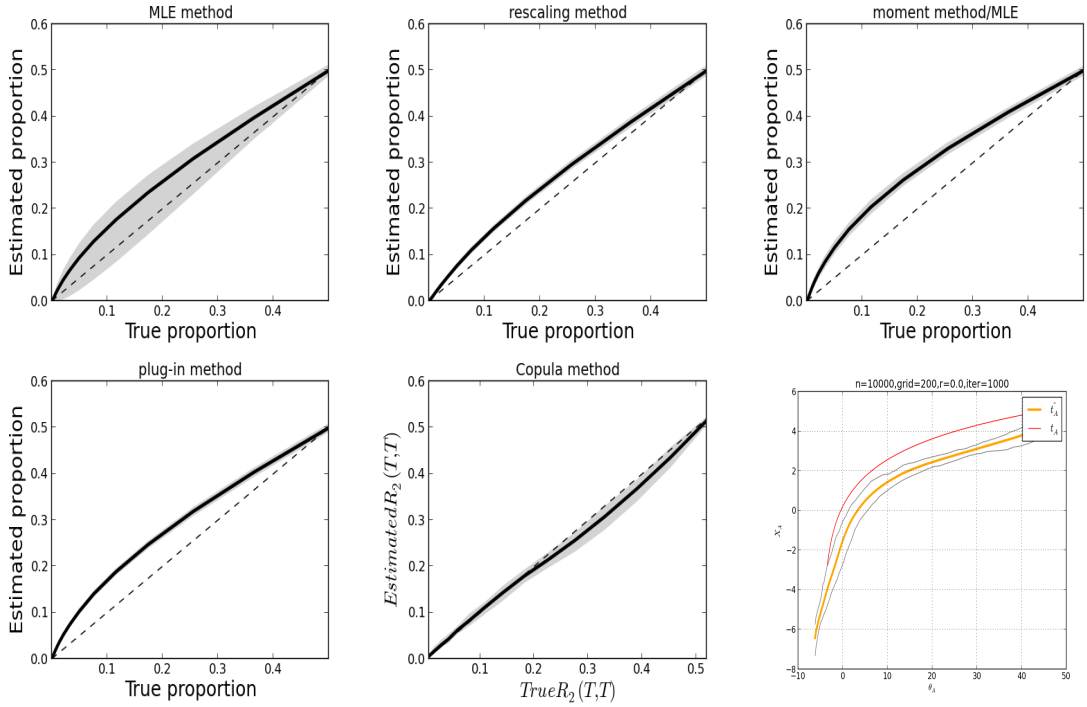

Figure 3.9: Plots compare the true standardized parameters θ_i^A, θ_i^B magnitude both greater than a sequence of thresholds T, $R_2(T,T)$ to the estimates of $R_2(T,T)$ for moments, mle, rescaling, copula and plug-in methods when the true standardized parameters θ_i^A, θ_i^B follow a marginal generalized normal distribution with parameters $\xi = -0.5, \alpha = 2, \kappa = -0.5$ with correlation 0.0. The lower right plot compare the true function t_A (red) with the estimated function $\hat{t}_A$ (orange) for copula method. Grey area is the approximate 95% confidence intervals for the estimators of $R_2(T,T)$, x axis is the true standardized parameter θ_A, y axis is the transformed standard normal vector $X_A = t_A(\theta_A)$.

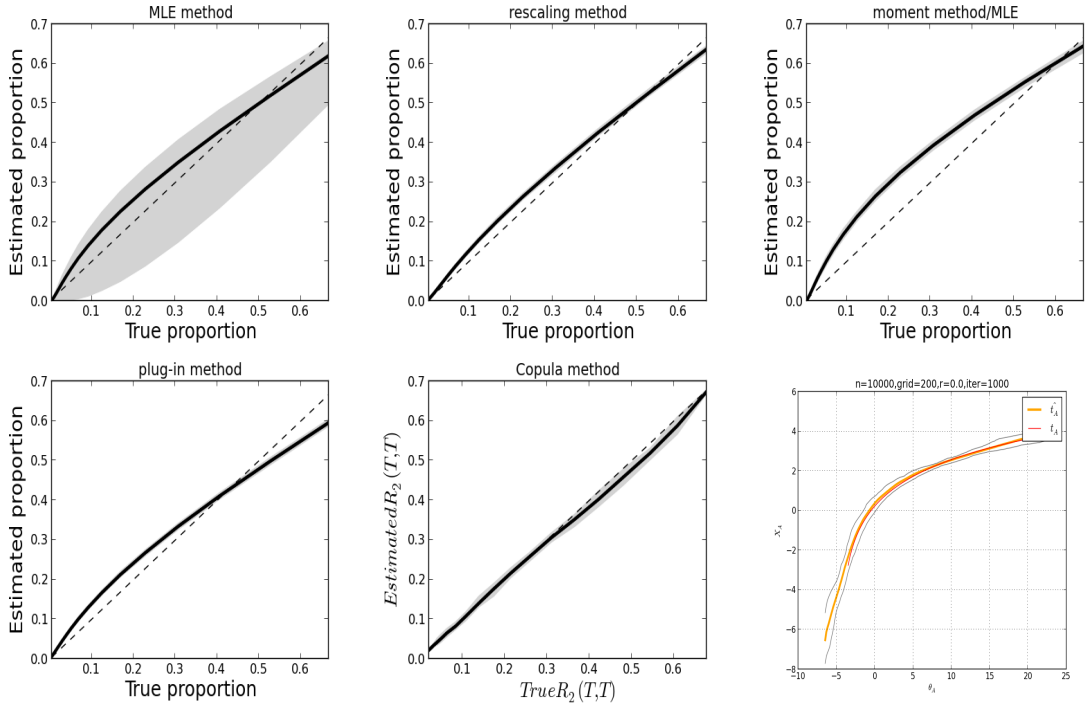

Figure 3.10: Plots compare the true standardized parameters θ_i^A, θ_i^B magnitude both greater than a sequence of thresholds T, $R_2(T,T)$ to the estimates of $R_2(T,T)$ for moments, mle, rescaling, copula and plug-in methods when the true standardized parameters θ_i^A, θ_i^B follow a marginal generalized normal distribution with parameters $\xi = -0.5, \alpha = 2, \kappa = -0.5$ with correlation 0.5. The lower right plot compare the true function t_A (red) with the estimated function $\hat{t}_A$ (orange) for copula method. Grey area is the approximate 95% confidence intervals for the estimators of $R_2(T,T)$,x axis is the true standardized parameter θ_A, y axis is the transformed standard normal vector $X_A = t_A(\theta_A)$.

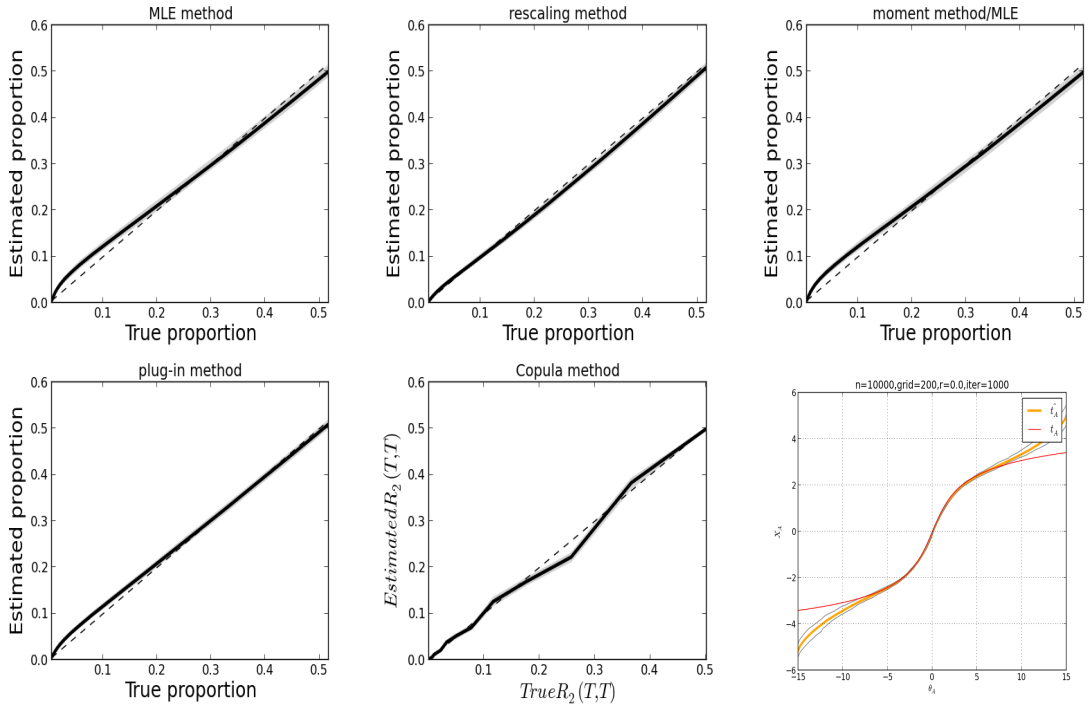

Figure 3.11: Plots compare the true standardized parameters θ_i^A, θ_i^B magnitude both greater than a sequence of thresholds T, $R_2(T,T)$ to the estimates of $R_2(T,T)$ for moments, mle, rescaling, copula and plug-in methods when the true standardized parameters θ_i^A, θ_i^B follow a marginal t distribution with $df = 3$ with correlation 0.0. The lower right plot compare the true function t_A (red) with the estimated function $\hat{t}_A$ (orange) for copula method. Grey area is the approximate 95% confidence intervals for the estimators of $R_2(T,T)$, x axis is the true standardized parameter θ_A, y axis is the transformed standard normal vector $X_A = t_A(\theta_A)$.

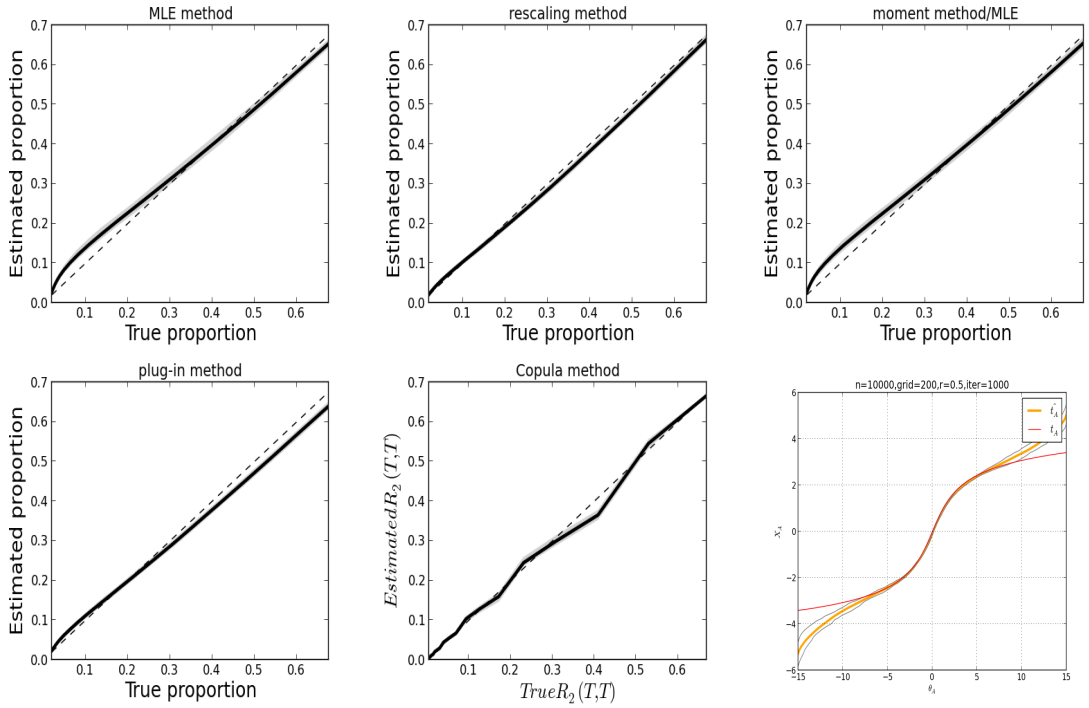

Figure 3.12: Plots compare the true standardized parameters θ_i^A, θ_i^B magnitude both greater than a sequence of thresholds T, $R_2(T,T)$ to the estimates of $R_2(T,T)$ for moments, mle, rescaling, copula and plug-in methods when the true standardized parameters θ_i^A, θ_i^B follow a marginal t distribution with $df = 3$ with correlation 0.5. The lower right plot compare the true function t_A (red) with the estimated function $\hat{t}_A$ (orange) for copula method. Grey area is the approximate 95% confidence intervals for the estimators of $R_2(T,T)$,x axis is the true standardized parameter θ_A, y axis is the transformed standard normal vector $X_A = t_A(\theta_A)$.

A direct way to investigate this is to estimate Pearson correlation coefficients $\hat{\rho}_i$ between the i^{th} marker and the outcome GFR based on available data, however, it is very hard to detect any interesting markers due to the small sample size from figure 3.13. Except for the subjects pooled together, we have 195 subjects which is fairly large and powerful to detect many interesting markers, the sample sizes for the disease subgroups individually are small. PIMA group has the largest number of subjects, which is 45 and LD group has the smallest number of subjects, which is only 10. From the right panel of figure 3.13, we could see that there are many markers detected from the false discovery rate analysis for the pooled group due to the large sample size, but the standard deviation of the effect sizes ρ_i for the pooled group is not the largest. However, RPGN disease subgroup has the highest standard deviation of the effect sizes but not much information from false discovery rate analysis. LD and DN disease subgroups have some effects while get nothing from false discovery rate analysis.

Since effect sizes ρ_i is invariant with different sample sizes, we will focus on the estimated correlation coefficient $\hat{\rho}_i^A$ of subgroup A to a variance stablized standardized statistic Z_i^A , by using the Fisher transformation

$$Z_i^A = \frac{\sqrt{n_A - 3}}{2} \log \frac{1 + \hat{\rho}_i^A}{1 - \hat{\rho}_i^A}.$$

Then

$$Z_i = \theta_i + \eta_i,$$
$$\theta_i^A \approx \tfrac{\sqrt{n_A-3}}{2} \log \tfrac{1+\rho_i^A}{1-\rho_i^A},$$

where η_i^A is approximately normal and θ_i^A is the true variance stablized standardized parameter. Now we are interested in estimating measure $R_1(t) = P(|\theta_i^A| > t)$ and $R_2(t_1, t_2) = P(|\theta_i^A| > t_1, |\theta_i^B| > t_2)$. Once the fisher transformed standardized

statistic Z_i^A for subgroup A is estimated from the real data set, we could use the same procedure as the simulation steps in univariate analysis to calculate the mle estimator, plug-in estimator and copula estimator of $R_1(T)$ for a sequence of thresholds T and the estimated marginal distribution of the true standardized parameter θ_i^A. For plug-in method, we just use the empirical distribution of the standardized statistic Z_i^A as the marginal distribution of θ_i^A. For mle method, we could estimate the parameters μ_A, σ_A of the estimated marginal distribution of θ_i^A, if we assume the distribution is normal. For copula method, we could first estimate the nonparametric function t_A that maps θ_i^A to X_i^A which follows standard normal distributions, then since $t_A = \Phi^{-1}F$, where F is the marginal distribution of θ_i^A, then $\hat{F} = \Phi(t_A)$ is the estimate of the marginal distribution of θ_i^A.

Figure 3.14 is the plots of the plug-in estimate, mle estimate and copula estimate of the CDF of true parameter θ_i^A for all the disease subgroups A and the subgroups pooled together. We could see that except for the pooled group, the mle estimate and the copula estimate of the CDF of θ_i^A are quite similar, implying that the true marginal distribution of θ_i^A is close to a normal distribution. Table 3.1 is the estimates of $R_1(2)$ of the three methods for all the disease subgroups and the subgroups pooled together. MLE estimator and copula estimator give similar result for most of the disease subgroups except for the pooled group while the plug-in estimator always bias up the true value of $R_1(2)$ since the sample statistic Z are always more dispersed than the true parameter θ which will lead to higher proportion of markers having effect sizes magnitude greater than some threshold t. Except for the pooled group, disease subgroups IgA, PIMA have the highest proportion (above 0.2) of markers having large effects, on the other side disease subgroups DN, MCD almost have no markers having effect size magnitude greater than 2.

Now we estimate the overlap measure $R_2(T, T)$. Once the fisher transformed standardized statistic (Z_i^A, Z_i^B) for each pair of disease subgroups A,B are estimated

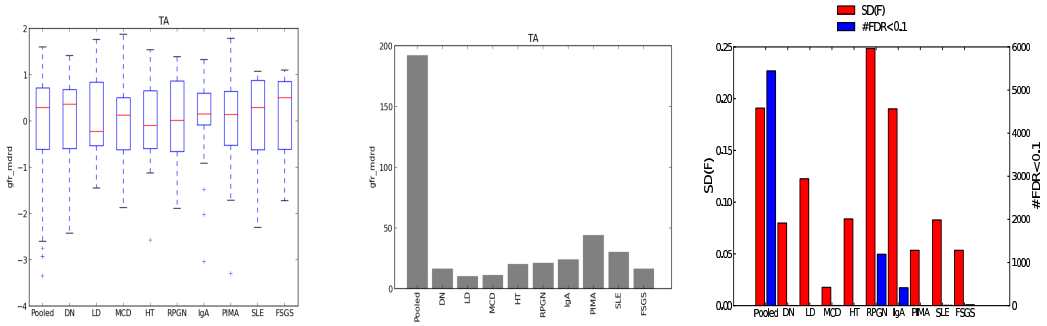

Figure 3.13: Right plot is a bar graph comparing results of false discovery rate analysis and standard deviation of the effect sizes for disease subgroups in CKD dataset; Middle plot is the bar graph of the number of subjects in disease subgroups in CKD data; Left plot is the boxplots of the outcome GFR in disease subgroups in CKD data.

Table 3.1: MLE, copula, plug-in estimate of $R_1(2)$ for different disease subgroups.

Disease subgroup	MLE estimate	Copula estimate	Plug-in estimate
Pooled	0.620	0.446	0.651
LD	0.087	0.07	0.218
DN	0	0	0.059
MCD	0.005	0.004	0.118
HT	0.042	0.050	0.167
RPGN	0.142	0.096	0.268
IgA	0.210	0.202	0.274
PIMA	0.021	0.015	0.146
SLE	0.130	0.128	0.244
FSGS	0.172	0.229	0.295

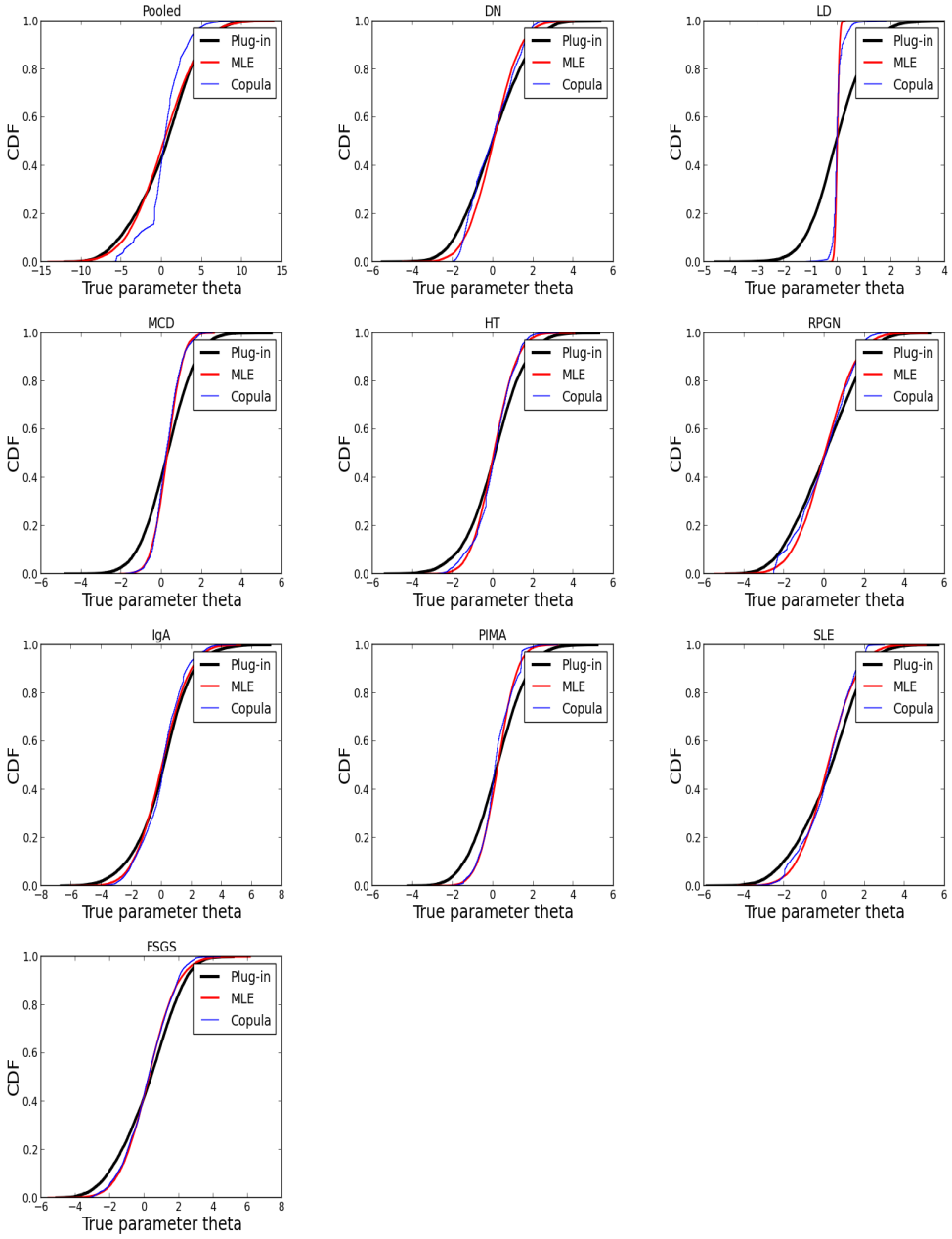

Figure 3.14: Plots of the estimated CDF of true parameters θ of mle, copula and plug-in method for disease subgroups and pooled together.

from the real data set, we could use the same procedure as the simulation steps in bivariate analysis to calculate the mle estimator, plug-in estimator and copula estimator of $R_2(T,T)$ for a sequence of thresholds T. Table 3.2 is the estimate of $R_2(2,2)$ from copula-based method for each pair of disease subgroups and table 3.3, 3.4 is the estimate of $R_2(2,2)$ from mle and plug-in method for each pair of disease subgroups.

From the copula estimates in table 3.1 we see that there are 9.6% of the markers having effect size magnitude greater than 2 in RPGN, 20.2% in IgA, 12.8% in SLE, and 22.9% in FSGS. From table 3.2, we see that 9.1% of the markers both having effect size magnitude greater than 2 in RPGN and IgA. Compare to the result in table 3.1, almost all the markers having effect size magnitude greater than 2 in RPGN also have effect size magnitude greater than 2 in IgA. From table 3.2, we also see that 10.9% of the markers both having effect size magnitude greater than 2 in RPGN and SLE. Compare to the result in table 3.1, almost all the markers having effect size magnitude greater than 2 in RPGN also have effect size magnitude greater than 2 in SLE. From table 3.2, we see that 6.9% of the markers both having effect size magnitude greater than 2 in RPGN and FSGS. Compare to the result in table 3.1, a large proportion of the markers having effect size magnitude greater than 2 in RPGN also have effect size magnitude greater than 2 in FSGS, but the proportion is lower than IgA and SLE.

For SLE, 10.5% of the markers both having effect size magnitude greater than 2 in SLE and IgA, and 8.2% of the markers both having effect size magnitude greater than 2 in SLE and FSGS. Then there is a larger proportion of markers having effect size magnitude greater than 2 in SLE have effect size magnitude greater than 2 in IgA than FSGS. In other words, there are more common associations in SLE and IgA than those in SLE and FSGS. At last, 11.2% of the markers both having effect size magnitude greater than 2 in FSGS and IgA, while 22.9% and 20.2% in each disease

Table 3.2: Copula estimate of $R_2(2,2)$ for different pairs of disease subgroups.

Copula estimate	DN	LD	MCD	HT	RPGN	IgA	PIMA	SLE	FSGS
Pooled	0.084	0.005	0.017	0.023	0.163	0.197	0.017	0.142	0.152
DN		0.0	0.0003	0.019	0.031	0.034	0.003	0.011	0.046
LD			0.0	0.0	0.0	0.0	0.0	0.0	0.0
MCD				0.0	0.0	0.001	0.001	0.005	0.0
HT					0.035	0.052	0.001	0.046	0.044
RPGN						0.091	0.0	0.109	0.069
IgA							0.005	0.105	0.114
PIMA								0.004	0.001
SLE									0.082

Table 3.3: MLE estimate of $R_2(2,2)$ for different pairs of disease subgroups.

MLE estimate	DN	LD	MCD	HT	RPGN	IgA	PIMA	SLE	FSGS
Pooled	0.075	0.0	0.003	0.034	0.116	0.191	0.013	0.106	0.149
DN		0.0	0.001	0.027	0.043	0.052	0.01	0.042	0.045
LD			0.0	0.0	0.0	0.0	0.0	0.0	0.0
MCD				0.003	0.003	0.004	0.001	0.005	0.002
HT					0.032	0.029	0.004	0.032	0.034
RPGN						0.110	0.005	0.103	0.111
IgA							0.008	0.104	0.132
PIMA								0.006	0.009
SLE									0.104

subgroup. This means that only a half of the markers have large effects in both FSGS and IgA.

3.7 conclusion and future direction

Many genomic studies involves large number of markers with small number of subjects, then it is powerless to detect any single effect. Such studies often involve populations that can be subdivided into several distinct subpopulations. Then we focus on the effect sizes of the marker/outcome associations which is invariant to the sample size and propose parametric and nonparametric methods to estimate the

Table 3.4: Plug-in estimate of $R_2(2,2)$ for different pairs of disease subgroups.

Plug-in estimate	DN	LD	MCD	HT	RPGN	IgA	PIMA	SLE	FSGS
Pooled	0.147	0.019	0.065	0.125	0.219	0.228	0.069	0.197	0.237
DN		0.005	0.012	0.044	0.06	0.075	0.024	0.064	0.069
LD			0.002	0.005	0.007	0.009	0.002	0.009	0.010
MCD				0.012	0.023	0.024	0.012	0.024	0.03
HT					0.061	0.08	0.016	0.061	0.074
RPGN						0.106	0.021	0.094	0.119
IgA							0.029	0.113	0.121
PIMA								0.022	0.027
SLE									0.115

overall distribution of the effect sizes and the magnitude of effect sizes greater than some thresholds both in univariate and bivariate populations. Especially, we proposed a copula-based nonparametric method to estimate the overlap measure of the magnitude of effect sizes both greater than some thresholds in two subpopulations.

In simulation study, we compare the accuracy of the estimate of the overlap measures through mle, moment, rescaling, copula and plug-in methods and find out that if the joint distribution of the true standardized parameter (θ_i^A, θ_i^B) is much deviated from bivariate normal distribution, the copula method performs better than the other parametric or nonparametric methods. Then we apply copula-based, mle and plug-in method to estimate the overlap measure of the common associations in each pairs of disease subgroups in CKD data. MLE estimator and copula estimator give similar result for most of the pairs disease subgroups, implying that the joint distribution of the effect sizes in any two disease subgroups is close to bivariate normal distribution.

CHAPTER IV

Statistical assessment of relationships between marginal properties of variables and their external correlations

4.1 Introduction

Modern medical studies often aim to identify genomic markers of individuals in a population that are associated with an external (i.e. non-genomic) trait. In hypothesis-generating research, these genomic markers must be identified from a large pool of candidate markers, most of which are irrelevant. For example, researchers in nephrology may be interested in identifying genes whose expression correlates with a measure of renal performance, such as the glomerular filtration rate (GFR). Simple statistics such as Pearson correlation coefficients or standardized group-wise mean differences are often used in this setting to identify potentially interesting markers. Screening analysis arise in many application areas, such as fraud detection (*Chen et al.* (2004)), astronomy (*Schreiber et al.* (2002)), and biomarker research, but here we will focus on applications in personalized medicine involving genomic markers.

Our setting is a screening study with n independent subjects, each of whom is assessed for a quantitative outcome $y_i \in \mathbb{R}$ $(i = 1, \cdots, n)$. In addition, each subject is assessed for gene expression on a large number of genes, we will write

X_{ij} for the measured expression of gene j in subject i. For each gene j, we can assess the association A_j between the expression levels of the gene and the outcomes. For a given measure Assoc of association, e.g. Pearson correlation, we have $\hat{A}_j = \text{Assoc}(y_1, \cdots, y_n; X_{1j}, \cdots, X_{nj})$. In a traditional "screening analysis", the aim is typically to identify a subset of markers that meet some level of statistical confidence such as a family-wise error rate or false discovery rate.

In many research settings, statistical power is low due to sample size limitations. Thus, while a few interesting markers may be found in a single study, there is often a sense that the data have much more to reveal. In this chapter, we focus on approaches for identifying global trends in the data that help us to understand what types of associations may be present. We illustrate that this can be accomplished even when the power is too low to attribute associations to specific variables.

Our main goal here is to ask whether properties of the marginal distributions of genomic markers can be identified that are statistically related to the strengths of their associations with the external trait. For our purposes, a property M is a function of the marginal distribution of one or more genes. For example, M_j could be the population mean of the j^{th} marker. A property/marker/outcome association is then any relationships between M_j and A_j.

Such relationships are completely empirical since association as measured by Pearson's correlation is location and scale invariant, there is no mathematical reason that a trend must exist between the marker/outcome correlations and the marginal properties of the markers. However, in genomic datasets, these trends often exist.

Some researchers have some related findings that genes with unique gene expression pattern may contain some useful information and of great interest. For example, gene pairs that have a large number of mutually exclusive outlier cancer samples are shown to be more likely involved in chromosomal translocations which are common in cancer and may be causal in the progression of the disease using COPA (Cancer

Outlier Profile Analysis) by *MacDonald and Ghosh* (2006). And Genes which have high connectivity (i.e. 'hub' genes) within a weighted co-expression network are significantly more likely to be essential for yeast viability demonstrated by *Zhang and Horvath* (2005).

Since the property/marker/outcome associations can be explicitly computed, it seems straightforward to assess whether these are somehow related. However, to fully understand this relationship, several challenges must be overcome. First, since the markers are highly correlated with each other and they all associate with the same outcome, then the marker/outcome correlations are highly dependent. This has the potential to bias the property/marker/outcome associations. We proposed a simulation-based approach to detect the bias and variability of this estimated association. A Second challenge is that there exist both monotone and non-monotone associations between genomic markers and the outcome. The monotone associations are largely captured by the Pearson correlation which has favorable statistical properties. However the non-monotone associations like u-shape association are not detected by the Pearson correlation coefficient. We develop a way that decomposes an association into a monotone component and a symmetric concave/convex component (plus a residual function) to see which association is dominant.

The chapter is organized as follows. In section 4.2, we will introduce the five marginal properties of the distribution of markers we are interested in including the definitions of the statistical measures of them. In section 4.3, we will introduce the quantile regression method we use to model the property/marker/outcome associations with the L_1 goodness of fit. Then we will talk about the statistical properties of the property/marker/outcome associations. At last, we will illustrate the method that decomposing the associations into monotone components and symmetric concave/convex components plus a residual function. In section 4.4, we show the simulation results for the properties of the property/marker/outcome associations. In

section 4.5, we use an example of CKD dataset to show the quantile regression results of the property/marker/outcome associations and find out that skewness is the most dominant property. And also we notice that monotone trends between GFR and symmetric genes are more likely to happen than symmetric concave trends by decomposing the association into monotone and symmetric concave components. Section 4.5 is the conclusion and future challenges for this chapter.

4.2 Marginal properties of variables involved in external associations

Our overall goal is to relate the marker/outcome associations A_j to the properties of the marginal distribution of markers M_j. This motivation builds on some standard practices used when analyzing large-scale genomic data. For example, filtering methods (*Hackstadt and Hess* (2009)) are usually used to reduce the number of hypothesis tests and therefore increase the power to detect associations among the non-filtered candidates. Common filtering methods include excluding genes with low variance or low mean (abundance) without referring to any non-genomic data. The rationale for filtering by variance is that small changes in absolute levels are less likely to be driving factors for changes in the phenotype. Moreover, such small changes are difficult to distinguish from measurement noise. For example, in the CKD data, we will show that genes with low variance tend to have weaker associations with GFR. Figure 4.1 shows the scatterplots of GFR and gene expression for genes with high and low variance. Similarly, one may argue that genes with very low absolute abundance are less likely to drive variation in the phenotype, and it is also a challenge to accurately measure the abundance of such genes with low expression.

The filtering method indicates that mean and variance of the distribution of the markers may relate to the marker/outcome associations. So mean and variance and

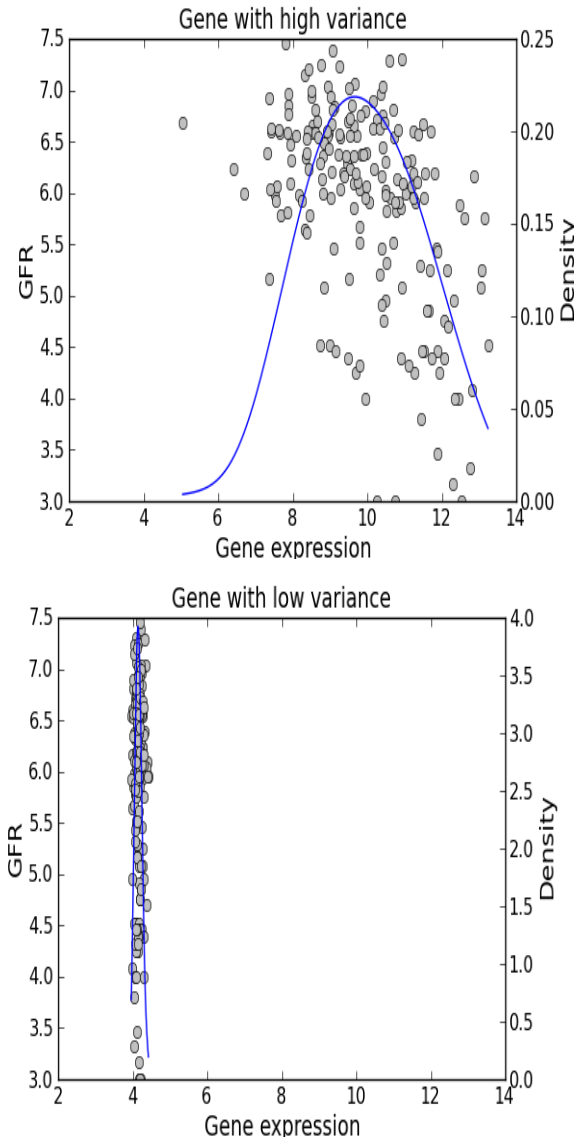

Figure 4.1: Scatterplots of GFR and gene expression for two specific genes with high and low variance.

other familiar statistics like the ordered moments in a distribution will be included in our marginal properties, so do some statistics specific to the genomic study. Also the properties could be some function of the observed values of each individual marker, denoted as $M(X_{1j}, \cdots, X_{nj})$ or some function of the observed values of one marker and the other markers, denoted as $M(X_{1j}, \cdots, X_{nj}; X)$. But all of them are using the empirical distributions of the markers without any information of the outcome.

To be concrete, we only focus on the gene expression data which measures the information from a gene which is used in the synthesis of a functional gene product. These products are often proteins or a functional RNA. In genetics, gene expression is the most fundamental level at which the genotype gives rise to the phenotype. So it is of essential importance for comparative investigations aiming at discovery of new genes, functional classification of genes, discovery of relationships between genes and their products.

4.2.1 Mean level of gene expression data

In microarray experiment, lowly expressed genes should be less important than highly expressed genes providing a simple and common explanation for the general relationship observed between gene expression and the different facets of gene evolution (*Gout et al.* (2010)). And also reliable measurement is more achievable for highly expressed genes in a target sample than for those expressed at low levels. Thus, most of the studies have focussed on high-expressing genes that have high signal intensities on microarrays. However, this approval may bias the conclusions. Some genes with low gene expression levels have been detected as important too. For example, low expression levels of soluble CD1d gene in patients with rheumatoid arthritis had been shown by *Kojo et al.* (2003). So the first marginal feature of gene expression data we are interested in is mean of expression levels for each gene across subjects. Figure 4.2 gives the distribution of mean expression levels of genes in some genomic data after

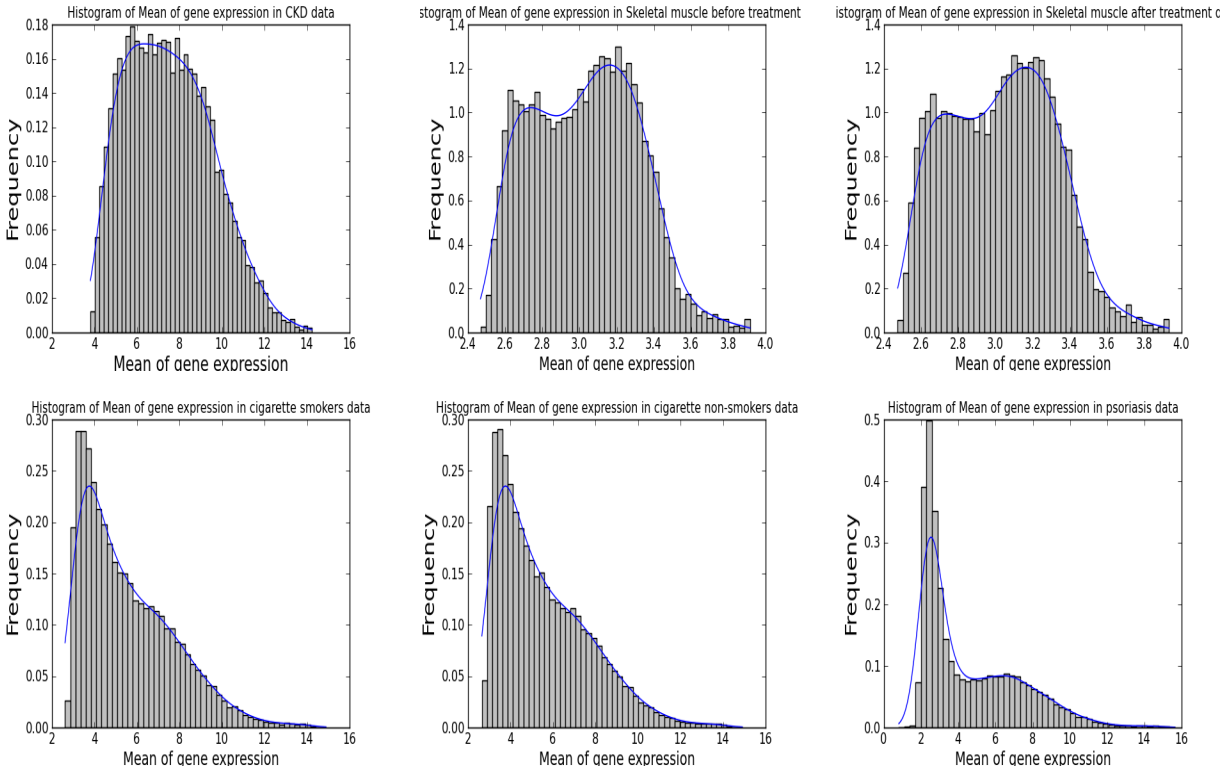

Figure 4.2: Distributions of mean expression levels of genes in CKD, Skeletal, Psoriasis and Cigarette datasets.

the $\log_2$ transformation of expression levels. From figure 4.2, we could see that the distributions of the mean property could be slightly skewed to the right or skewed to the left in different datasets. Whether genes with high mean expression levels tends to have more strong relations with the outcome is our question.

4.2.2 Variance of gene expression data

Much of our understanding of biological system is based on interpreting average behavior, variance has been largely ignored because it has been considered solely in the context of experimental reproducibility. Now, there is evidence that biological sources of variance may play an important role in determining cellular and organismal phenotypes, as well as in helping to explain a wide range of biological phenomena ranging from reduced penetrance to evolutionary fitness (*Mar et al.* (2011)). If the

genes have very low variance, a natural interpretation is that those genes are themselves highly constrained, and then are less likely to be a driven factor for the changes in the phenotype. We will check if this is always true.

Here we define our measure of variance of each gene to be:

$$\text{IQR}(p) = \text{pth Quantile} - (1-p)\text{th Quantile}$$

which is robust to outliers, and when $p = 0.75$, this measure is IQR and when $p = 0.9$, this measure is IDR.

We may identify the parameter p that has the largest marginal association between IQR(p) and their external correlations or just choose $p = 0.75$ as usual. Figure 4.3 is the distribution of IQR of genes in some genomic data. From figure 4.3, we could see that the distribution of IQR have a very long right tail, most of the IQRs of genes are around 0-1, with a few genes with very high IQR. Then whether genes with higher IQR tend to have stronger signals and whether variance is the most important properties associates with external correlations of all the properties we considered is our main interest.

4.2.3 Outliers of gene expression data

We should mention that while the term "outlier" has a pejorative meaning in statistics, it is a very meaningful concept in a biological sense. As noted by *Lyons et al.* (2004) and subsequently by *Tomlins et al.* (2005), the biology of oncogenesis permits that unique sets of genes may be involved in tumor development across patients. While statistical outliers refer to measurements that exceed the expected variation in a set of data, the oncogenetic outliers we seek to find will be putatively related to cancer processes.

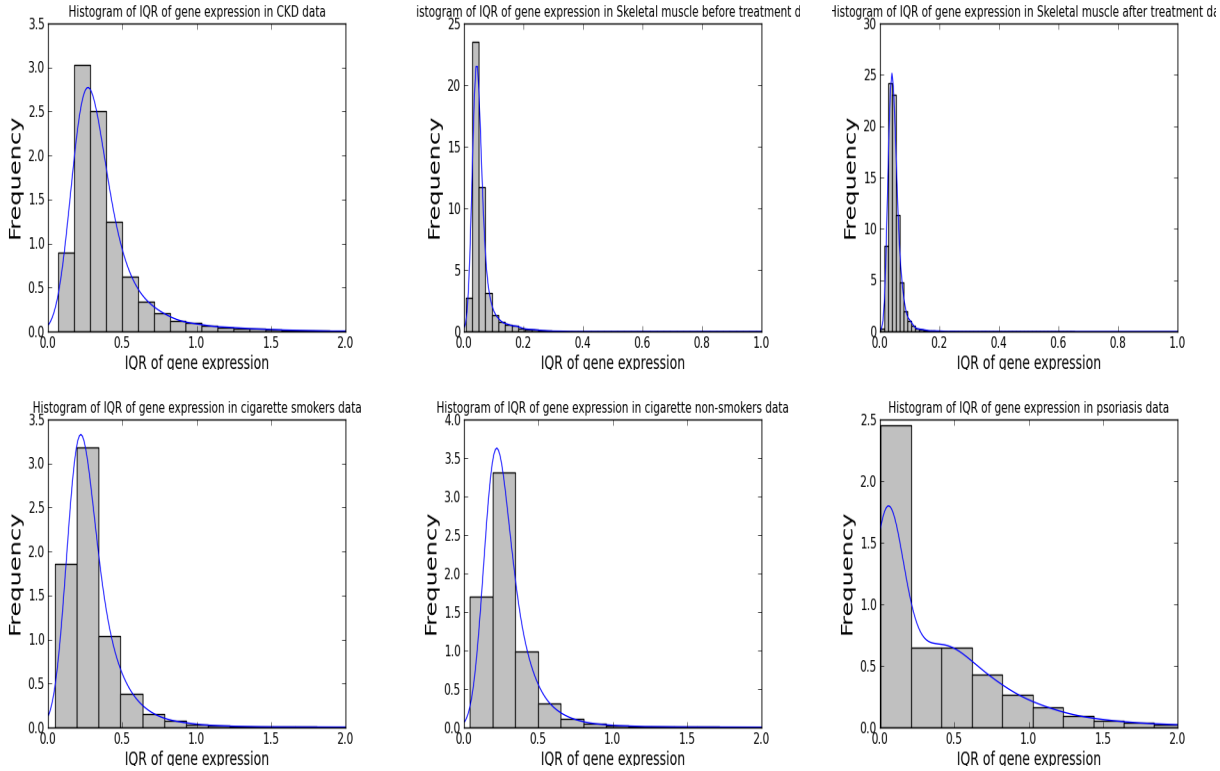

Figure 4.3: Distributions of IQR of genes in CKD, Skeletal, Psoriasis and Cigarette datasets.

The measure to quantify how many and how strong the outliers is:

$$\text{Outlier} = \sum_{j=1}^{n}(|\frac{X_{ij} - \bar{X}_i}{SD(X_i)}| - k)I\{(|\frac{X_{ij} - \bar{X}_i}{SD(X_i)}| - k) > 0\},\ i = 1, \cdots, m,$$

here we standardized our gene expression data and compare it with some threshold k ($k = 2$ in this paper), and to make it robust to outliers, we use IQR/1.349 instead of SD(X). Since if we have many outliers, the SD(X) tends to be large, this will make the standardized value to be small, then have small outlier measure, which kind of cancel each other. Some researcher might want to know if more outliers will lead to higher correlations between outcome and gene markers. Figure 4.4 shows the distribution of outlier measures of genes in some genomic data. From figure 4.4, we could see that the distribution of outlier measures is strongly skewed to the right, with most of the outlier measures around 0-20, while the others have very large outlier measures. Outlier measures equals 20 means that 20 subjects in this genes have absolute standardized gene expression value to be 3 with threshold $k = 2$.

4.2.4 Skewness of gene expression data

Gene expression data tend to have a large proportion of skew and heavy tailed genes, so we usually take log transformation of the gene expression data to obtain normality. But there are still some genes are highly skewed after log transformation. The measure to quantify skewness is:

$$\text{Skew} = \sum_{i=1}^{n}(X_i - \bar{X}_i)^3/SD(X_i)^3.$$

Here we also use IQR/1.349 instead of SD(X) to make it robust to outliers. Variance, outlier and skewness are three measures of the variation of gene expression data set. They measure different aspects but have some relationships, like high skewness will lead to high variance and outlier measure, but high outlier may not lead to high

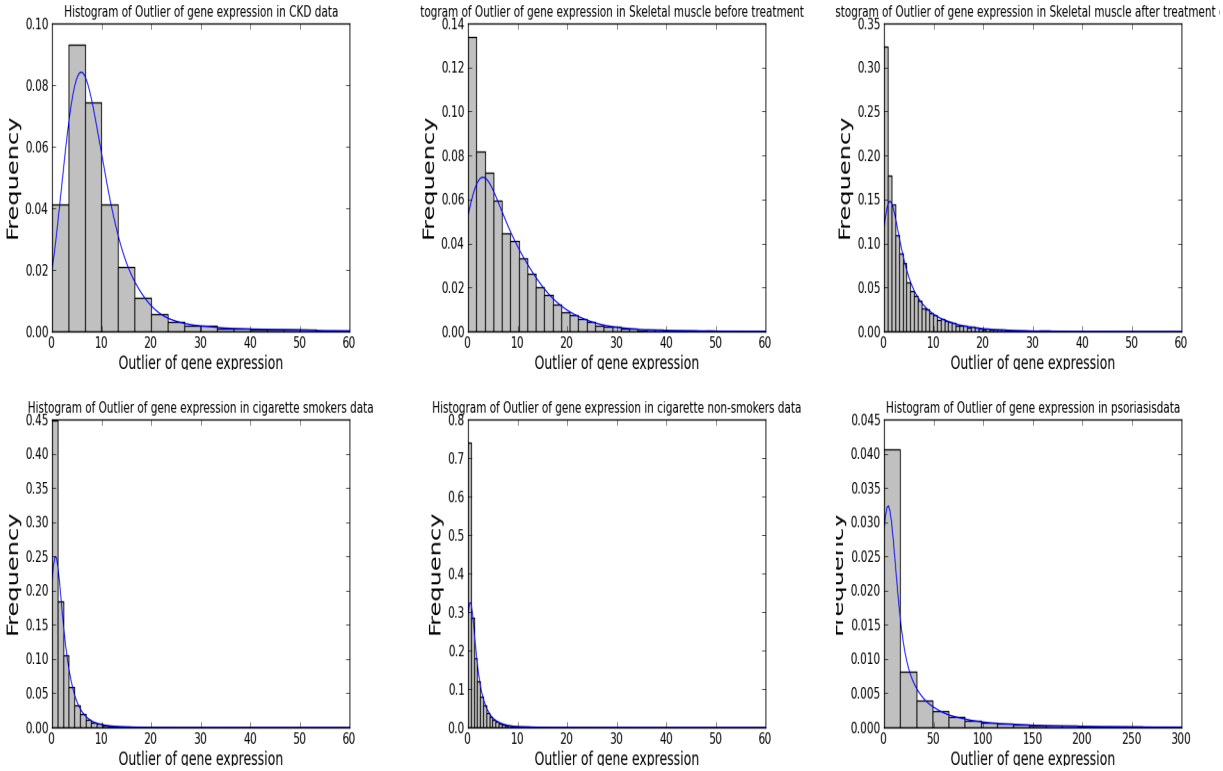

Figure 4.4: Distributions of outlier measures of genes in CKD, Skeletal, Psoriasis and Cigarette datasets.

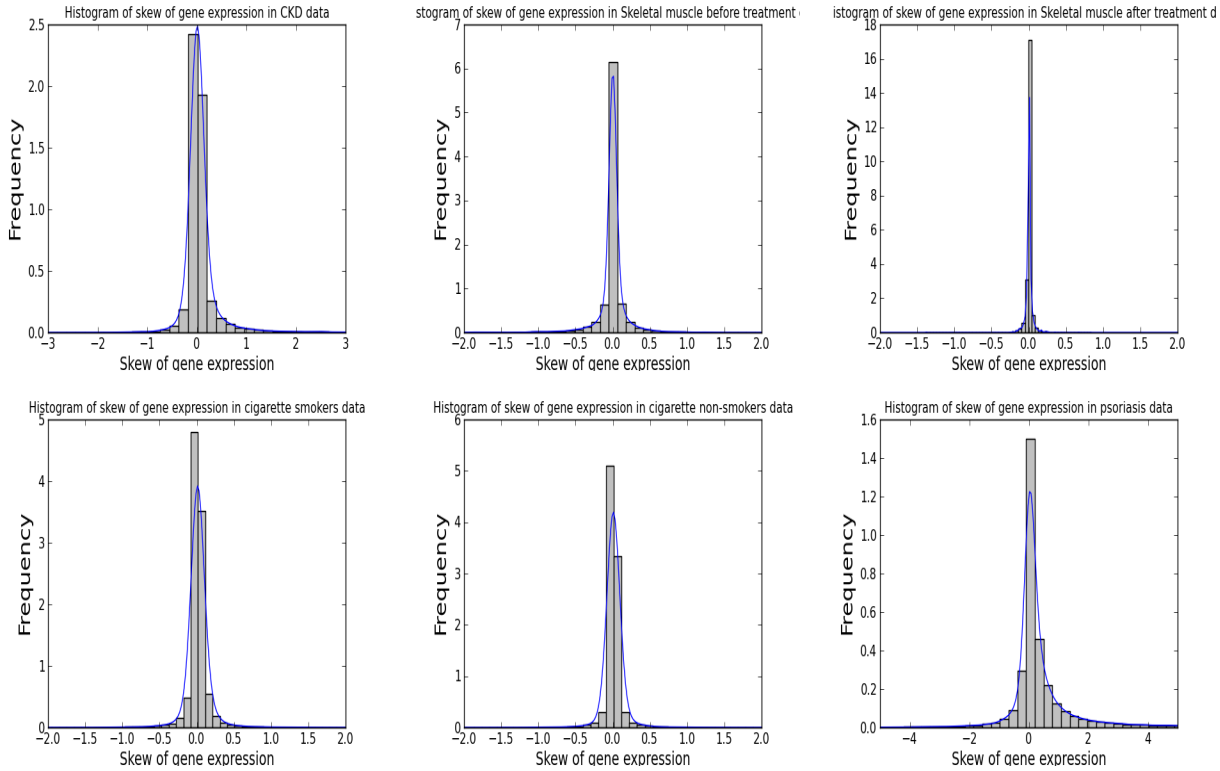

Figure 4.5: Distributions of skewness of genes in CKD, Skeletal, Psoriasis and Cigarette datasets.

skewness. Controlling the marginal association between these three measures to be not so strong is necessary before include them into regression model.

Figure 4.5 is the distribution of skewness of genes in some genomic data. The distribution of skewness is almost symmetric around 0 and most of the genes have skewness around -0.2 to 0.2, which means that most of the genes are symmetric with only a few genes have very strong positive or negative skewness. Then we are interested in whether genes with high absolute skewness tend to have stronger association with the external trait, and within these genes, whether subjects who have expression levels in the tail of the skewed distribution are in bad or good condition.

4.2.5 Gene connectivity

Genes and their protein products carry out cellular processes in the context of functional modules and are related to each other through a complex network of interactions. Correlation of gene expression across a wide variety of experimental perturbations has been shown to cluster genes of similar function. A gene which is highly correlated with many other genes based on gene expression level is called highly connected nodes in Network and has been found to be relatively more important. For example, Genes which have high connectivity (i.e. 'hub' genes) within a weighted co-expression network are significantly more likely to be essential for yeast viability demonstrated by *Mar et al.* (2011).

In gene co-expression networks, each gene corresponds to a node. The neighbors of a node i are the nodes that are connected to the node i. Two genes are connected by an edge with a weight indicating the connection strength. A gene co-expression network can be represented by an adjacency matrix $A = [a_{ij}]$, where a_{ij} is the weight of a connection between two nodes i and j. The connectivity equals the sum of connection weights substract some threshold.

The choice of the adjacency function determines whether the resulting network will be weighted (soft thresholding) or unweighted (hard thresholding). A widely used adjacency function is the signum function which implements 'hard' thresholding involving the threshold parameter τ. Specifically,

$$a_{ij} = I(|\text{cor}(x_i, x_j)| > \tau).$$

Zhang and Horvath (2005) proposed a 'soft' power adjacency function:

$$a_{ij} = |\text{cor}(x_i, x_j)|^{\beta}$$

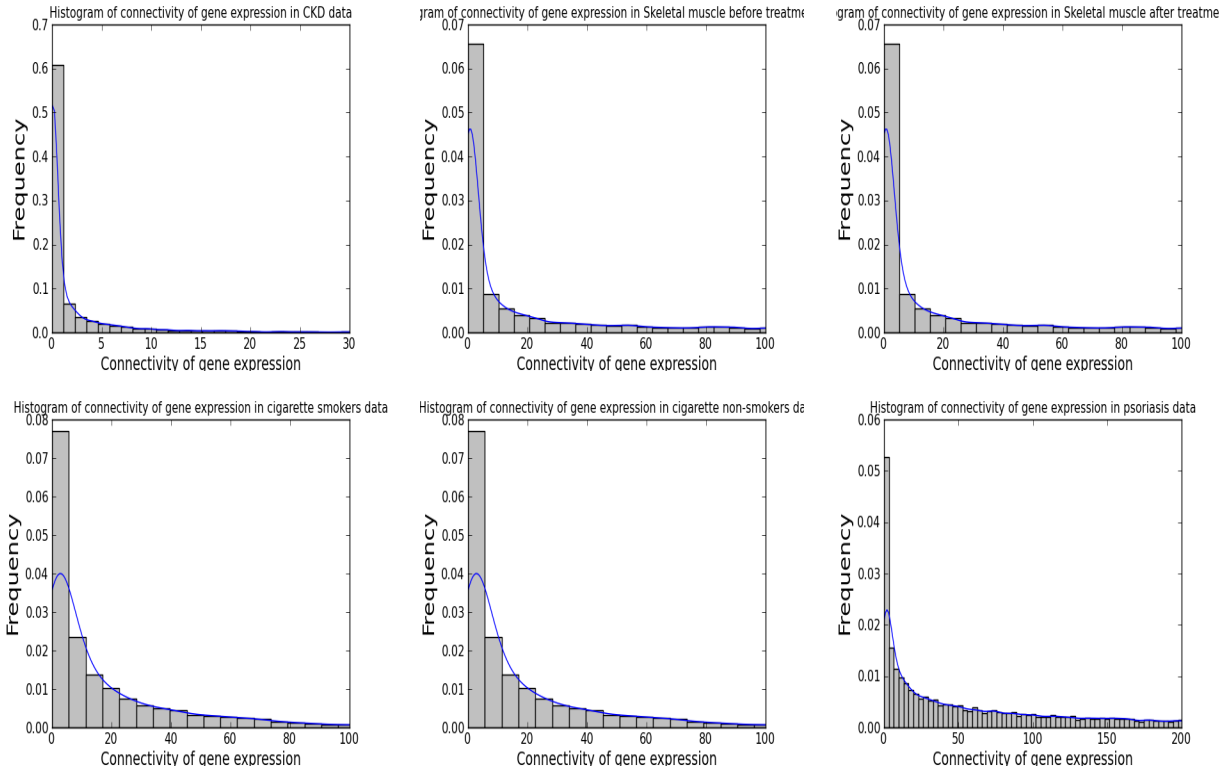

Figure 4.6: Distributions of connectivity of genes in CKD, Skeletal, Psoriasis and Cigarette datasets.

with the single parameter β. To choose the parameters of an adjacency function: Only those parameter values that lead to a network satisfying scale-free topology at least approximately were considered (e.g. signed $R^2 > 0.80$).

Here we choose

$$a_{ij} = (|\text{cor}(x_i, x_j)| - \tau)I(|\text{cor}(x_i, x_j)| - \tau > 0)$$

Then our connectivity measure is:

$$\text{Connectivity} = \sum_{j=1}^{n} a_{ij} = \sum_{j=1}^{n} (|\text{cor}(x_i, x_j)| - \tau)I(|\text{cor}(x_i, x_j)| - \tau > 0),$$

similar with hard thresholding function with parameter $\tau = 0.6$.

Figure 4.6 shows the distribution of measures of connectivity of genes in some

genomic data. The distribution is strongly skewed to the right, with most of genes have connectivity measures equal 0, which means that they are not highly connected with other genes. Connectivity measure equals 1 means that this gene is highly correlated with other 10 genes with absolute correlation 0.7 with threshold $\tau = 0.6$. Then genes with high connectivity are more interesting to us.

4.3 Methodology

4.3.1 Introduction

The project presented in this chapter contributed three novel methodological ideas. The first contribution is a new framework for understanding marginal marker/outcome associations in large datasets. This framework involves familiar summary statistics such as the Pearson correlation coefficients, but applies it in a non-standard way to derived quantities, rather than directly to the observations. The second contribution addresses the challenge of assessing the uncertainty in statistics that are aggregated over large data sets within complex and poorly understand dependencies. We show that commonly used randomization approaches, while intuitive, can give misleading results, and we provide a simulation based alternative approach that appears to perform well in a variety of situations. The third contribution addresses the issue of marker/outcome relationships that are strongly non-monotonic. We propose a decomposition of such relationships into monotonic and "u-shaped" components. This decomposition allows us to assess the prevalence of these two types of dependency in large datasets.

4.3.2 Statistical property of the property/marker/outcome associations

If we use Pearson correlation to represent the marker/outcome association, then the sample association between gene j and the outcome is

$$\hat{A}_j = \widehat{Cor}(X_j, Y) = \sum_{i=1}^{n} \frac{(y_i - \bar{y})(X_{ij} - \bar{X}_j)}{\hat{\sigma_y}\hat{\sigma_{X_j}}}, \ j = 1, \cdots, m,$$

where m is the number of genes and n is the number of subjects. The marginal property of the distribution of gene j is M_j. Then we are interested in the relationship between the sample marker/outcome association $\hat{A}$ and marginal property $M(X)$. The most obvious way is to look at the Pearson correlation between marker/outcome association and marginal property, which is

$$\theta = Cor(\hat{A}, M(X)),$$

then the estimated correlation is

$$\hat{\theta} = \widehat{Cor}(\hat{A}, M(X)).$$

Since the markers X_j are highly correlated with each other and they all associate with the same outcome Y, then the marker/outcome associations A_j are highly dependent with each other. If we look at the correlation between marker/outcome associations and marginal properties, there might be some "build-in effect" that $E(\hat{\theta}) \neq \theta$ and the stability of the $\hat{\theta}$ is of concern.

If marker/outcome association is zero, or X and Y are independent, the population correlation θ is:

$$
\begin{aligned}
\theta &= \mathrm{Cor}(\sum_{i=1}^{n} \frac{(y_i - \bar{y})(X_i - \bar{X})}{\hat{\sigma_y}\hat{\sigma_X}}, M(X))) \\
&= E(\sum_{i=1}^{n} \frac{(y_i - \bar{y})(X_i - \bar{X})}{\hat{\sigma_y}\hat{\sigma_X}} M(X)) - E(\sum_{i=1}^{n} \frac{(y_i - \bar{y})(X_i - \bar{X})}{\hat{\sigma_y}\hat{\sigma_X}}) E(M(X)) \\
&= \sum_{i=1}^{n} E(\frac{y_i - \bar{y}}{\hat{\sigma_y}}) E(\frac{(X_i - \bar{X})M(X)}{\hat{\sigma_X}}) - (\sum_{i=1}^{n} E(\frac{y_i - \bar{y}}{\hat{\sigma_y}}) E(\frac{X_i - \bar{X}}{\hat{\sigma_X}})) E(M(X)) \\
&= 0
\end{aligned}
$$

The expected sample correlation $\hat{\theta}$ is:

$$
\begin{aligned}
E(\hat{\theta}) &= E(\widehat{Cor}(\sum_{i=1}^{n} \frac{(y_i - \bar{y})(X_i - \bar{X})}{\hat{\sigma_y}\hat{\sigma_X}}, M(X))) \\
&= E(\sum_{j=1}^{m}\sum_{i=1}^{n} \frac{(y_i - \bar{y})(X_{ij} - \bar{X_j})}{\hat{\sigma_y}\hat{\sigma_{X_j}}} M(X_j)) \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n} E(\frac{(y_i - \bar{y})(X_{ij} - \bar{X_j})}{\hat{\sigma_y}\hat{\sigma_{X_j}}} M(X_j)) \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n} E_x E(\frac{(y_i - \bar{y})(X_{ij} - \bar{X_j})}{\hat{\sigma_y}\hat{\sigma_{X_j}}} M(X_j)|X) \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n} E_x(\frac{(X_{ij} - \bar{X_j})M(X_j)}{\hat{\sigma_y}\hat{\sigma_{X_j}}} E(y_i - \bar{y}|X)) \\
&= 0
\end{aligned}
$$

Then we know that there is no bias of $\hat{\theta}$ when X and Y are independent. Further to check the variability of $\hat{\theta}$, we will propose both a simulation-based approach and a data-based approach.

Usually people will do a permutation test (also called a randomization test) to

obtain the distribution of the test statistic under the null hypothesis by calculating all possible values of the test statistic under the rearrangement of the labels on the observed data points. Here the test statistic is the sample correlation $\hat{\theta}$ and the null hypothesis is that there is no marker/outcome association and therefore no property/marker/outcome association, $\theta = 0$. But this will always mislead the result that the standard error for $\hat{\theta}$ is always much smaller than the real one. So in this chapter, we will use simulation-based approach to detect the sampling distribution of $\hat{\theta}$ when $\theta = 0$.

4.3.3 Function decomposition

Since monotone trend is just one type of marker/outcome relationships which could be represented by Pearson correlation, in our studies, there exists at least one other type of marker/outcome relationships, which is called "u-shaped" or symmetric convex relationship. Then both the lower value and the higher value of marker X will lead to high/low outcome Y, while the middle values of X will not. These types of non-monotone associations could not be represented by Pearson correlation, researchers proposed other measures like R^2 from fitting natural cubic spline models to the marker/outcome relationships (*Lin et al.* (2008)) to represent the non-monotonic associations. Here we focus on detecting the "u-shaped" associations from the linear/monotone association and assess the prevalence of these two types of associations by using a decomposition method.

We will decompose Y into a monotone function and a symmetric convex function of X (plus a residual term). Then,

$$E(Y|X) = f_{sc} + f_m$$

where f_{sc} is a symmetric convex function which is a combination of many symmetric

convex basis functions. f_m is a monotone function which is also a combination of many monotone basis functions. In our situations, we will use the general form

$$sign(X - \bar{X})|X - \bar{X}|^p$$

as the monotone basis function and

$$|X - \bar{X}|^p$$

as the symmetric convex basis function. Here we choose $p = 0.5, 1, 2, 3$ for monotone basis and add anther basis $\arctan(X - \bar{X})$ for monotone function and choose $p = 0.5, 1, 1.5, 2$ for non-monotone basis and add another basis $\log(X - \bar{X} - 1)$ for non-monotone function. Also we need a constraint for the regression model that the coefficients for monotone basis functions have the same sign and also the coefficients for the symmetric convex basis functions have the same sign since the linear combination of monotone functions are not always a monotone function unless the scalars are all non-negative or non-positive. We use the nnls package in R program which make the signs of all the coefficients in the model to be non-negative. By changing the sign of the predictors in the model, we could regress the model in four situations, when the sign of the coefficients of the monotone function is positive/negative and the sign of the coefficients of the symmetric convex function is positive/negative. Then choose the situation of the highest R^2.

Partial R^2 for each combination of basis functions is used to quantify the prevalence of monotone and non-monotone associations.

4.3.4 Quantile regression model and B-spline basis

In genomic studies, it is known that the distribution of the correlations between gene expression and traits is heavy-tailed due to the exist of genes have strong effects.

Also the high positive and high negative correlations are more important than the average correlation, so the conditional mean model can then become an inappropriate measure just focusing on central tendencies but fail to capture informative trends in the response distribution. It is quite natural to go beyond location and scale effects of predictor variables on the response and ask how changes in the predictor variables affect the underlying shape of the distribution of the response.

Quantile regression, which models conditional quantiles as function of predictors, specifies changes in the conditional quantile of the dependent variable associated with a change in the covariates. Since multiple quantiles can be modeled, it is possible to achieve a more complete understanding of how the marker/outcome correlations are affected by marginal properties of the markers, including information about shape change. As in linear regression, the methodology we present is easily adapted to more complex model specifications, including interaction terms and polynomial or spline functions of covariates.

The quantile regression model can be expressed as:

$$Q_Y(\tau|X) = \alpha(\tau) + \beta(\tau)X,$$

where τ is the possible quantiles of the outcome we are interested in. We may estimate the coefficients directly by minimizing the weighted sum of absolute residuals:

$$V(\tau) = \sum_{i=1}^{n} \rho_\tau(y_i - \alpha - \beta X_i),$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$. Compared to linear regression that the coefficients are estimated by minimizing the sum of squared residuals

$$S = \sum_{i=1}^{n} (y_i - \alpha - \beta X_i)^2.$$

Since the goodness of fit R^2 for linear model is defined by using 1 substract the sum of squared residuals of the full model $\hat{S}$ over the sum of squared residuals of the model with no covariate $\tilde{S}$:

$$R^2 = 1 - \hat{S}/\tilde{S} = 1 - ||Y - \hat{Y}||^2/||Y - \bar{Y}||^2.$$

We may proceed in the same manner for quantile regression, we define $R_1(\tau)$ as the goodness of fit for quantile regression with quantile τ which is defined by using 1 substract the weighted sum of absolute residuals of the full model $\hat{V}(\tau)$ over the weighted sum of absolute residuals of the model with no covariate $\tilde{V}(\tau)$,:

$$R^1(\tau) = 1 - \hat{V}(\tau)/\tilde{V}(\tau),$$

where $\hat{V}(\tau) = \sum_{i=1}^{n} \rho_\tau(y_i - \hat{\alpha} - \hat{\beta}X_i)$, $\tilde{V}(\tau) = \sum_{i=1}^{n} \rho_\tau(y_i - a)$, constant a is the τ's quantile of y. The partial $R^1(\tau)$ is defined by

$$\text{partial } R^1(\tau) = \frac{R_2^1(\tau) - R_1^1(\tau)}{1 - R_1^1(\tau)},$$

where $R_1^1(\tau)$ is the goodness of fit R^1 of a quantile regression model without including the variables you are interested in for a particular quantile τ, and $R_2^1(\tau)$ is the goodness of fit R_1 of a full quantile regression model for a particular quantile τ.

Like R^2, it is immediately apparent that $\hat{V}(\tau) \leq \tilde{V}(\tau)$, and this $R^1(\tau)$ lies between 0 and 1. Unlike R^2, which measures the relative success of two models for the conditional mean function in terms of residual variance. $R^1(\tau)$ measures the relative success of the corresponding quantile regression models at a specific quantile in terms of an appropriately weighted sum of absolute residuals. Thus $R^1(\tau)$ constitutes a local measure of goodness of fit for a particular quantile rather than a global measure of goodness of fit over the entire conditional distribution, like R^2.

4.4 Simulation approaches

4.4.1 Simulation steps

To estimate the sampling distribution of $\hat{\theta}$, the sample correlation between marker/outcome association A and marginal properties M under the null hypothesis that $\theta = 0$, we could provide a simulation-based approach to see if there is any bias of $\hat{\theta}$ to θ and the variability of $\hat{\theta}$. The following are the simulation steps.

1. Generate outcome Y_i, $i = 1, \cdots, n$ from standard normal distribution and Fisher transformation of marker/outcome association Z_j, $j = 1, \cdots, p$ from a normal distribution with mean μ_z and standard deviation σ_z, and

$$Z_j = \frac{\sqrt{n-3}}{2} log(\frac{1+A_j}{1-A_j}).$$

   Then the marker/outcome association

$$A_j = \frac{\exp 2Z_j/\sqrt{n-3} - 1}{\exp 2Z_j/\sqrt{n-3} + 1}.$$

2. Generate covariate X_j which has Pearson correlation A_j with Y by using

$$X_j = A_j \times Y + \sqrt{1 - A_j^2}\epsilon_j, \tag{4.1}$$

   where ϵ has mean 0 and variance 1. To make covariates X to be correlated with each other, the covariance matrix of ϵ, Σ_ϵ has the form that it has k diagonal blocks with equal size p/k and it is compound symmetric structure in each block with parameter a. Then

$$\epsilon_{ij} = a \times U_i + \sqrt{1 - a^2}\eta_{ij},\ i = 1, \cdots, k, j = 1, \cdots, p/k,$$

where $U_i \sim N(0,1), \eta_{ij} \sim N(0,I)$. Then $cor(\epsilon_{ij}, \epsilon_{il}) = a^2$, $cor(\epsilon_{ij}, \epsilon_{ql}) = 0, i \neq q$. Then the correlation of each X_i, X_j pair would be $\rho_{ij} = A_i A_j + \sqrt{1-A_i^2}\sqrt{1-A_j^2}a^2$ if i, j in the diagonal blocks and $\rho_{ij} = A_i A_j$ otherwise. if $a = 0$, $\epsilon \sim N(0, I)$.

3. Generate marginal mean property $M^1 \sim N(\mu_1, \sigma_1)$, SD property $M^2 \sim N(\mu_2, \sigma_2)$, and skew property $M^3 \sim N(\mu_3, \sigma_3)$. To make the covariate X to have these properties, first we need η_{ij} which is used to construct ϵ_{ij} to be skewed and standardized, let η_{ij} follows Gamma distribution with shape s and scale 1, then the skewness of η_{ij} equals $2/\sqrt{s}$ and the skewness of X_j equals $2(\sqrt{1-A_j^2}\sqrt{1-a^2})^3/\sqrt{s}$ which should equals M_j^3. Then $s = 4(1-A_j^2)^3(1-a^2)^3/(M_j^3)^2$. After X_j is constructed using formula

$$X_j = A_j \times Y + \sqrt{1-A_j^2}\epsilon_j$$
$$\epsilon_{ij} = a \times U_i + \sqrt{1-a^2}\eta_{ij},\ i = 1, \cdots, k, j = 1, \cdots, p/k,$$

scale X_j by M_j^2 and linear transform X_j by M_j^1, then covariate X_j will have desired marginal properties M^1, M^2, M^3.

4. Sample marker/outcome association $\hat{A}_j = \widehat{Cor}(Y, X_j)$ is calculated and $\hat{M}_j$ which is the sample marginal property of X_j is calculated too. Then the sample correlation $\hat{\theta}$ between $\hat{A}_j$ and $\hat{M}_j$ is calculated. We are interested in the change of mean and standard error of $\hat{\theta}$ with different standard deviation of fisher transformation of marker/outcome correlation σ_z. Also other aspects that could affect the amount of change of standard error of $\hat{\theta}$ caused by σ_z, like number of subjects n, number of variables p, correlations within covariate X and parameters of marginal properties.

5. Also we would like to compare this simulation result with permutation re-

sult. There are only two steps different with simulation. One is that the marker/outcome associations A_j and the marginal properties M_j has some correlations, so the Fisher transform Z-score is generated by using

$$Z = r_0 \mathrm{M} + \sqrt{1 - r_0^2}\lambda,$$

where M is the marginal property, λ follows a standard normal distribution and is independent with M, r_0 is the correlation between Z and M. The other step is that when X and Y are generated, permute Y with replacement while make X fixed, which give the assumption that population property/marker/outcome correlation $\theta = 0$. Then the sample correlation $\hat{\theta}$ between sample marker/outcome association $\hat{A}_j$ and marginal mean $\hat{M}_j$ is calculated, and also we are interested in change of mean and standard error of $\hat{\theta}$ with different standard deviation of fisher transformation of marker/outcome correlation σ_z.

4.4.2 Simulation results for $SD(\hat{\theta})$

Overall the expectation of $\hat{\theta}$ always equals to θ when $\theta = 0$ which is consistent with the theoretical derivation of $E(\hat{\theta})$. The standard error of $\hat{\theta}$ will increase when the standard deviation of marker/outcome association, σ_z increase. Furthermore, the amount of increase of standard error of $\hat{\theta}$ caused by σ_z is affected by some other attributes.

First we look at the standard deviation of mean property, σ_1 when holding other attributes fixed. From figure 4.7, we know that the amount of increase of standard error of $\hat{\theta}$ between $\hat{A}_j$ and mean property M^1 caused by σ_z will decrease with increasing σ_1, and the amount of increase of standard error of $\hat{\theta}$ between $\hat{A}_j$ and skew property M^3 caused by σ_z will become constant with increasing σ_1, while the standard error of $\hat{\theta}$ between $\hat{A}_j$ and SD property M^2

will not change with different σ_z and σ_1.

Secondly, we look at the mean of SD property, μ_2 when holding the other attributes fixed. From figure 4.8, we know that the amount of increase of standard error of $\hat{\theta}$ between $\hat{A}_j$ and mean property M^1 caused by σ_z will increase with increasing μ_2, and the amount of increase of standard error of $\hat{\theta}$ between $\hat{A}_j$ and skew property M^3 caused by σ_z will become constant with increasing μ_2, while the standard error of $\hat{\theta}$ between $\hat{A}_j$ and SD property M^2 will not change with different σ_z and μ_2.

Actually we could see that the SD of mean property and the mean of SD property is just the between variance and within variance of covariate X. So we combine these two attributes to one attribute called "within/between variance V", defined as the mean of the variance of X_j over the variance of mean of X_j. From figure 4.9, we know that the amount of increase of standard error of $\hat{\theta}$ between $\hat{A}_j$ and mean property M^1 caused by σ_z will increase with increasing within/between variance, and the amount of increase of standard error of $\hat{\theta}$ between $\hat{A}_j$ and skew property M^3 caused by σ_z will become constant with increasing within/between variance, while the standard error of $\hat{\theta}$ between $\hat{A}_j$ and SD property M^2 will not change with different σ_z and within/between variance.

Next, we look at the correlation within covariate X, the average ρ_{ij}^2, where ρ_{ij} represent the Pearson correlation between X_i and X_j, will be determined by the number of diagonal blocks k and the parameter a. From figure 4.10, we could see that the value of the correlation between X_i, X_j pairs will not change the amount of increase of standard error of $\hat{\theta}$ caused by σ_z.

Then, we look at how number of subjects n and number of variables p influence the amount of increase of standard deviation of $\hat{\theta}$. From figure 4.11, we know

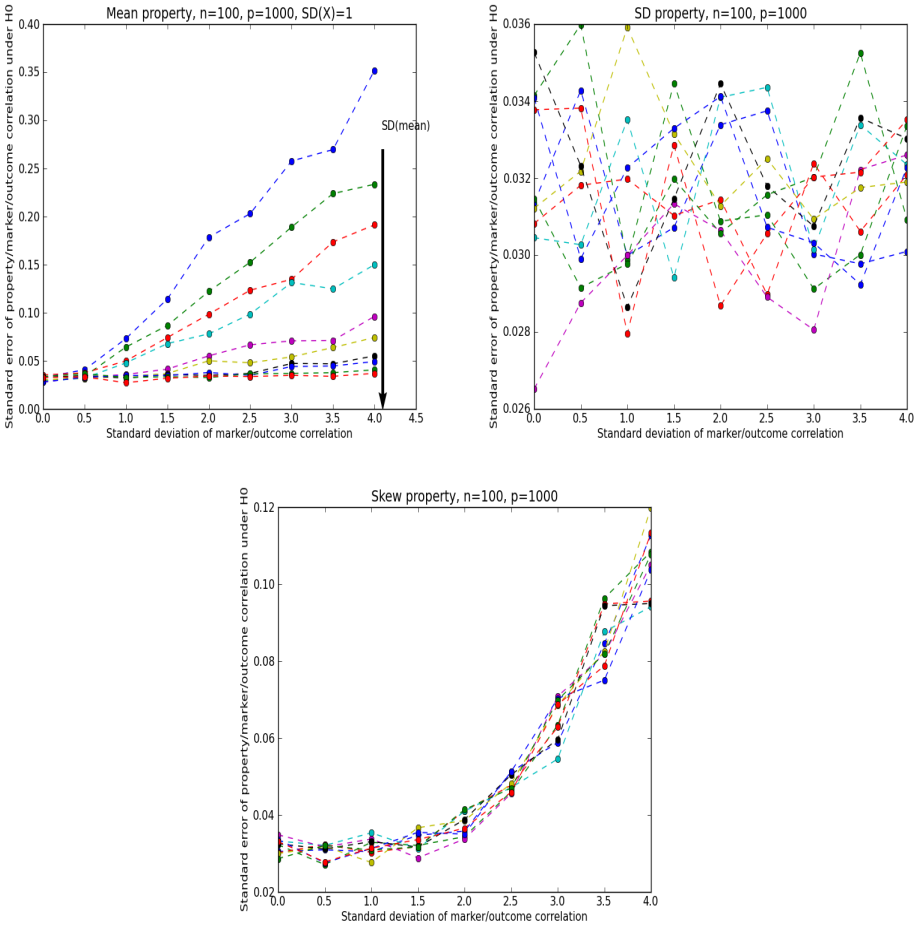

Figure 4.7: Plots of standard error of $\hat{\theta}$ and standard deviation of fisher transformation of marker/outcome association with different level of standard deviation of mean property. In the left plot, $\hat{\theta}$ is the correlation between marker/outcome association and mean property ; In the middle plot, $\hat{\theta}$ is the correlation between marker/outcome association and SD property; in the right plot, $\hat{\theta}$ is the correlation between marker/outcome association and skewness property. n=100, p=1000, a=0, $\mu_1 = 0$, $\mu_2 = 1$, $\sigma_2 = 0$, $\mu_3 = 0$, $\sigma_3 = 0$.

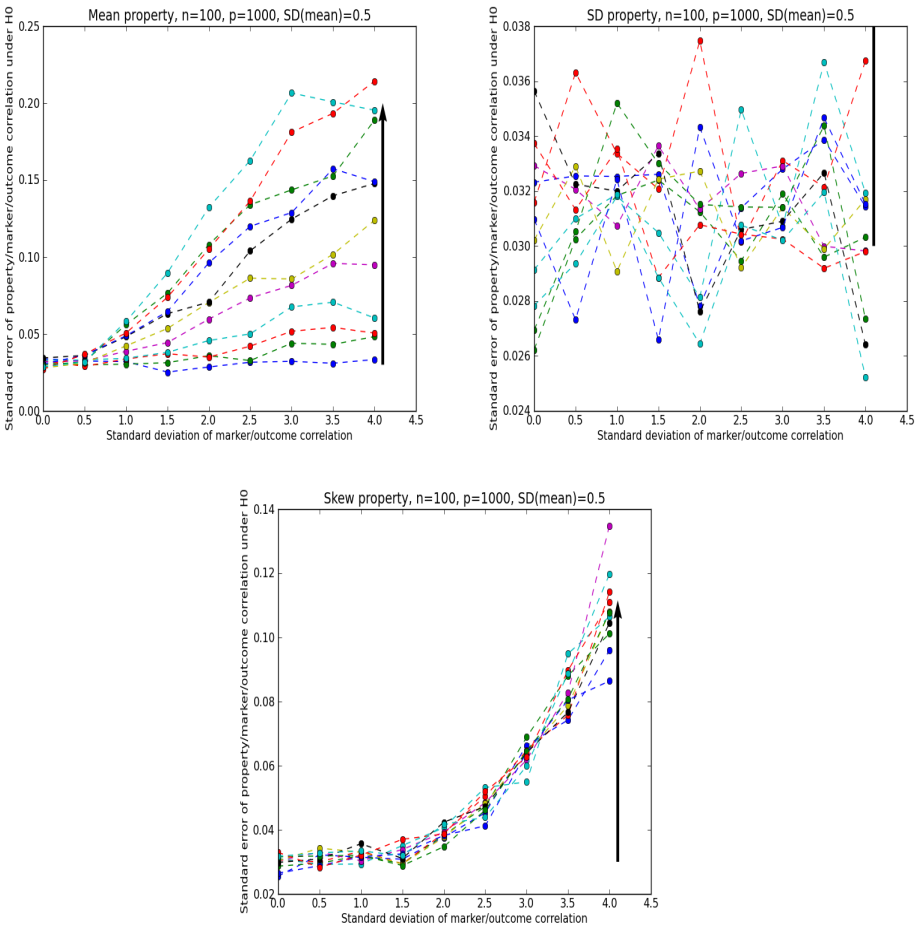

Figure 4.8: Plots of standard error of $\hat{\theta}$ and standard deviation of fisher transformation of marker/outcome association with different level of mean of SD property.In the left plot, $\hat{\theta}$ is the correlation between marker/outcome association and mean property; in the middle plot, $\hat{\theta}$ is the correlation between marker/outcome association and SD property; in the right plot, $\hat{\theta}$ is the correlation between marker/outcome association and skewness property. n=100, p=1000, a=0, $\mu_1 = 0$, $\sigma_1 = 0.5$, $\sigma_2 = 0$, $\mu_3 = 0$, $\sigma_3 = 0$.

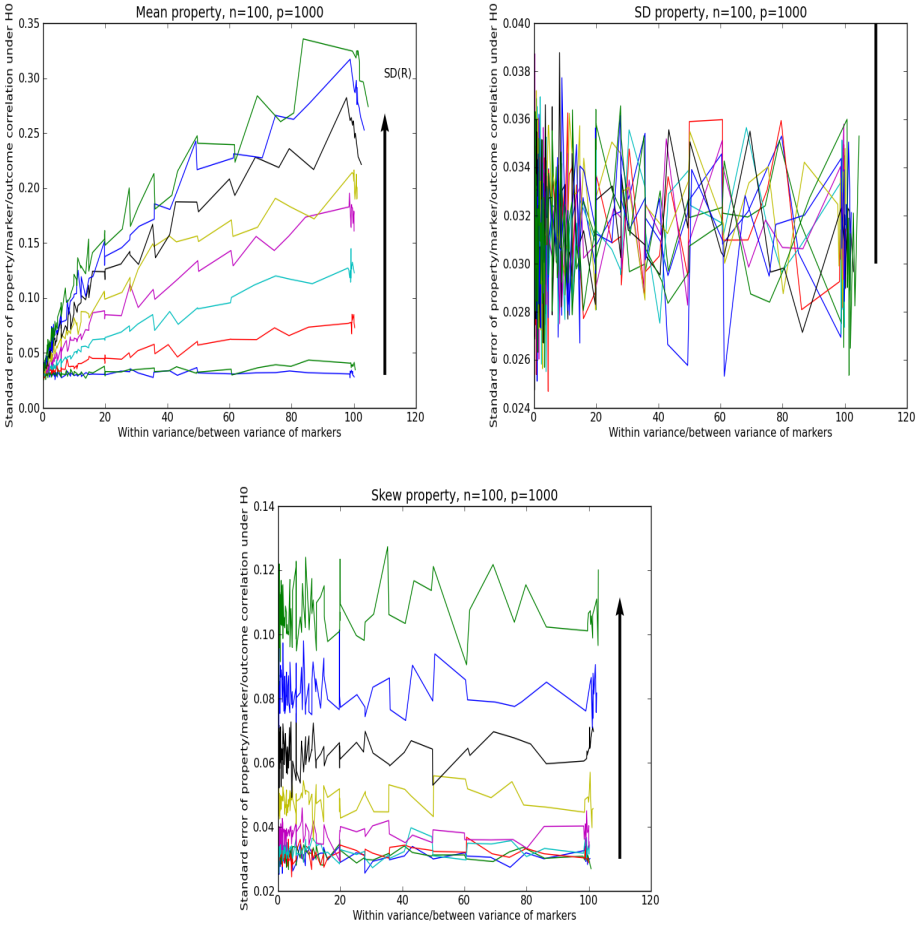

Figure 4.9: Plots of standard error of $\hat{\theta}$ and standard deviation of fisher transformation of marker/outcome association with different level of within/between variance. In the left plot, $\hat{\theta}$ is the correlation between marker/outcome association and mean property; in the middle plot $\hat{\theta}$ is the correlation between marker/outcome association and SD property; in the right plot, $\hat{\theta}$ is the correlation between marker/outcome association and skewness property. n=100, p=1000, a=0, $\mu_1 = 0$, $\sigma_2 = 0$, $\mu_3 = 0$, $\sigma_3 = 0$.

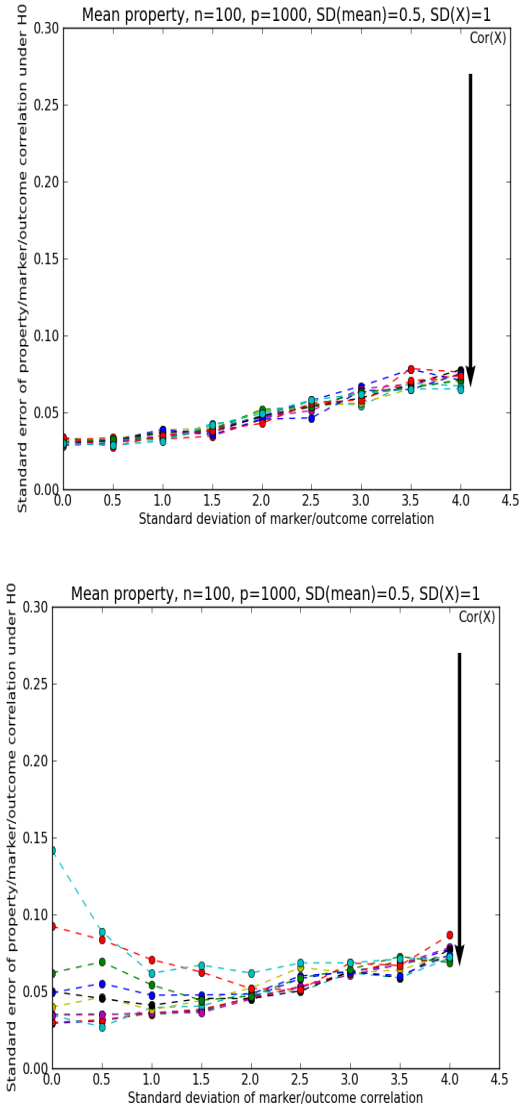

Figure 4.10: Plots of standard error of $\hat{\theta}$ and standard deviation of fisher transformation of marker/outcome association with different level of within correlation of covariate X. $\hat{\theta}$ is the correlation between marker/outcome association and mean property, number of diagonal blocks k=1 in the left plot and k=2 in the right plot. n=100, p=1000, $\mu_1 = 0$, $\sigma_1 = 0.5$, $\mu_2 = 1$, $\sigma_2 = 0$, $\mu_3 = 0$, $\sigma_3 = 0$.

that the amount of increase of standard error of $\hat{\theta}$ between $\hat{A}_j$ and mean property M^1 caused by σ_z will decrease with increasing number of subjects n, so does skew property M^3, while the standard error of $\hat{\theta}$ between $\hat{A}_j$ and SD property M^2 will not change with different σ_z and number of subjects n.

From figure 4.11, we know that the sampling errors of the correlation between marker/outcome association and marginal property follows a normal distribution with mean 0 and standard deviations $1/\sqrt{p}$, so the standard deviation of $\hat{\theta}$ will always increase with decreasing number of variables p. So the standard error of $\hat{\theta}$ between $\hat{A}_j$ and SD property M^2 is just due to the sampling error while the the standard error of $\hat{\theta}$ between $\hat{A}_j$ and mean property M^1 is not just caused by the sampling error, it is also caused by the standard deviation of marker/outcome association σ_z. When σ_z is small, the standard deviation of $\hat{\theta}$ is mainly due to the sampling error. Eith increasing σ_z, the standard deviation of $\hat{\theta}$ will increase, and the amount of increase will increase when p increase. In the end, when σ_z is extremely large, the standard deviation of $\hat{(\theta)}$ will converge no matter the number of p is. And the property of the standard deviation of $\hat{\theta}$ between $\hat{A}_j$ and Skew property M^3 is somewhere between SD property and Mean property.

Overall in the simulation study, the standard deviation of $\hat{\theta}$ between marker/outcome association A and Mean property M^1 will increase when the standard deviation of marker/outcome association, σ_z increase. The amount of increase will be enhanced when the number of subjects n decrease, the number of variables p increase and the within/between variance V of covariate X increase. The standard deviation of $\hat{\theta}$ between marker/outcome association $\hat{A}_j$ and Skew property M^3 will increase when the standard deviation of marker/outcome association, σ_z increase. But the amount of increase will be not affected by the within/between variance V of covariate X, the other patten is the same with Mean property. At

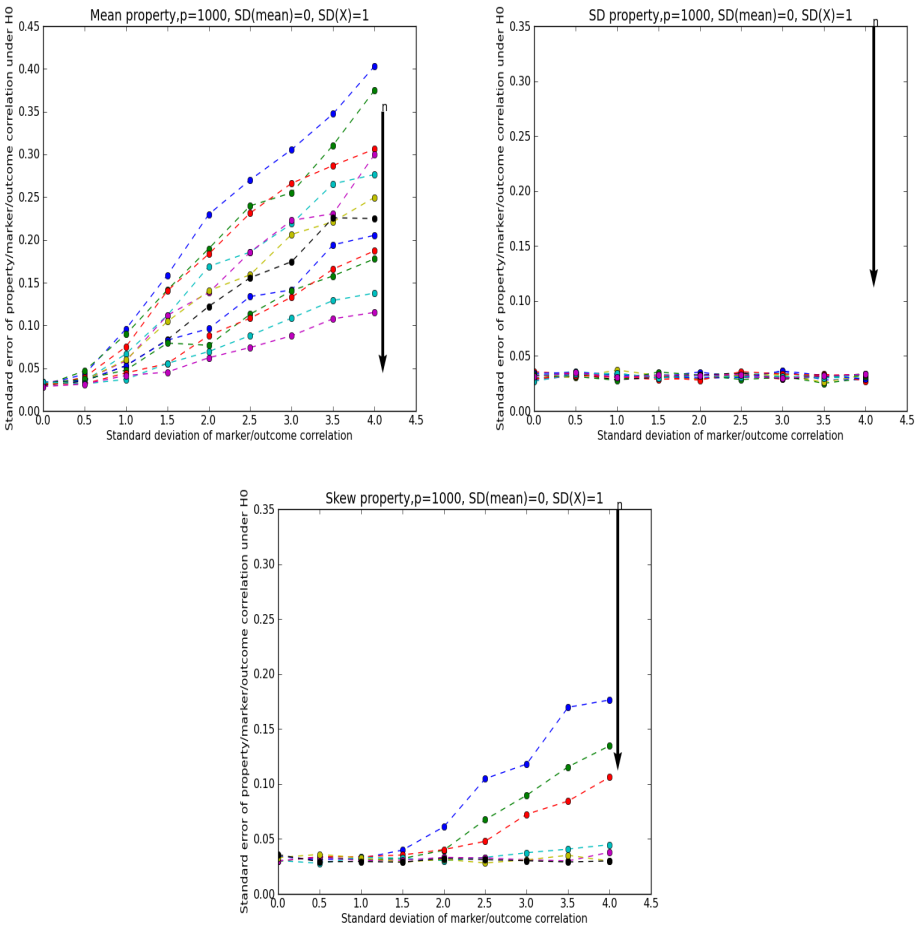

Figure 4.11: Plots of standard error of $\hat{\theta}$ and standard deviation of fisher transformation of marker/outcome association with different number of subjects n. In the left plot, $\hat{\theta}$ is the correlation between marker/outcome association and mean property; in the middle plot, $\hat{\theta}$ is the correlation between marker/outcome association and SD property; in the right plot, $\hat{\theta}$ is the correlation between marker/outcome association and skewness property. p=1000, a=0, $\mu_1 = 0$, $\sigma_1 = 0.5$, $\mu_2 = 1$, $\sigma_2 = 0$, $\mu_3 = 0$, $\sigma_3 = 0$.

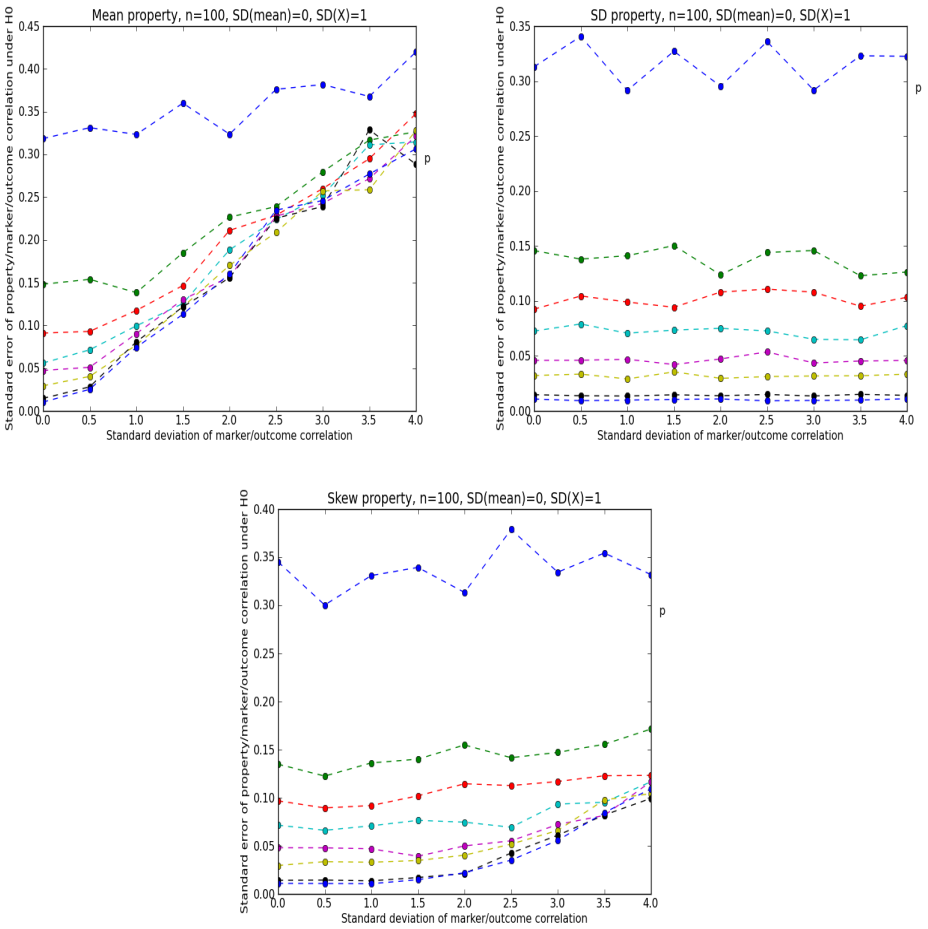

Figure 4.12: Plots of standard error of $\hat{\theta}$ and standard deviation of fisher transformation of marker/outcome association with different number of variables p. In the left plot, $\hat{\theta}$ is the correlation between marker/outcome association and mean property; in the middle plot, $\hat{\theta}$ is the correlation between marker/outcome association and SD property; in the right plot, $\hat{\theta}$ is the correlation between marker/outcome association and skewness property. n=100, a=0, $\mu_1 = 0$, $\sigma_1 = 0.5$, $\mu_2 = 1$, $\sigma_2 = 0$, $\mu_3 = 0$, $\sigma_3 = 0$.

last, the standard deviation of $\hat{\theta}$ between marker/outcome association $\hat{A}_j$ and SD property M^2 will be not affected by σ_z, it is only due to the sampling error of the correlation of two vectors.

4.4.3 Simulation results for $SD(\hat{\theta})$ under permutation analysis

In permutation analysis of the simulated data set, we permute Y while keeping covariate X fixed, then the marker/outcome association has mean 0 and standard deviation $\sigma_z = 0$ and also the population correlation between marker/outcome association and marginal property θ equals 0. This is just the case in simulation study when $\sigma_z = 0$, known that the standard error of $\hat{\theta}$ will increase when σ_z increase, then the standard error of $\hat{\theta}$ calculated by permutation is smaller than the case when σ_z is not zero, which is always true in real case. Then we will be more likely to reject the null hypothesis that $\theta = 0$ and conclude that the marginal property has some association with the marker/outcome association.

But we should realize that in real data set, there is always some sample correlation $\hat{\theta}$ could be detected between marker/outcome association and marginal property, which is not 0, or we don't need to do the permutation analysis to construct the standard error of $\hat{\theta}$ under null hypothesis and test for significance. So we need to add one step before simulation step 1 that the marker/outcome association A_j should be correlated with the marginal property M_j at first. Though by permutation analysis, the expected correlation $E(\hat{\theta})$ between A_j and M_j is forced to be 0, there is always more standard errors of $\hat{\theta}$ when the correlation between $\hat{A}_j$ and $\hat{M}_j$ is large in real case.

From figure 4.13, we see that the the amount of increase of the standard error of $\hat{\theta}$ between marker/outcome association $\hat{A}_j$ and marginal property $\hat{M}$ caused by σ_z will increase when the absolute value of the real correlation between

marker/outcome association $\hat{A}_j$ and marginal property $\hat{M}$ is increased.

4.4.4 Simulation results of $SD(\hat{\theta})$ using the property of the CKD data

As shown above, the standard deviation of the estimated property/marker/outcome associations $\hat{\theta}$ is affected by many factors of the data generating models, like the standard deviation of the marker/outcome associations σ_z, the within/between variance of covariate V, the correlation structure of the gene pairs and so on. So it is hard to decide whether the standard deviation of the estimated property/marker/outcome associations $\hat{\theta}$ is underestimated or overestimated under the permutation technique that researchers usually use to detect the significance of the property/marker/outcome associations.

Here we tried to match the factors of the simulated data to the real CKD data and then compare the standard deviation of $\hat{\theta}$ under permutation analysis with the true simulation analysis. The procedure is similar with the simulation steps in 4.4.1. We use $n = 195, p = 12000$, the marginal mean property, SD property and skew property is calculated from the CKD data. For a grid of τ from -1 to 1, assume that the marker/outcome associations A_j is correlated with marginal mean property with correlation τ and follows a normal distribution with mean 0 and standard deviation 0.2. The covariance matrix Σ_x has 5 diagonal blocks with equal sample size 2400, and in each block there is a compound symmetric structure with correlation parameter $a = 0.7$ to make the the average squared correlation of gene pairs X_i, X_j equals 0.04 which is consistent with the CKD data. Then using the approaches in simulation steps 3, we make the covariate X_j has the same marginal properties.

Now the simulated data (X, Y) is generated, we could calculate the sample marker/outcome association $\hat{A}_j$, the sample marginal properties $\hat{M}$, and then the estimated property/marker/outcome association $\hat{\theta}$ is calculated. For permutation analysis, we need to permute our outcome Y while holding covaraites X fixed, then calculate the esti-

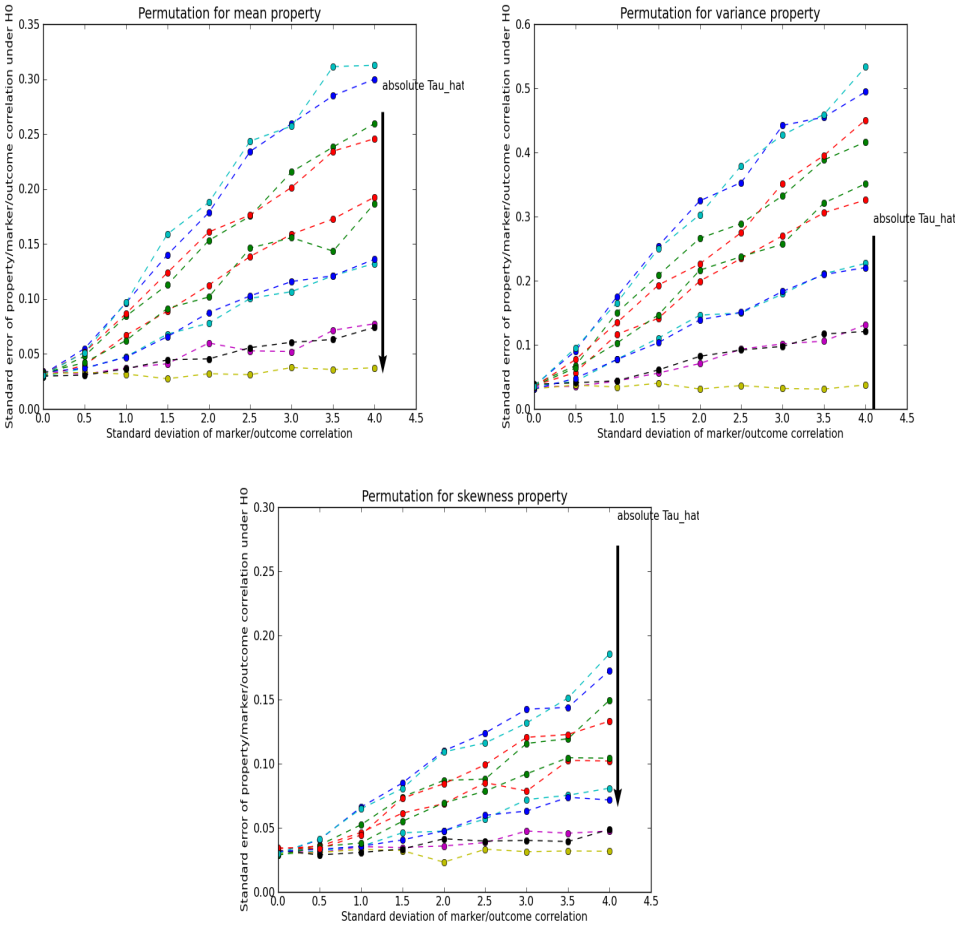

Figure 4.13: Plots of standard error of $\hat{\theta}$ and standard deviation of fisher transformation of marker/outcome association with different levels of the absolute value of the real $\hat{\theta}$ in permutation analysis. In the left plot, $\hat{\theta}$ is the correlation between marker/outcome association and mean property; in the middle plot, $\hat{\theta}$ is the correlation between marker/outcome association and SD property; in the right plot, $\hat{\theta}$ is the correlation between marker/outcome association and skewness property. n=100, p=1000, a=0, $\mu_1 = 0$, $\sigma_1 = 0.5$, $\mu_2 = 1$, $\sigma_2 = 0$, $\mu_3 = 0$, $\sigma_3 = 0$.

mated property/marker/outcome association $\hat{\theta}$. Figure 4.14 compares the standard deviation of $\hat{\theta}$ for simulation and permutation analysis, and shows that the permutation technique will overestimate the standard deviation of $\hat{\theta}$ in CKD data set, then lead to unsignificant result while it is more likely to be significant in truth.

In the approaches above, we try to match the average squared correlation of gene pairs X_i, X_j in simulated data to the real data, and assume that it is good measure of the dependency between covariates X. But it may not capture the most property of the dependence structure of covariates X. Then we illustrate another method to capture the dependence structure of covariates X. We calculate the residuals of $X_j|Y$ by regressing Y on each gene marker X_j, then calculate the covariance matrix of the residuals, Σ_r. Let ϵ_j in equation 4.1 has the same covariance structure Σ_r. The remaining procedures are the same with the procedures above. Figure 4.15 compares the standard deviation of $\hat{\theta}$ for simulation and permutation analysis, and shows that the permutation technique will still overestimate the standard deviation of $\hat{\theta}$ in CKD data set. The difference between these two procedures is that the magnitude of overestimate for permutation technique is different and we believe that the second procedure is more close to the real case.

4.5 CKD data Example

In CKD dataset, 12023 genes and 195 subjects are involved and the glomerular filtration rate (GFR) is used as external trait Y. Glomerular filtration rate (GFR) is a test used to check how well the kidneys are working. Specifically, it estimates how much blood passes through the tiny filters in the kidneys, called glomeruli, each minute. Lower GFR means kidney is not working very well represents patients with severer kidney disease. So it is reasonable that GFR is left skewed since there are few people with very bad GFR and the other people with normal GFR which is consistent with the distribution of GFR in figure 4.16.

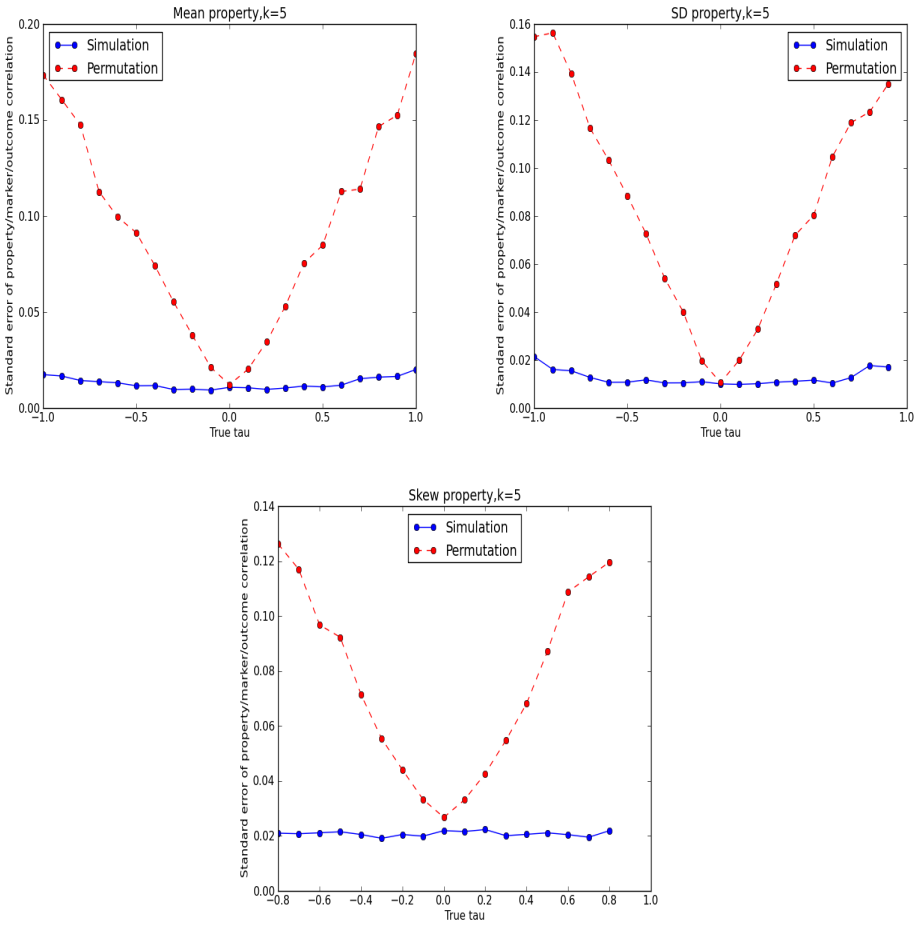

Figure 4.14: Plots of standard error of $\hat{\theta}$ and standard deviation of fisher transformation of marker/outcome association with different levels of real θ in both simulation and permutation analysis for three marginal properties. All the factors of the simulated data are matched to the CKD data and the average squared correlation of gene pairs X_i, X_j is used to represent the covariance structure of X.

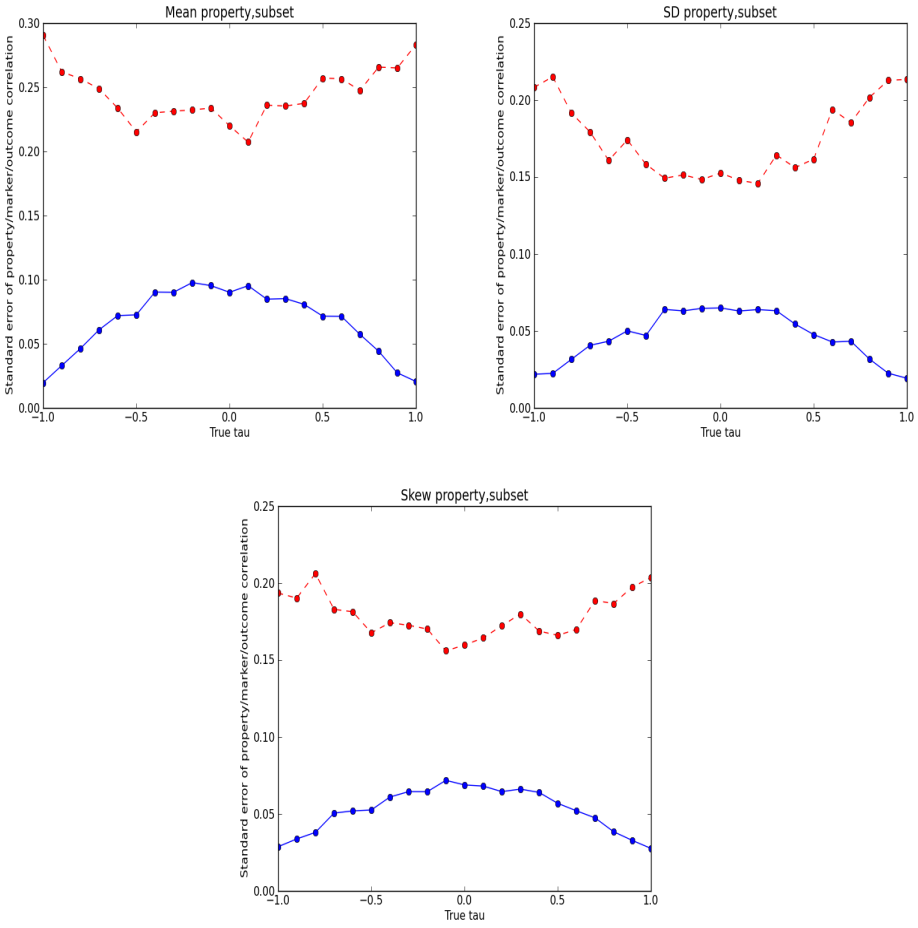

Figure 4.15: Plots of standard error of $\hat{\theta}$ and standard deviation of fisher transformation of marker/outcome association with different levels of real θ in both simulation and permutation analysis for three marginal properties. All the factors of the simulated data are matched to the CKD data and the covariance matrix of the residuals of $X_j|Y$ is used to represent the covariance structure of X. The red line is for permutation analysis, the blue line is for simulation analysis.

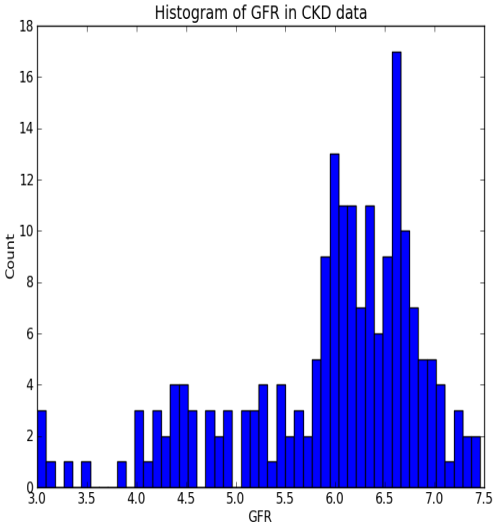

Figure 4.16: Histogram of distribution of GFR in CKD data.

The five marginal features of gene expression across subjects in CKD data are calculated based on the definition on section 4.2. Here we choose $p = 0.75$ for the measure of variance, which is IQR. After each marginal feature is calculated, we construct cubic B-spine with (k=3, df=3) for each measurement and then use them as the covariates and the sample pearson correlations $\hat{r}$ between gene expression and GFR as the outcome in the quantile regression model with quantiles vary from 0.05 to 0.95.

4.5.1 Relations between external correlations and each feature

First, marginal quantile regression of $\hat{r}$ and B-spine of each marginal feature is made and figures 4.17-4.21 are the plots of predicted quantiles vary from 0.05 to 0.95 and each marginal feature and the goodness of fit R_1 is calculated for each regression.

Figure 4.17 shows that skewness has a negative relationship with the quantiles of correlations between gene expression and GFR. The predicted quantiles are parallel, then we could use the center of the predicted quantiles, the median quantile to represent the whole pattern. The average external correlations for genes decrease

linearly when the skewness of the genes increase and the external correlations are usually negative when their corresponding skewness are positive and the external correlations are usually positive when their corresponding skewness are negative. If a gene with expression value skew to the right, it has positive skewness and then leads to negative correlation between GFR and gene expression. Since lower GFR is bad, which means higher gene expression level is bad for this particular gene, then people have higher expression level on this gene or have expression level in the right tail are in poor situation. On the other hand, if a gene with expression value skew to the left, it has negative skewness and then leads to positive correlation between GFR and gene expression. Then people have lower expression level on this gene or in the left tail are in poor situation.

Figure 4.22 give examples of genes in CKD data that are highly skewed and have high correlation between gene expression and GFR. The left scatterplot is for gene which is positively skewed and then have high negative external correlation. The subjects on the right tail have lower GFR, then in bad situation. The right scatterplot is for gene which is negatively skewed and then have high positive external correlation. The subjects on the left tail have lower GFR, then in bad situation. In summary, for genes with non-symmetric distribution across subjects, people have gene expression value in the tail of the distribution are in poor situation. Does the conclusion still true that for genes with symmetric distribution, people have gene expression value in the tail are also in poor situation? We will discuss it later. Also in this CKD dataset, genes are more likely to be right skewed, which is the same with what people expect that there may be more right skewed genes since there is low boundary for gene expression level that gene expression values are always greater than 0. If we choose 0.4 as the threshold, then there are 841 genes have skewness greater than 0.4, while only 218 genes have skewness smaller than -0.4.

Then we look at figure 4.18, the plot of predicted quantiles of external correlations

and IQR, there is a fan pattern of the predicted quantiles, that the tail distribution of external correlations is more spread with the larger IQR. Then genes with larger IQR tend to be more correlated with GFR, either negative or positive but the tend is weaken when gene's IQR are relatively large. If one gene is positively correlated with GFR, then people have low expression level on this gene tends to have low GFR, then in poor situation, on the other hand, if one gene is negatively correlated with GFR, then people have high expression level on this gene are in poor situation. But if one gene does not have much linear relationship with GFR (low correlation with GFR), it could still be interesting, since there could be a symmetric concave/convex relationship between this gene and GFR, then people have high expression level and low expression level are both in poor or good situation.

Figure 4.23 gives examples of genes in CKD data that are symmetric distributed across subjects and have strong linear relationship with GFR and also examples of genes in CKD data that have symmetric concave/convex relationship with GFR. Then the question is, for genes with symmetric distribution, people have gene expression level in the tail are more likely to be both in poor situation or just one tail is bad, the other is good. In other words, are genes more likely to have strong linear trend with GFR or strong symmetric convex/concave relation with GFR?

From figure 4.19, the overall pattern of mean and external correlation is that for genes with lower expression levels, genes are more likely to be highly negatively correlated with the GFR than highly positively correlated, but when expression level goes up, the pattern disappears, there are equally highly positively correlated genes and highly negatively correlated genes and overall, genes with high mean expression level are more likely to be highly correlated with GFR. Then not only the genes with high mean expression level are of interest, genes with low mean expression level could still be interesting due to our finding. The overall pattern of connectivity and external correlation from figure 4.20 is that genes with higher connectivity measure are more

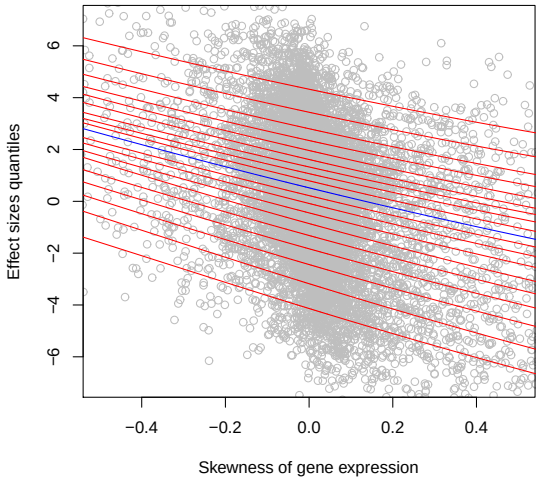

Figure 4.17: Plot of predicted quantiles from 0.05 to 0.95 of external correlations and skewness of gene expression in CKD data.

likely to be highly correlated with GFR and the number of highly positively correlated genes is much higher than the number of highly negatively correlated genes.

At last, we look at the overall pattern of outlier features and external correlations from figure 4.21. There are slightly decreasing trend of external correlations with the increasing of outlier measure, which means for gene with more outliers, it is more likely that the gene is highly negatively correlated with GFR. Since there are more positively skewed genes than negatively skewed, outliers are more likely to be on the right side of the gene, then subjects who are the outliers of one gene are more likely to have high gene expression levels and then in poor situation. The conclusion for outlier measure is quite similar with skewness, and the pearson correlation between outlier and skewness are 0.63 in CKD data. Here comes our next question, will the effect of outlier on the external correlations mainly due to the effect of skewness or in the opposite way that the effect of outlier is dominant? Of all the five marginal features, which is the most dominant feature?

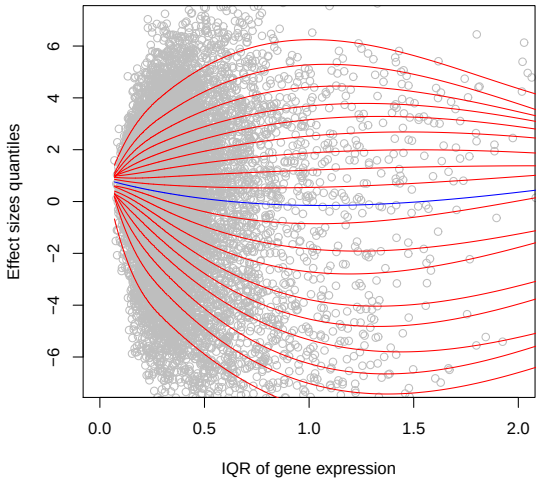

Figure 4.18: Plot of predicted quantiles from 0.05 to 0.95 of external correlations and IQR of gene expression in CKD data

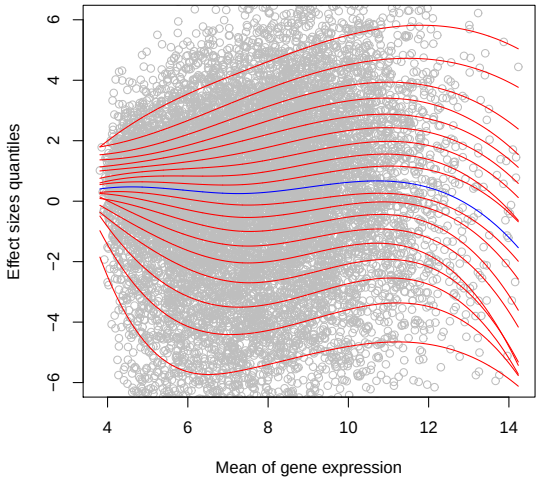

Figure 4.19: Plot of predicted quantiles from 0.05 to 0.95 of external correlations and mean of gene expression in CKD data

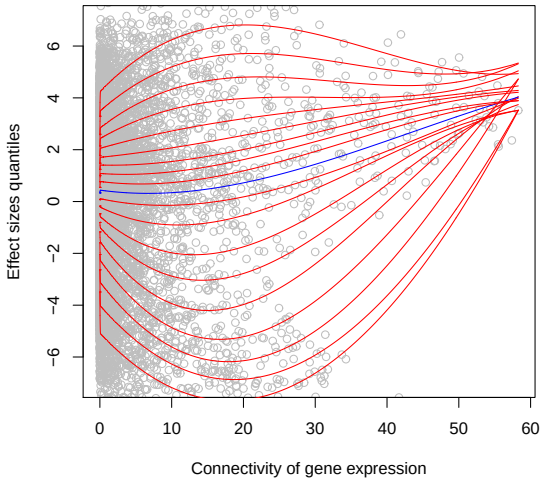

Figure 4.20: Plot of predicted quantiles from 0.05 to 0.95 of external correlations and connectivity of gene expression in CKD data

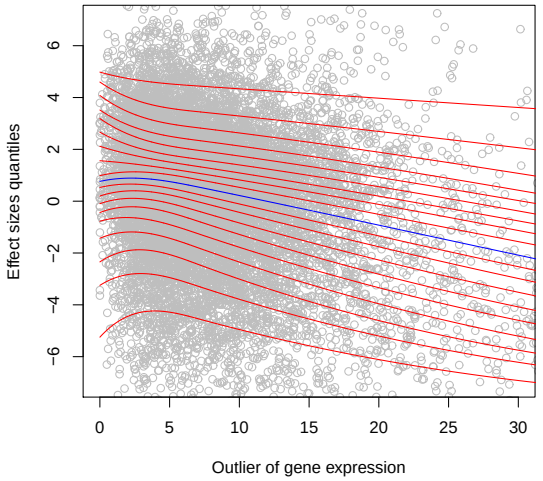

Figure 4.21: Plot of predicted quantiles from 0.05 to 0.95 of external correlations and outlier of gene expression in CKD data

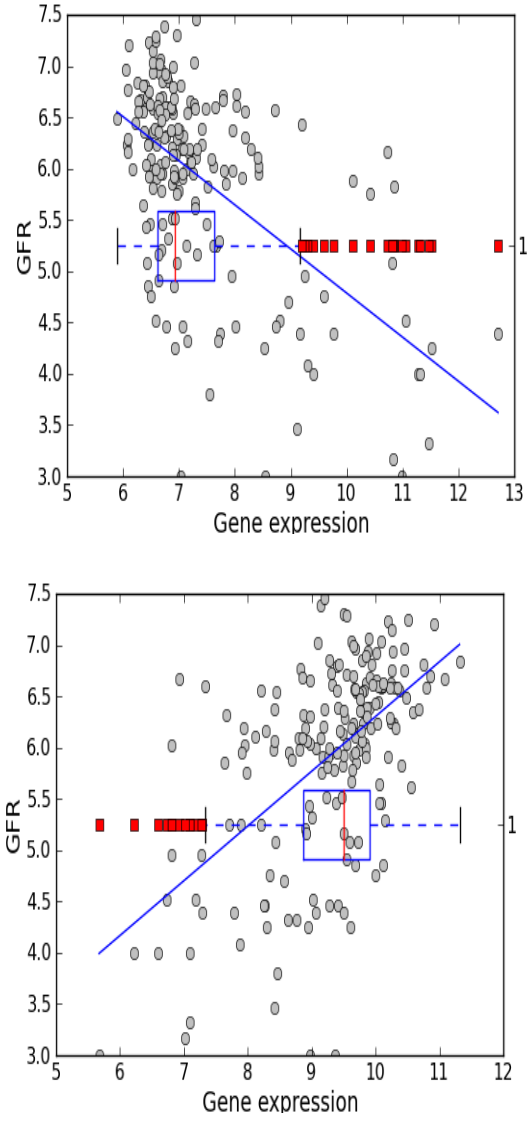

Figure 4.22: Examples of genes in CKD data that are highly skewed and have strong linear relationship with GFR.

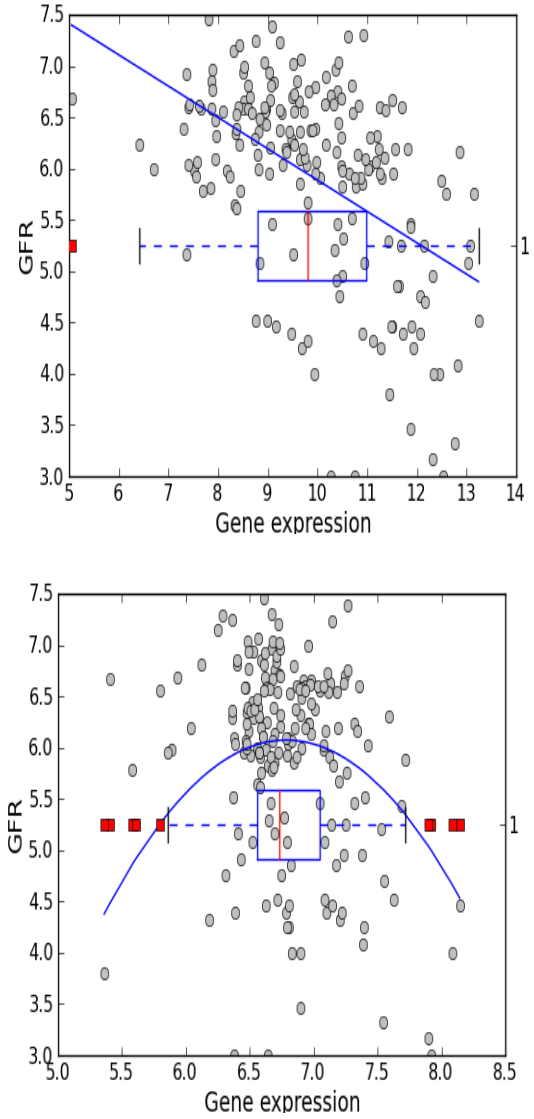

Figure 4.23: Left plot is an example of genes in CKD data that are symmetric and have strong linear relationship with GFR. Right plot is an example of genes in CKD data that are symmetric and have symmetric convex relationship with GFR.

4.5.2 Relations between marginal features

Though the five marginal features of genes are mathematically uncorrelated. For example, genes with high mean could have low variance, genes with high variance could be symmetric or skewed to the left or right. But in real case, there always be some correlations between these five marginal features. Table 4.1 shows the pearson correlation between any two of the marginal features, and the tolerance score which gives the strength of multicollinearity. Tolerance is calculated by using $1 - R_j^2$, where R_j^2 is the coefficient of determination of a regression of explanator j on all the other explanators. A tolerance of less than 0.2 or 0.1 indicates a multicollinearity problem.

Except the CKD data set, we use other three data sets, the skeletal muscle data, psoriasis data and cigarette data. The skeletal muscle data is used to analysis of vastus lateralis muscle biopsies from insulin-sensitive subjects, insulin-resistant subjects and diabetic patients following insulin treatment with 12626 genes and 110 samples where 60 samples are measured before Insulin treatment and 50 samples after Insulin treatment. Then this data could be used as two datasets before and after Insulin treatment. Psoriasis data is used to analyze lesional and non-lsional skins from patients with psoriasis with 54675 genes and 82 samples. We just use 61 samples excluding 21 control samples. Cigarette data analyze the cigarette smoke effect on the oral mucosa with 54675 genes and 79 samples divided into 39 smokers and 40 non-smokers. Then we could still use this data as two datasets with smokers and non-smokers.

From table 4.1, we see that in CKD dataset, there is some positive relationship between mean and variance, and outlier and skewness has the highest correlation 0.63 since higher skewness will lead to higher outlier measure, and also the outliers are more likely to on the right side, not on both sides. We could still see that variance and connectivity have some relationship because if one gene has very low variance, which means that every subject has very similar gene expression value then it should have no correlations with other genes. Since there is no high multicollinearity between our

Table 4.1: Relations among marginal features of gene expression data sets

CKD	IQR	Skewness	Outlier	Connectivity	Tolerance	
n=195	0.28	-0.16	-0.2	0.26	0.84	Mean
		-0.09	0.09	0.34	0.79	IQR
			0.63	0.06	0.58	Skewness
				0.01	0.56	Outlier
					0.84	Connectivity
Skeletal muscle	IQR	Skewness	Outlier	Connectivity	Tolerance	
before treatment	-0.04	0.17	0.25	-0.26	0.85	Mean
n=60		0.07	-0.30	-0.15	0.85	IQR
			-0.72	-0.2	0.35	Skewness
				0.57	0.24	Outlier
					0.52	Connectivity
Skeletal muscle	IQR	Skewness	Outlier	Connectivity	Tolerance	
after treatment	-0.21	-0.19	0.21	0.24	0.84	Mean
n=50		-0.08	-0.12	0.12	0.85	IQR
			0.16	0.01	0.90	Skewness
				0.69	0.46	Outlier
					0.46	Connectivity
Psoriasis	IQR	Skewness	Outlier	Connectivity	Tolerance	
n=61	0.49	-0.06	-0.25	0.34	0.81	Mean
		-0.05	-0.24	0.32	0.85	IQR
			0.63	-0.08	0.70	Skewness
				-0.12	0.79	Outlier
					0.73	Connectivity
Cigarette smokers	IQR	Skewness	Outlier	Connectivity	Tolerance	
n=40	0.25	-0.02	-0.04	0.08	0.99	Mean
		-0.01	-0.13	0.53	0.69	IQR
			0.64	-0.03	0.72	Skewness
				-0.10	0.69	Outlier
					0.71	Connectivity
Cigarette non-smokers	IQR	Skewness	Outlier	Connectivity	Tolerance	
n=39	0.25	-0.02	-0.05	0.10	0.99	Mean
		-0.01	-0.14	0.44	0.78	IQR
			0.48	-0.002	0.78	Skewness
				-0.10	0.76	Outlier
					0.80	Connectivity

5 marginal properties, we could include all of them in the multiple regression model.

The finding is similar in psoriasis data and cigarette data. There is some positive correlation between mean and variance, and outlier and skewness has the highest positive correlation. In cigarette data, there is also some positive relationship between variance and connectivity, which means genes have higher variance are more likely to be highly connected with other genes.

While in skeletal muscle data, the relationships between marginal features might be different. First, mean and variance has some negative relationship, this may cause some problem when you filter the genes with low mean and low variance, then some genes with high mean expression level are also removed, which will lead to losing some important information. Also the high positive correlation between outlier and connectivity which is 0.69 for skeletal muscle data after treatment and 0.57 before treatment means that genes with more outliers tend to be more correlated with other genes. There is high negative correlation between skewness and outliers which is -0.72 for skeletal muscle data before treatment which is just opposite with the relationship we found in CKD dataset, which means genes are more likely to have outliers on the left side or genes are more likely to skewed to the left, then there are some genes have very low expression values.

When we do simple regression, we must be cautious in looking at the effect of one predictor to the outcome. It is commonly accepted that effect of factor A to the outcome is weaken if there is an alternative factor that is related to factor A as well as the outcome. For example, in CKD dataset, there may be trends between external correlations and both variability and skewness. But variability and skewness are also correlated. Then it is not acceptable if we just do marginal simple regression of external correlations on variability or skewness and conclude that genes with certain level of variability or skewness tend to show high external correlations. Statistical control of "confounding" is to include it as a covariate in a quantitative model. So if

all the five marginal features are correlated and each of them has relationship with the outcome, we should include all of the five marginal properties in one multiple regression model and then we could look at the partial R^2 which quantify how much unique information about outcome in one covariate is not captured by the other covariates, then predictor with the largest partial R^2 is considered to be the most important predictor. Another way is controlling for one factor, then look at the relationship between the other factor and outcome. For example, we could control for the effect of skewness by dividing the genes into several groups with different value of absolute skewness, then see if the relationship between variance and external correlations still exists.

4.5.3 The most dominant feature

First, we include all the marginal features in one quantile regression model, the partial R_1 is calculated by using the formula

$$partial R_1 = (R_{1,total} - R_1^\star)/(1 - R_1^\star)$$

where $R_{1,total}$ is the total R_1 when all the features are included in the model and $R_1^\star$ is the R_1 when one specific feature is not included in the model while others are. Then it quantifies how unique information of one specific feature to the external correlations. Figure 4.24 shows the marginal R_1 and $partial R_1$ for our five features, we could see that the average $partial R_1$ across all the quantiles for mean is 0.007, IQR is 0.008, Outlier is 0.015, Skewness is 0.05 and Connectivity is 0.012. The importance of skewness is 3-4 times more than the other marginal features on average of the multiple quantile regression. So the most important predictor in quantile regression is skewness and the second important predictor is outlier, the third is connectivity.

Second, we do the marginal regression of one feature while controlling the value of

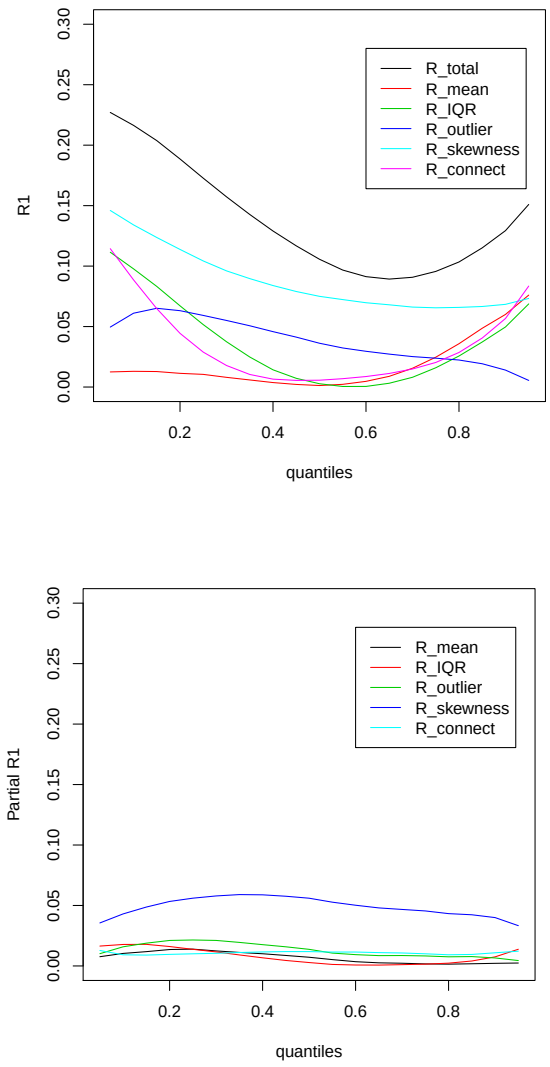

Figure 4.24: Left plot is R1 for quantile regression of external correlations and each feature; Right plot is partial R1 for multiple quantile regression, comparing full model and full model without one feature at each time.

another feature. It is well known that if feature A has strong relation with outcome and also feature A and feature B are highly correlated with each other, then feature B will also have some relation with the outcome, though in fact, feature B is independent with the outcome. Then if we control for the value of feature B, though the domain and density of feature A will change, the effect of feature A to the outcome will not change. On the other hand, if we control for the value of feature A, the effect of feature B to the outcome will disappear due to the true relationship between feature B and outcome is independent. From the results of multiple quantile regression, we know that skewness is the most dominant feature. To check that, let's control for the other features to see if the relationship between skewness and external correlations changes.

We first control for the effect of IQR to see how the relationship between skewness and external correlations changes due to different levels of IQR. We order genes with their IQR values and divide genes into three groups based on their IQR values and then the low IQR group contain the lowest 1/3 genes with IQR smaller than 0.25. Middle IQR group contains the middle 1/3 genes with IQR between 0.25 and 0.38 and high IQR group contains the top 1/3 genes with IQR greater than 0.38. Within each IQR group, we will do the marginal quantile regression on skewness feature. Figure 4.25 shows that the relationship between skewness and external correlations does not change when controlling for the effect of IQR. Then we also control for the effect of mean, outlier and connectivity with the same procedure, from figure 4.26-4.28, we know that the relationship between skewness and external correlations remains the same. These results are consistent with what we find using $partialR_1$.

4.5.4 Function deconvolution result

The second topic for this section is for symmetric genes, are they more likely to be linearly correlated with GFR or have a symmetric concave/convex relationship with

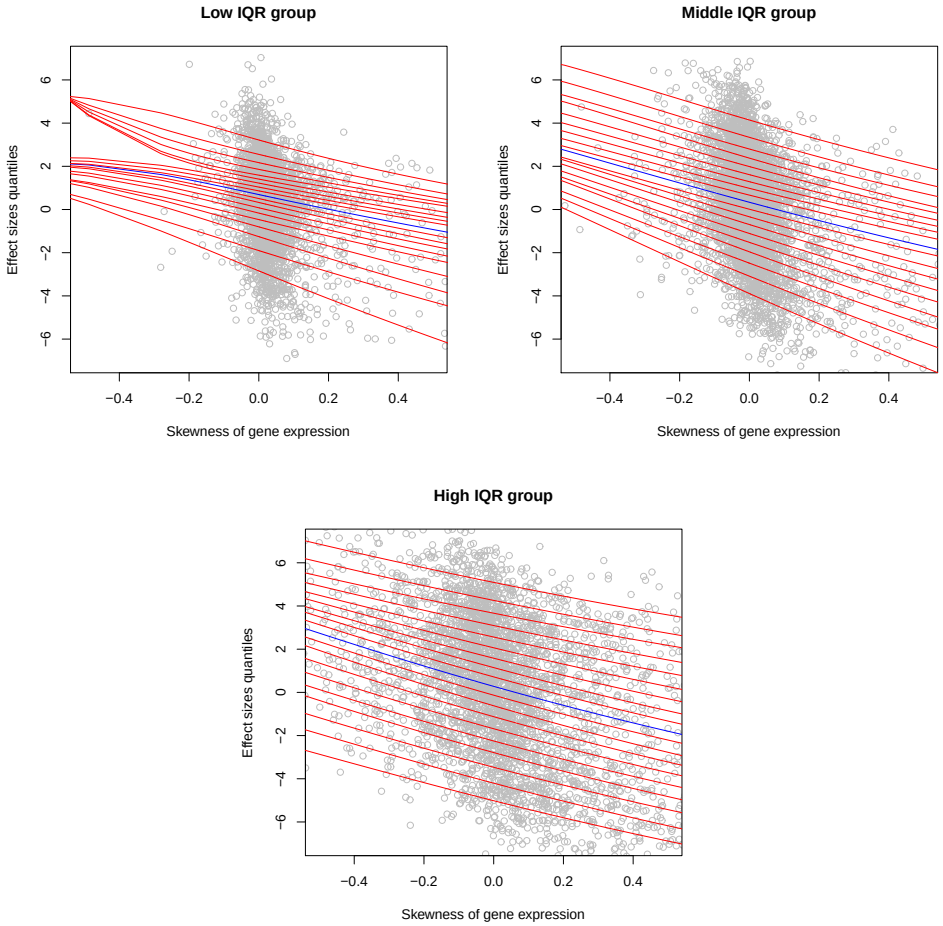

Figure 4.25: Plot of predicted quantiles of external correlations and skewness of gene expression for three gene sets with different level of IQR.

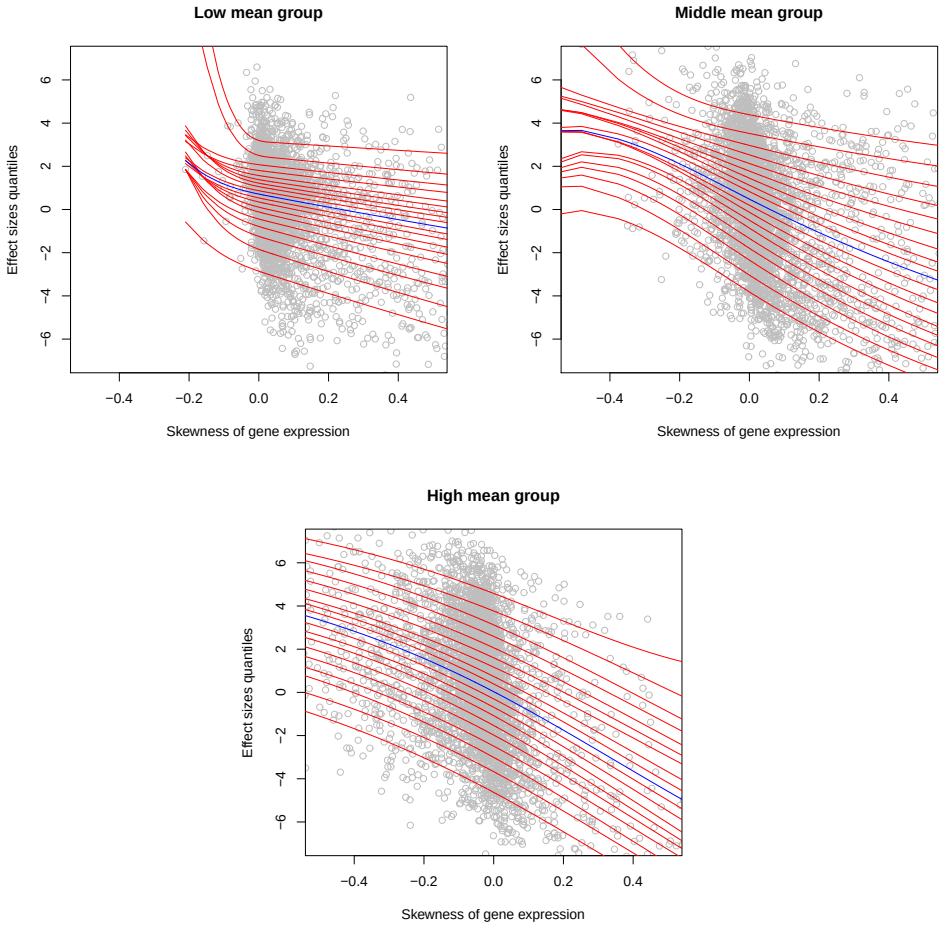

Figure 4.26: Plot of predicted quantiles of external correlations and skewness of gene expression for three gene sets with different level of mean.

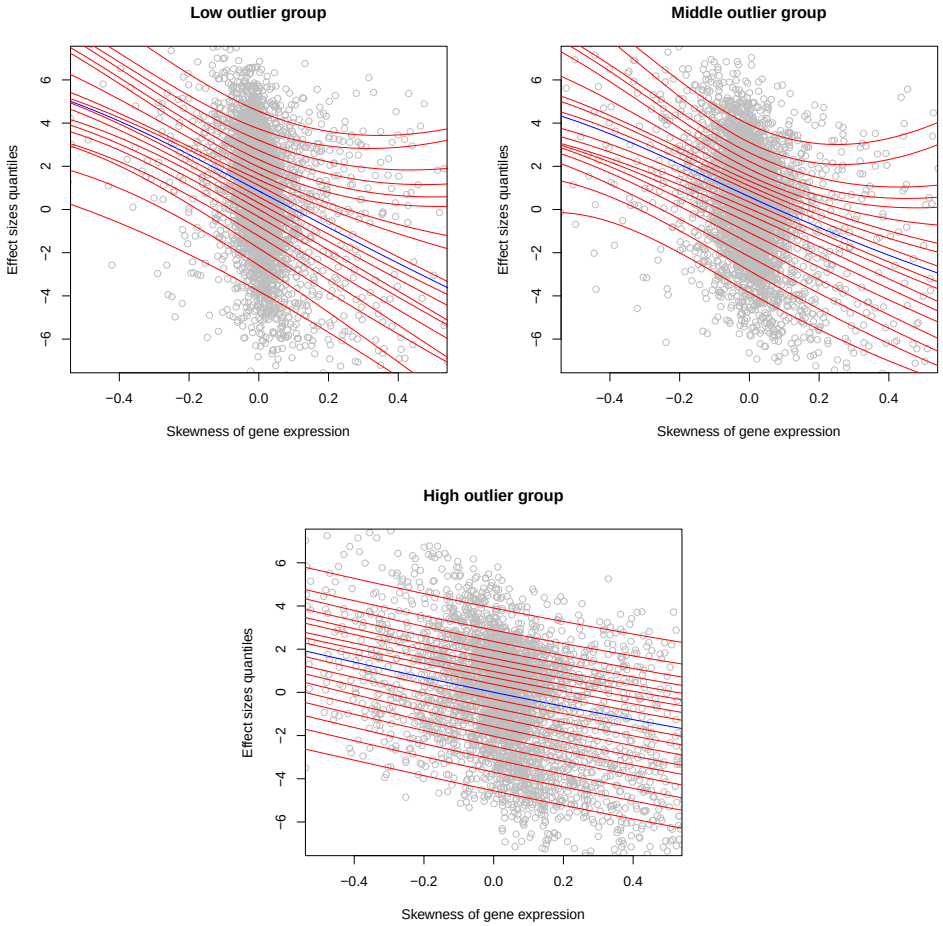

Figure 4.27: Plot of predicted quantiles of external correlations and skewness of gene expression for three gene sets with different level of outlier.

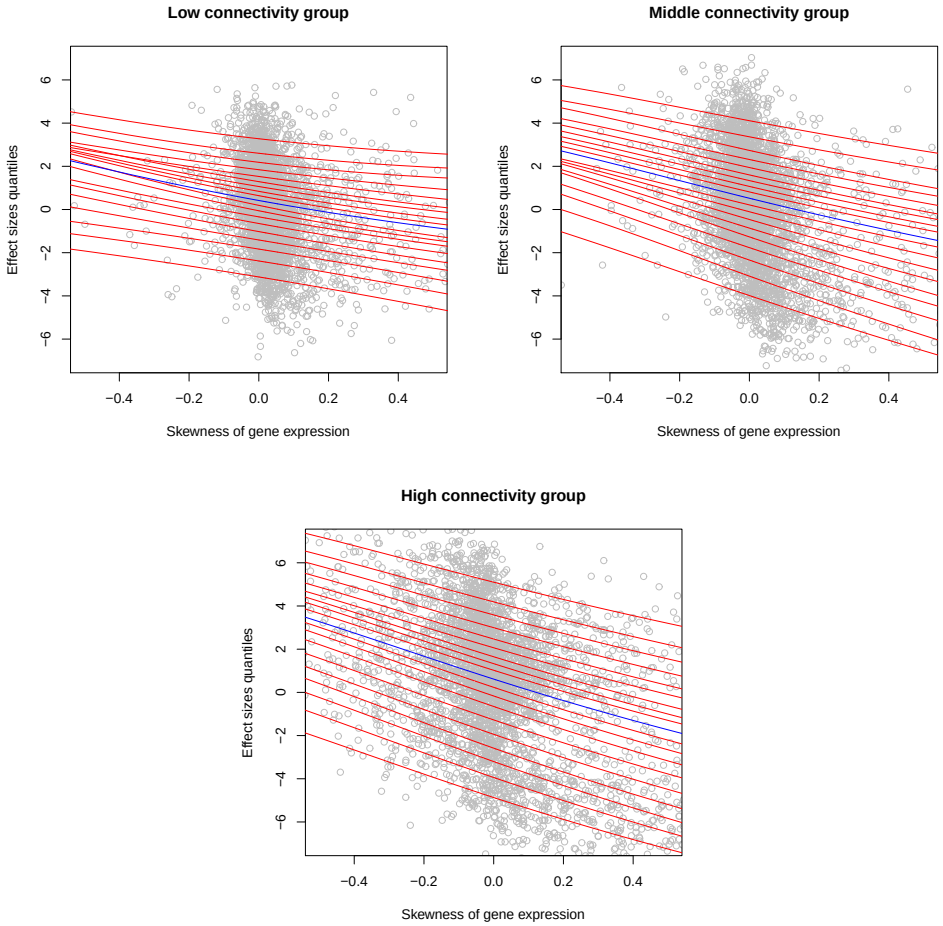

Figure 4.28: Plot of predicted quantiles of external correlations and skewness of gene expression for three gene sets with different level of connectivity.

GFR? Here we deconvolve the function of $E(Y|X)$ into a monotone function and a symmetric convex function and a residual term. Then,

$$E(Y|X) = f_{sc} + f_m + r$$

In our situation, outcome Y is GFR and X is gene expression value for each gene and we use $X - \bar{X}$, $(X - \bar{X})^3$, $sign(X - \bar{X})(X - \bar{X})^2$, $sign(X - \bar{X})\sqrt{X - \bar{X}}$, $\arctan(X - \bar{X})$ as the basis monotone functions and $\sqrt{|X - \bar{X}|}$, $|X - \bar{X}|$, $|X - \bar{X}|^{1.5}$, $|X - \bar{X}|^2, \log(X - \bar{X} - 1)$ as the basis symmetric convex function. Then our regression model becomes:

$$\begin{aligned} Y = \beta_0 + \beta_1(X - \bar{X}) + \beta_2(X - \bar{X})^3 + \beta_3 \arctan(X - \bar{X}) \\ +\beta_4 sign(X - \bar{X})(X - \bar{X})^2 + \beta_5 sign(X - \bar{X})\sqrt{X - \bar{X}} + \beta_6|X - \bar{X}| + \beta_7|X - \bar{X}|^{1.5} \\ +\beta_8|X - \bar{X}|^2 + \beta_9\sqrt{|X - \bar{X}|} + \beta_{10} \log X - \bar{X} - 1 + \epsilon \end{aligned}$$

where $E(\epsilon|X) = 0$, $Var(\epsilon|X) = \sigma^2$. And we need $sign(\beta_1) = sign(\beta_2) = sign(\beta_3) = sign(\beta_4) = sign(\beta_5)$ and $sign(\beta_6) = sign(\beta_7) = sign(\beta_8) = sign(\beta_9) = sign(\beta_{10)}$. We use the nnls package in R program and regress the model in four situations, when the sign of the coefficients of the monotone function is positive/negative and the sign of the coefficients of the symmetric convex function is positive/negative. Then choose the situation of the highest R^2 for each symmetric gene (in the low skewness group). Then partial R^2 for monotone function and partial R^2 for symmetric convex function are calculated for each gene and they are plotted in figure 4.29. Using threshold 0.1, then about 4% of the symmetric genes are demonstrated to be linearly correlated with GFR, while 0.2% genes show strong symmetric convex/concave relation with GFR.

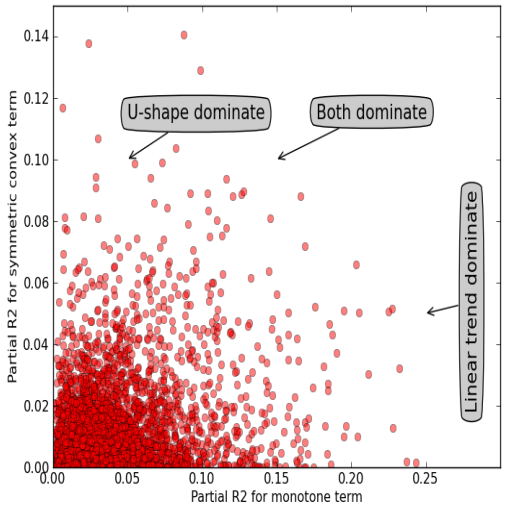

Figure 4.29: Plot of partial R2 of linear term and partial R2 of quadratic term.

4.6 Challenge and conclusion

We focus on a new question that relate the effect sizes (marker/outcome associations) to the properties of marginal distribution of the markers. This new framework for understanding marginal marker/outcome associations in large datasets involves familiar summary statistics such as the Pearson correlation coefficients, but applies it in a non-standard way to derived quantities, rather than directly to the observations. We use the quantile regression with spline basis technique to model the relationships between the marker/outcome associations and the five marginal properties with a L_1 goodness of fit R_1 in real data analysis. We figure out that the skewness property has the strongest association with the marker/outcome association and dominate the other marginal properties.

Then we addresses the challenge of assessing the uncertainty in statistics that are aggregated over large data sets within complex and poorly understand dependencies. We show that commonly used randomization approaches, while intuitive, can give misleading results, and we provide a simulation based alternative approach

that appears to perform well in a variety of situations. Also applied this approaches to the real data, we compare the standard deviation of the statistic of simulation based approach to the randomization approach and find out that the randomization approach overestimates the standard deviation of the statistic and tends to make the property/marker/outcome association unsignificant.

Then we addresses the issue of marker/outcome relationships that are strongly non-monotonic. We propose a decomposition of such relationships into monotonic and "u-shaped" components. We find out that the monotone marker/outcome associations are more likely to exist in CKD data than the "u-shaped" associations, then use Pearson correlation coefficients to represent the marker/outcome associations is reasonable.

There is also a challenge for the regression model, that the external correlations between GFR and gene expression data are themselves correlated since we calculated the correlations with the same outcome GFR, then the dependent variable in regression model is correlated or the error term in the regression model is correlated which violate the assumption for quantile regression that the quantiles of the error terms given covariates are independent. Therefore the estimated coefficients of the regression will be biased.

CHAPTER V

Conclusions

In this thesis, we consider several challenging issues that arise when analyzing genomic data. Difficulties that arise in this area commonly result from the effects of covariate measurement error, complicated dependence structure, and data sets with high dimension and small sample size. Chapter 2 focuses on the statistical assessment of predictive performance due to covariate measurement error. Chapter 3 focuses on proposing a new method to measure the overlap effect sizes in two subpopulations in genomic study. Chapter 4 focuses on a new question that relate the effect sizes (marker/outcome associations) to the properties of marginal distribution of the markers.

In chapter 2, we first demonstrate that the predictive performance is negatively affected by the increase of magnitude of measurement error. The effect is also influenced by other factors of data generating model related to the true regression coefficient β, the covariance matrix of true covariates X, Σ_x, and the covariance matrix of the measurement error, Σ_η. Then we identify these factors from the theoretical derivation of predictive accuracy AUC, we find that there are four factors might influence the decline of predictive accuracy and similar findings is shown for the linear case. Also in the simulation study, we find that $E(S_1)$ has a negative relationship with the decline of $\widetilde{AUC}$, while $\mathrm{Var}(S_2)$, $\mathrm{Var}(S_1)$, $\mathrm{Cov}(S_1, X_1 - X_2)$ have positive relationships with

the decline of $\widetilde{AUC}$.

To apply this to practical use, we propose a SIMEX procedure to estimate these two factors from real data, though the estimate is not very accurate. Then we define a ratio of the decline of predictive accuracy due to measurement error compare to the overall decline of predictive accuracy. If the ratio is large, the effect of measurement error dominate the decline of predictive accuracy, otherwise, we do not need to worry much about the measurement error. This could help researchers to decide whether to improve technologies to measure the data more accurately or to use more advanced regression techniques, find more relevant covariates or collect more samples to reduce other errors causing the decline of predictive accuracy.

In chapter 3, we first define the overlap measure of the effect sizes of the marker/outcome associations in two subpopulations, and then propose parametric and nonparametric methods to estimate it. In simulation study, we compare the accuracy of the estimate of the overlap measures through mle, moment, rescaling, copula and plug-in methods and find out that if the joint distribution of the true standardized parameter (θ_i^A, θ_i^B) is much deviated from bivariate normal distribution, the copula method performs better than the other parametric or nonparametric methods. Then we apply copula-based, mle and plug-in method to estimate the overlap measure of the common associations in each pairs of disease subgroups in CKD data. MLE estimator and copula estimator give similar result for most of the pairs of disease subgroups, implying that the joint distribution of the effect sizes in any two disease subgroups is close to bivariate normal distribution.

In chapter 4, we address four issues. First, we assess the uncertainty of the property/marker/outcome associations that are aggregated over large data sets within complex and poorly understand dependencies. We show that commonly used randomization approaches, while intuitive, can give misleading results, and we provide a simulation based alternative approach that appears to perform well in a variety of

situations. Second, we address the issue of marker/outcome relationships that are strongly non-monotonic. We propose a function decomposition method and find out that the monotone marker/outcome associations are more likely to exist in CKD data than the "u-shaped" associations. Third, We use the quantile regression with spline basis technique to model the relationships between the marker/outcome associations and the five marginal properties with a L_1 goodness of fit R_1 in real data analysis. We figure out that the skewness property has the strongest association with the marker/outcome association and dominate the other marginal properties. At last, the scientific meaning of the property/marker/outcome association is considered. We conclude that for the highly skewed genes, people have gene expression level on the tail distribution of these genes are more likely to be a poor condition of the disease.

BIBLIOGRAPHY

BIBLIOGRAPHY

A, B., B. MT, L. Best, and G. Stoica (2008), Accuracy of effect size estimates from published psychological research, *Perceptual and Motor Skills*, *106*(2), 645–649, doi:10.2466/PMS.106.2.645-649.PMID 18556917.

Carroll, R. J., D. Ruppert, S. L. A., and C. M. Crainiceanu (2006), *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*, Chapman and Hall/CRC, United States of America.

Chen, R., M. Chiu, and C. L. (2004), Detecting Credit Card Fraud by Using Questionaire-Responded Transaction Model Based on Support Vector Machines, *Proc. of IDEA*, pp. 800–806.

Dodd, L. E., and M. S. Pepe (2003), *Partial AUC Estimation and Regression*, UW Biostatistics Working Paper Series, United States of America.

Draghici, S., P. Khatri, A. C. Eklund, and Z. Szallasi (2005), Reliability and reproducibility issues in DNA microarray measurements, *TRENDS in Genetics*.

Efron, B. (2007), Size, power and false discovery rates, *Ann. Statist.*, *35*(4), 1351–1377.

Fuller, W. A. (1987), *Measurement Error Models*, John Wiley & Sons, United States of America.

Gout, J.-F., D. Kahn, and D. L (2010), Filtering for increased power for microarray data analysis, *Paramecium Post-Genomics Consortium*, *6*(5), doi: 10.1371/journal.pgen.1000944.

Hackstadt, A. J., and A. M. Hess (2009), Filtering for increased power for microarray data analysis, *BMC Bioinformatics*, *10*, doi:10.1186/1471-2105-10-11.

Hanley, J. A., and B. J. McNeil (1982), The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve, *Radiology*, pp. 29–36.

He, Z., and J. Zhou (2008), Empirical Evaluation of a New Method for calculating Signal-to-Noise Ratio for Microarray Data Analysis, *Applied and Environmental Microbiology*, *74*(10), doi:10.1128/AEM.02536-07.

Hedges, l. V., and I. Olkin (1985), Statistical Methods for Meta-Analysis, *Orlando: Academic Press*.

Ioannidis, J. P., T. A. Trikalinos, and K. M. J. (2005), Implications of Small Effect sizes of Individual Genetic Variants on the Design and Interpretation of Genetic Association Studies of Complex Diseases, *American Journel of Epidemiology*, *164*(7), doi:10.1093/aje/kwj259.

Kojo, S., A. Tsutsumi, G. D, and S. T (2003), Low expression levels of soluble CD1d gene inpatients with rheumatiod arthritis, *J Rheumatol*, *30*(12).

Larkin, J. E., B. C. Frank, H. Gavras, S. Razvan, and J. Quackenbush (2005), Independence and reproducibility across microarray platforms, *Nature Methods*, *2*, 337–344, doi:10.1038/nmeth757.

Lin, R., S. Dai, and R. D. Irwin (2008), Gene set enrichment analysis for non-monotone association and multiple experimental categories, *BMC Bioinformatics*, *9*(481), doi:10.1186/1471-2105-9-481.

Lyons, W., J. Patel, and M. Becich (2004), Testing for finding complex patterns of differential expression in cancer: towards individuallized medicine, *BMC Bioinformatics*, *5*.

MacDonald, J. W., and D. Ghosh (2006), COPA–cancer outlier profile analysis, *Bioinformatics*, *22*(23), 2950–2951, doi:10.1093/bioinformatics/bt1433.

Mar, J., N. Matigian, and A. Mackay-Sim (2011), Variance of Gene Expression Identifies Alter Network Constraints in Neurological Disease, *PLoS Genet*, *7*(8).

Nakagawa, S., and I. C. Cuthill (2007), Effect size, confidence interval and statistical significance: a practical guide for biologists, *Biological Review Cambridge Philosophical Society*, *82*(4), 591–605, doi:10.1111/j.1469-185X.2007.00027.x.PMID 17944619.

Park, J.-H., M. H. Gail, and W. C. R. (2011), Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants, *PNAS*, *108*(44), 18,026–18,031.

Schreiber, M. B., L. Beiganowski, and K. Andrzej (2002), Digitial Analysis of Ophthalmological Images - a Proposal for a Screening Stategy, *Polish J Med Phys and Eng*, *8*(2), 89–98.

So, H.-C., and P. C. Sham (2010), Effect Size Measures in Genetic Association Studies and Age-Conditional Risk Prediction, *Human Heredity*, *70*, 205–218, doi: 10.1159/000319192.

Tomlins, S., D. Rhodes, and S. Perner (2005), Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer, *Science*, *310*.

Xia, Y. (2006), Asymptotic distribution for two estimators of the single-index model, *Econometric Theory*, *22*, 1112–1137.

Zhang, B., and S. Horvath (2005), A General Framework for Weighted Gene Co-expression Network Analysis, *Statistical Applications in Genetics and Molecular Biology*.