STATISTICAL TOOLS FOR GENERAL ASSOCIATION TESTING AND CONTROL OF FALSE DISCOVERIES IN GROUP TESTING

Pratyaydipta Rudra

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2015

Approved by:

Fred A. Wright

Andrew Nobel

Wei Sun

Yun Li

Karen Mohlke

ABSTRACT

Pratyaydipta Rudra: Statistical Tools for General Association Testing and Control of False Discoveries in Group Testing
(Under the direction of Fred A. Wright and Andrew Nobel)

In modern applications of high-throughput technologies, it is important to identify pairwise associations between variables, and desirable to use methods that are powerful and sensitive to a variety of association relationships. In the first part of the dissertation, we describe *RankCover*, a new non-parametric association test for association between two variables that measures the concentration of paired ranked points. Here 'concentration' is quantified using a disk-covering statistic that is similar to those employed in spatial data analysis. Analysis of simulated datasets demonstrates that the method is robust and often powerful in comparison to competing general association tests. We also illustrate *RankCover* in the analysis of several real datasets. Using *RankCover*, we also propose a method of testing the association of two variables while controlling the effect of a third variable.

In the second part of the dissertation, we describe statistical methodologies for testing hypotheses that can be collected into groups, with each group showing potentially different characteristics. Methods to control family-wise error rate or false discovery rate for group testing have been proposed earlier, but may not easily apply to expression quantitative trait loci (eQTL) data, for which certain structured alternatives may be defensible and enable the researcher to avoid overly conservative approaches. In an empirical Bayesian setting, we propose a new method to control the false discovery rate (FDR) for grouped hypothesis data. Here, each gene forms a group, with SNPs anno-

tated to the gene corresponding to individual hypotheses. Heterogeneity of effect sizes in different groups is considered by the introduction of a random effects component. Our method, entitled *Random Effects model and testing procedure for Group-level FDR control* (*REG-FDR*) assumes a model for alternative hypotheses for the eQTL data and controls the FDR by adaptive thresholding.

Finally, we propose *Z-REG-FDR*, an approximate version of *REG-FDR* that uses only Z-statistics of association between genotype and expression at each SNP. Simulations demonstrate that *Z-REG-FDR* performed similarly to *REG-FDR*, but with much improved computational speed. We further propose an extension of *Z-REG-FDR* to a multi-tissue setting, providing a basis for gene-based multi-tissue analysis.

ACKNOWLEDGEMENTS

I wish to express my sincerest gratitude to my advisor Dr. Fred Wright for guiding and mentoring me. I have profoundly benefited from his guidance over the past five years. His enthusiasm for pursuing interesting statistical problems at the highest level of scientific integrity and rigor has been a constant source of inspiration for my research. I deeply thank him for allowing me the space to think myself and for fostering my capacity as an independent researcher. I feel very fortunate to have him as my advisor and the lessons that I learnt through this journey will always stay with me.

I would like to thank my co-advisor Dr. Andrew Nobel for his support and critical remarks on my work. His constant invigilation and assistance helped me a lot.

I would also like to thank all the other members in my thesis committee, namely, Drs Wei Sun, Yun Li and Karen Mohlke, for careful review and providing valuable suggestions in improving the contents of this thesis.

I am extremely thankful and indebted to Dr. Pranab Kumar Sen for providing support whenever I needed. I thank Dr. and Mrs. Sen also for sponsoring the Kalyani Sen International Student Scholarship in Biostatistics that helped me to overcome the financial burden of studying in a foreign country.

Life as an international graduate student is not easy. Mine was no exception. Things would have been very difficult for me without the helping hands of Rinku Majumder and Samarpan Majumder. I would remain indebted forever to their concern and care. The warmth of their love made me feel at home. They have always been there in my difficulties.

I sincerely thank my friend Sayan Dasgupta for his academic and non-academic guidance throughout the past five years.

I am fortunate enough to be surrounded by a wonderful set of friends, who deserves special mention. Sujatro Chakladar, Lopamudra Kundu, Anjishnu Banerjee, Wangsuk Choi, Suprateek Kundu, Swarnava Mukherjee, Ranajoy Bhattacharjee, Anabil Gayen, Vivek Atal, Anjan Mandal, Rakesh Krishnan Poduval, Avijit Shee, Projjwal Das, Jishnu Datta - they all were beside me any time I needed.

I have been blessed to be friends with Gen Li, Sayantan Banerjee, Abhishek Pal Majumder, Subhamay Saha and Pourab Roy, the academic discussions with whom highly enriched me.

I would like to thank my parents for their faith in me and allowing me to be as ambitious as I wanted. It was under their watchful eye that I gained so much drive and an ability to tackle challenges in life. I would not be here without their sacrifices, patience, continuous support and unconditional love.

Last but not the least, I express my heartiest thanks to my wife Sreemala Das Majumder, who has provided immense support in tough times. Without her constant encouragement and love, this journey would not have been possible. Even after remaining thousands of miles away, she has always tried her best to provide me company and support.

Finally, I dedicate this thesis to my wife and my parents for making me who I am.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Testing of General Association

The need for statistical methods to identify general pairwise association measured between variables is increasingly recognized, as evidenced by recent attention to methods such as distance correlation (dCor) (Székely et al. 2007, Székely and Rizzo 2009), Maximal Information Coefficient (MIC) (Reshef et al. 2011), and the Heller-Heller-Gorfine (HHG) method (Heller et al. 2013). The term *general association* refers to any departure from independence among random variables, and methods differ in the types of departures to which they are sensitive. The need for general association tests is perhaps greatest for analysis of large datasets, for which discovery-based approaches are needed, without prior hypotheses regarding the form or structure of dependence. In addition to the need to test dependence among pairs of variables as a primary analysis, dependencies can invalidate inference for downstream methods that require independence among input variables (Albert et al. 2001).

The methodologies for detecting general association are numerous and consist of several ways to approach the problem. We consider only non-parametric procedures since the methods with parametric assumptions are not 'general' in the true sense. Here, we discuss the most relevant and applicable ones from each approach with special attention to some methods that are relatively new, easy to interpret, computationally less expensive and at the same time most useful in terms of their robustness and power to detect different forms of general associations.

1.1.1 Classical non-parametric tests

Classical tests attempting to detect general association date back to the early part of the last century with Spearman's rank correlation (Spearman 1904) and Kendall's tau (Kendall 1938). Standard tests based on these rank correlations assume values are not tied, and are primarily designed for monotone relationships, but are not principally different in spirit from Pearson's product moment correlation.

Many trend tests (Mann 1945, Kendall 1975, Cuzick 1985, Hamed and Ramachandra Rao 1998) were devised over the years for testing linear and non-linear trends, primarily in time series data. However, they also suffer from insensivity to non-monotone relationships.

1.1.2 Methods in spatial statistics

The spatial statistics literature is abundant with tests of complete spatial randomness (CSR), which is closely related to the general association of two variables. Complete spatial randomness as defined by Diggle (1983) occurs when

1. the number of events in any planar region A with area $|A|$ follows a Poisson distribution with mean $\lambda|A|$.

2. given n events x_i in a region A, the x_i's form an independent random sample from the uniform distribution on A.

The self-consistency of the above two conditions is a non-trivial fact that can be proved. If two variables are associated, their scatter plot is expected to deviate from such CSR since the points will be more clustered as compared to the independent case. However, for testing general association to be exactly equivalent to testing CSR, the marginal distributions of the two random variables must be uniform. Also, CSR can be violated if the occurrence of a point is either encouraged or inhibited the occurrence

of other points in the neighborhood of it, but the alternative of inhibition is not very relevant for testing general association. These differences can be somewhat reduced by using the ranks of the two variables while testing general association, since each component of $rank(X)$ and $rank(Y)$ has a discrete uniform distribution for any two jointly distributed random variables X and Y. However, note that these components are not independent since a rank vector needs to be a permutation of $(1, 2, ..., n)$, where n is the sample size. One sided tests will be appropriate in this case to test for the association (and not for inhibition).

A number of testing procedures sensitive to local clustering have been devised in the field of spatial statistics (Holgate 1965b;a, Diggle et al. 1976, Donnelly 1978, Ripley and Silverman 1978, Hines and Hines 1979, Ripley 1979, Grabarnik and Chiu 2002, Smith 2004, Torabi and Vahidi-Asl 2009). Among the most popular ones, the G and F functions by Diggle (Diggle 1983) use nearest neighbor distances to devise a test against the hypothesis of complete spatial randomness. The two functions are closely related and are proved to be asymptotically equivalent (Diggle 1983). Diggle suggested the use of Monte Carlo simulations to obtain the distributions of empirical versions of the whole curves $G(x)$ and $F(x)$, but it is computationally expensive. Clark and Evans (1954) suggested a test based on mean nearest neighbor distances and an asymptotic distribution was proposed by Donnelly (1978). However, it assumes joint uniformity of the two variables and hence cannot be used in the context of general association.

Coverage processes are somewhat related to such spatial statistics ideas and find potential applications in ballistics, queueing theory, statistical mechanics, molecular biology and so on. In the theory of coverage process, each of the spatial points is assumed to be generated by a stochastic point process, which is not necessarily a Poisson process (discussed in more details in Section 2.6). However, the theory does not directly apply to the general association testing.

1.1.3 Other methods of detecting general association

Some other important methods for detecting general association are maximal correlation (Hirschfeld 1935, Gebelein 1941, Rényi 1959), Hoeffding's D (Hoeffding 1948), and mutual information. The maximal correlation, also known as Renyi correlation, between two random variables X and Y, is defined as $\max_{f(x),g(y)} E(f(X)g(Y))$ subject to $E(f(X)) = E(g(Y)) = 0$ and $E(f(X)^2) = E(g(Y)^2) = 1$. The maximal correlation enjoys various desirable theoretical properties including that it is zero if and only if X and Y are independent. However, there is no explicit formula to calculate it. Breiman and Friedman (1985)'s Alternating Conditional Expectations (ACE) algorithm is the most common algorithm to approximate it. Bickel and Xu (2009) provided another way to approximate the maximal correlation and a test based on it.

Hoeffding's D measures the difference between the joint ranks and the product of their marginal ranks. It can identify even non-monotone associations, but fails to identify non-functional relationships like circle or cross (Fujita et al. 2009, de Siqueira Santos et al. 2013).

The mutual information of two random variables X and Y is defined as

$$MI(X,Y) = \int\int f_{X,Y}(x,y) log_2(\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)})dxdy$$

The mutual information is 0 if and only if $X$and Y are independent. Several methods to estimate the mutual information have been proposed (Paninski 2003, Daub et al. 2004, Kraskov et al. 2004, Moon et al. 1995). The test of general association using these estimators of mutual information are observed to be powerful when the sample size is large, but not satisfactory for small samples (de Siqueira Santos et al. 2013).

1.1.4 Recent advancements

Recently, a number of methods using Reproducing Kernel Hilbert Spaces (RKHS) have been proposed (Fukumizu et al. 2007, Gretton et al. 2008, Gretton and Györfi 2008). These methods have some desirable properties (Gretton et al. 2009), but are complex in nature and not always easy to compute. On the other hand, three methods developed very recently have been extremely popular due to their simplicity, desirable theoretical properties, relative ease of computation and power to detect several forms of association. We discuss these methods in greater detail.

Distance correlation (dCor), introduced by Székely et al. (2007) is motivated by consideration of distances between the empirical characteristic function under the null vs. under the alternative. For observed data, the dCor statistic is the Pearson correlation of distances (after some adjustments) between all pairs of samples. For an observed random sample $(x, y) = \{(x_k, y_k) : k = 1, 2, ..., n\}$, the distances between pairs of samples are defined as $a_{kl} = |x_k - x_l|$ and $b_{kl} = |y_k - y_l|$; $k, l = 1, 2, ..., n$. The approach is intuitively sensible when the relationship is monotone, as sample pairs that are close on the x-axis should also be close on the y-axis. However, for non-monotone relationships, pairs of points that are close on the x-axis can be quite distant on the y-axis (Figure 1.1).

dCor satisfies several ideal theoretical properties (Székely et al. 2007). It is zero if and only if the two variables are independent and is the only method with an explicit formula to enjoy such property. Also, dCor can be used in higher dimensions and has an interpretation related to Brownian distances (Székely and Rizzo 2009).

The maximal Information Coefficient (MIC), proposed by Reshef et al. (2011) measures the largest possible mutual information achievable by any x-y grid applied to the data. Reshef et al. (2011) provided a quick algorithm to calculate the MIC and showed that it has two desirable properties: (i) It is *general* in the sense that with sufficient

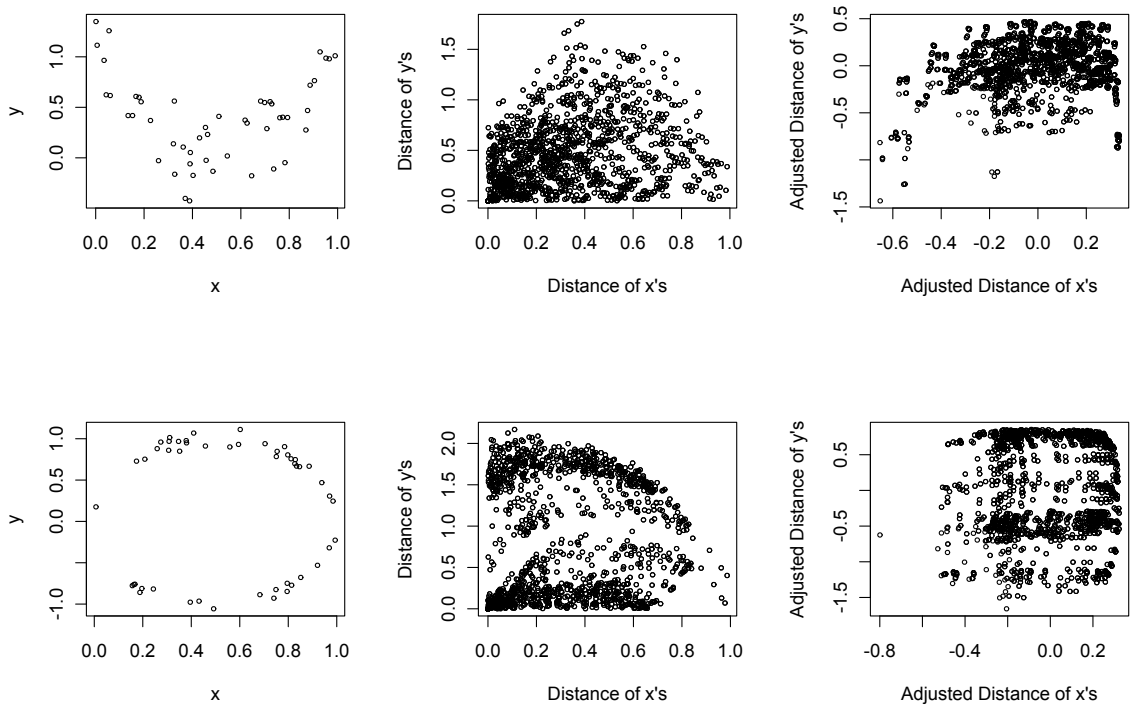

Figure 1.1: Illustration of paired adjusted distances underlying dCor. (top row) Illustration of dCor for a quadratic relationship between x and y.(bottom row) A circular relationship between x and y. The adjusted paired distances show little correlation.

sample size it is able to detect a wide range of associations without being limited to any specific form. (ii) It is *equitable* in the sense that the value of the coefficient is similar for various forms of association that are equally 'noisy' in their departure from a functional relationship.

Simon and Tibshirani (2014) argued that such *equitability* may not be a desirable property while testing for general association, as it might lead to lower power of the test. However, recently there have been debates over the appropriate definition of *equitability* and whether MIC truly enjoys that property (Kinney and Atwal 2014a;b).

Heller et al. (2013) proposed HHG, a new test of general association based on a simple geometric idea that if X and Y are associated then there will a point (x_0, y_0) and radii around R_x and R_y such that the joint distribution of X and Y will differ from

the product of marginal distributions in the Cartesian product of balls around (x_0, y_0). The test uses as the test statistic a sum of n Pearson chi-square statistics where n is the number of paired observations. It can also be extended to higher dimensions. The method has been shown to be consistent as n grows larger and simulation studies were presented to demonstrate that it has high power against several alternatives.

1.1.5 Summary

To summarize, several tests of association have been found to perform well in terms of power in different situations. MIC, dCor and HHG are probably the most appealing in terms of their power against different alternatives, desirable theoretical properties and computational efficiency. However, their performance for small samples against various forms of associations has been relatively unexplored. In Chapter 2 we will present a comparison of these methods with our newly proposed method *RankCover*.

Our method *RankCover* is robust and powerful against different forms of association. The method has been applied on both simulated and real datasets and has been observed to perform better than competing methods in many situations. It is truly 'general' in the sense that it does not depend on the distributions of the two variables under consideration, and has the potential to detect any departure from independence.

1.2 Control of False Discovery Rate for Grouped Hypotheses

Modern scientific technology has given rise to large scale simultaneous inference problems where thousands of tests are carried out at the same time. Special care is needed to ensure that the incorrect rejection of null hypotheses are kept under control. Such control of false positives can be achieved in different ways. The false discovery rate (FDR) approach (Benjamini and Hochberg 1995) is contemporary and has been proved to have advantages over other approaches like controlling the family-wise error rate (FWER). The Benjamini and Hochberg method has been refined to better understand

behavior under dependency, and to accommodate certain dependency structures. Often such hypotheses form into groups that exhibit different properties. The control of FDR without considering the group classification has the potential problem of over or under sensitivity as significant instances of one group might be hidden among the nulls of another group, and insignificant instances might look like significant (Cai and Sun 2009, Efron 2008).

FDR based approaches have also been studied in the domain of interval estimation (Benjamini and Yekutieli 2005, Jung et al. 2011, Zhao and Gene Hwang 2012). However, we will focus on the methods controlling FDR in the grouped hypothesis setting, especially considering the applications for expression quantitative trait loci (eQTL) data.

1.2.1 Classical methods and family-wise error rate

Methods to control type I error, after considering the effects of multiple testing, are more than fifty years old and include the Bonferroni method (Dunn 1961) and Sidak method of multiple comparison (Šidák 1967), being proposed after the works of Tukey and Scheffe in the 1950's. The Sidak method assumes the hypotheses to be independent and can be highly conservative if the correlations are positive. The Bonferroni method does not assume independence and can be even more conservative. Holm (1979) introduced the concept of a stepped procedure that can be used to improve the Bonferroni or Sidak method to obtain less conservative control. Using a similar approach, Hochberg (1988) proposed a step up procedure to obtain higher power. The concept of 'Family-Wise Error Rate (FWER)' was formalized by Westfall and Young (1993). They also introduced a permutation based procedure, applicable to many datasets, which can control the FWER exactly at the target level under a permutation null.

The idea of FWER can be understood from Table 1.1. Suppose we have a total of

m hypotheses and m_0 of them are true null. Based on a particular rejection criterion, let R of them be rejected. The cross-classification of the truth and the decision is as shown in Table 1.1. Then, FWER is defined as the probability of making at least one false discovery, i.e. $P(V \geq 1)$.

	True Null	True Alternative	Total
Rejected	V	S	R
Accepted	U	T	$m - R$
Total	m_0	$m - m_0$	m

Table 1.1: Showing the cross-classification of true and false null hypothesis against the decision to accept or reject

1.2.2 The false discovery rate approach

Benjamini and Hochberg (1995) argued that the FWER may not be the error criterion that should be used for multiple hypothesis testing. They introduced the idea of the False Discovery rate (FDR) and claimed that it is a quantity that is desirable to control. The FDR is the expected proportion of false positives among all rejected cases. In the light of Table 1.1, FDR can be defined as $E(\frac{V}{max\{1,R\}})$. Benjamini and Hochberg (1995) proposed a linear step up (LSU) procedure for controlling FDR and showed that the control of FDR at the same target level as an FWER-controlling method will result in a less conservative procedure and higher power to detect significant cases. Also, controlling FDR assures the weak control of FWER when all the null hypotheses are true. Even though the original work assumed that the hypotheses are independent (Benjamini and Hochberg 1995), later Benjamini and Yekutieli (2001) showed that the same procedure gurantees control of FDR even when the hypotheses are positively dependent in a certain way (positive regression dependence from a subset, PRDS). They also showed that under completely unspecified dependence structure, the LSU procedure still controls the FDR if the target level is adjusted by $(1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{m})$. To be more specific, the Benjamini-Hochberg procedure with target level q works in the

following way.

$FDR \leq \frac{m_0}{m} q$ when hypotheses are independent,

$FDR \leq \frac{m_0}{m} q$ when hypotheses are positively dependent (PRDS),

$FDR \leq \frac{m_0}{m} q(1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{m})$ for general dependence.

It is clear that the FDR control using the Benjamini-Hochberg LSU procedure can thus be extremely conservative if the proportion of true null hypotheses (m_0/m) is not close to 1. Even though controlling FDR at the exact level may not always lead to the most powerful procedure (Cao et al. 2013), in most cases the power is reduced when a procedure controls the FDR at a lower level than the target. This observation inspired the idea of 'adaptive' procedures, where m_0 is first estimated from the data and then the LSU procedure is used for a target level $qm/\hat{m_0}$ (Benjamini and Hochberg 2000, Storey 2002, Black 2004). Such plug-in type procedures, even though valid as 'oracle' procedures, might not always control the FDR when m_0 is estimated from the same data. Especially under dependency, the variability of the estimate of $1/m_0$ can be very high (Farcomeni 2007b, Blanchard and Roquain 2009). Benjamini et al. (2006), Blanchard and Roquain (2009), Benjamini et al. (2009), Gavrilov et al. (2009) proposed several adaptive methods which can be proved to control the FDR at the target level.

1.2.3 Extension and different approaches to FDR

There has been considerable research in the field of multiple hypothesis testing using FDR over the last two decades including many studies regarding the properties of the FDR approach under different scenarios (Green and Diggle 2007, Ferreira et al. 2006, Sarkar 2008; 2002, Farcomeni 2007a). The effect of dependence among the hypotheses has been the topmost concern for the researchers. Even when the Benjamini-Hochberg procedure controls the FDR, it might be overly conservative under dependence (Qiu

and Yakovlev 2006, Schwartzman and Lin 2011). Owen (2005) has noted that the variance of the number of false discoveries might be greatly inflated under dependence. Yekutieli and Benjamini (1999) proposed a permutation based approach to take care of the dependence, but it is computationally burdensome for large number of hypotheses. Other procedures to take care of the dependence have been proposed including a hidden Markov model based approach by Sun and Cai (2009). They propose an 'oracle' procedure as well as an asymptotically optimal data-driven procedure, but the entire procedure requires a natural ordering of the hypotheses such that dependencies of null/alternative hypotheses may be exploited. Genovese and Wasserman (2004; 2002) extended the FDR approach and also introduced the idea of 'False Negative Rate' (FNR) which is the expected proportion of false negatives among all non-rejections (Genovese and Wasserman 2002). They proposed an optimal method which minimizes the FNR subject to a bound on FDR. Sun and Cai (2007) also provided an 'oracle' method based on a decision theoretic framework that minimizes FNR while controlling the FDR. They showed that when the method is data driven, it asymptotically attains the performance of the 'oracle' procedure.

Storey (2003) introduced a Bayesian approach to FDR by considering the hypotheses to be Bernoulli random variables with probability π_0, where $\pi_0 = P(H_0)$ for each hypothesis. For a rejection region $\mathcal{R}$ and observed data z, FDR is defined from the Bayesian veiwpoint as $P(H_0|z \in \mathcal{R})$. Storey and Tibshirani (2003) introduced the concept of 'q-values', the FDR-equivalent of p-values, which can be used in multiple testing without the prior fixing of a target FDR level. Multiple other error rates have been proposed including 'positive False Discovery Rate' (pFDR) defined as $E(V/R|R > 0)$ (Storey 2002), 'Fdr' defined as $E(V)/E(R)$ (Benjamini and Hochberg 1995), 'k-FWER' defined as $P(V \geq k)$ (Lehmann et al. 2005), and tail probaility $P(V/R \geq q)$ of the false discovery proportion (van der Laan et al. 2004). Benjamini (2010) argued that such multiplicity of error rates is welcome as they find applications in different

situations. However, the FDR is widely seen as the most useful one, having the desirable properties under the most general conditions (eg Fdr or pFDR cannot be controlled when all the null hypotheses are true).

1.2.4 The empirical Bayes approach and local false discovery rate

The empirical Bayes approach uses a Bayesian set up assuming the null hypotheses to be Bernoulli random variables, but estimates the prior probability π_0 instead of assuming prior belief. Empirical Bayes methods use the advantages of both classical and Bayesian approaches and can be superior to both in many cases (Casella 1985). Efron et al. (2001b) introduced the empirical Bayes approach for controlling FDR in microarray datasets and mentioned that such approach has an easy appeal and interpretation. The model, known as two-groups model, can be used in other applications as well. With such a model, for a given data z related to a hypothesis, the density can be written as a mixture density:

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z) \tag{1.1}$$

where f_0 and f_1 are the densities under null and alternative, respectively. The adaptivity is in inherent to such procedures since the estimatiion π_0 is equivalent to estimating m_0 in the classical FDR setting.

The local false discovery rate (lfdr) (Efron et al. 2001a) is defined as the posterior probability $P(H_0|z)$ of the true null given the data. Efron et al. (2001a) showed that lfdr has a natural connection with the Benjamini-Hochberg FDR controlling method that allows one to control the FDR by an adaptive step up method (see Theorem 1). The empirical Bayes approach using lfdr has, in principle, the advantage of inherently taking care of the dependencies (Efron et al. 2001a). Thus, one doesn't have to worry about the dependency structure of the p-values like the Benjamini-Hochberg LSU procedure.

The difficulty of using the empirical Bayes approach is to estimate the lfdr's. The requirement of the estimation of the null density f_0 has been discussed by many researchers (Efron 2004, Jin and Cai 2007, Schwartzman 2008) although in some cases it might be assumed to be a known distribution (Efron et al. 2001a).

1.2.5 Grouped Hypotheses

Grouped hypothesis testing is a special case of multiple testing where the hypotheses have a natural stratification and adjustments for multiple comparison is required not only within each group, but also for the existence of multiple groups. For instance, gene expression data can be grouped according to the ontologies (Ashburner et al. 2000). For cis-eQTL analysis, there is a natural grouping in terms of the different genes. Within each gene, there are several SNPs local to the gene with which the associations are tested. For eQTL studies such as GTEx (Lonsdale et al. 2013), it is often useful to find out whether there is any eQTL within a particular gene since genes are believed to be directly associated with the diseases. An example for expression data is presented by Heller et al. (2009) where gene-sets are thought of as units of interest and a method to find out gene-sets that are differentially expressed has been developed. Benjamini and Heller (2007) reports an example where the clusters are of more interest than individual locations in a neuro-imaging study. They propose an adaptive procedure to control the FDR for clusters, i.e. to control the proportion of clusters erroneously rejected out of all rejected clusters.

Another important factor for grouped testing is the heterogeneity of the groups. Different groups might have different properties, and ignoring that fact might lead to overly conservative or overly anti-conservative results (Cai and Sun 2009, Efron 2008). Efron (2008) demonstrated that pooling all the groups together is not recommended for such heterogeneous groups. He also showed that separate analysis controlling FDR at α for each group and then combining the results ensures that the overall control of FDR

at same target level α. However, such choice of $\alpha_i = \alpha$ for each group is not optimal (Cai and Sun 2009). Yang and Jeong (2013) has applied such a separate analysis approach to RNAseq data. The conditional lfdr based 'oracle' procedure (when the distributional information of each group is known) introduced by Cai and Sun (2009), when applicable, has been shown to be optimal in the sense that it controls the overall FDR and minimizes the overall FNR. When the parameters are unknown, they propose a data-driven procedure that is asymptotically equivalent to the 'oracle' procedure.

Most of the other methods use weighted p-value based approaches to combine p-values from different groups (Benjamini and Hochberg 1997, Genovese et al. 2006, Hu et al. 2010). Roeder and Wasserman (2009) showed that such weighted p-value based methods are robust to weight misspecification. Hu et al. (2010) proposed a 'Group Benjamini Hochberg' method, but it is limited by the assumption that the non-null distribution of different groups are same. Zhao and Zhang (2014) proposed another weighted p-value method where the weights are obtained by maximizing a power-related objective function. Wang et al. (2010) introduced a Hidden Markov Model based method for group testing and succesfully applied it to GWAS data. Another method targeted at GWAS data was proposed by Sun et al. (2006).

A different way to approach grouped testing is to adopt a hierarchical structure and sequentially test at different levels. One such example might be to split a genome-wise data into chromosomes, which can be further split into arms, then into genes and so on. Only the chromosomes found to be significant in the first stage will be tested at the next level. There exist several methods controlling FWER in such tree-like set up (Goeman and Finos 2012, Meinshausen 2008), Yekutieli (2008) proposed a method that controls for the overall FDR. However, Benjamini and Bogomolov (2014) have cautioned that such selective procedures may not control FDR at the group levels unless some adjustments are made. The authors specifically mentioned different adjustment methods for controlling different error rates.

1.2.6 Application in eQTL studies

There have been a lot of studies regarding eQTL data over the past decade. eQTL mapping methods have rapidly moved from classical genetic methods for linkage or association mapping to modern computationally efficient algorithms. Wright et al. (2012) provides a review of the different eQTL mapping methods. While some of the researchers emphasize on the statistical modeling aspect (Kendziorski et al. 2006, Chen and Kendziorski 2007, Gelfond et al. 2007), other methods focus on developing fast and efficient algortithms for the huge eQTL datasets (Gatti et al. 2009, Shabalin 2012, Purcell et al. 2007).

The FDR controlling procedure due to Benjamini and Hochberg (1995) and the q-value approach by Storey and Tibshirani (2003) are the most common approaches to control FDR in eQTL studies. Among other approaches, some clustering methods are used by Jia and Xu (2007) and Chun and Keleş (2009). However, the natural grouping defined by the genes in the eQTL data is relatively unexplored. Due to the large number of groups and large number of hypotheses within the groups, many group-testing methods become computationally burdensome for eQTL datasets. However, the methods might be simplified by making further assumptions considering the special structure of the eQTL data.

1.2.7 Summary

The past two decades have seen extensive studies on multiple hypothesis testing using the FDR controlling approach. Different situations like grouped hypotheses and mutually dependent hypotheses have been considered by researchers and methodologies to tackle them have been proposed. However, appropriate approaches to avoid conservativeness under dependence are still somewhat unclear. While there has been lot of research on both FDR control in grouped hypothesis testing and analysis of eQTL data

separately, the application of grouped hypothesis testing for eQTL data has not been well explored. The natural grouping of the eQTL data using the genes as groups has been largely ignored when applying multiple comparison techniques, except using computationally intensive method such as permutation (Ardlie et al. 2015). There might be assumptions that do not hold in general for grouped hypotheses, but hold in eQTL data due to its special structure. In Chapter 3, we will discuss how such special structure of the data can be used to develop new group testing methodologies for eQTL datasets.

Our method *Random Effects model and testing procedure for Group-level FDR control* (*REG-FDR*) models the alternative for the eQTL data and controls the FDR by adaptive thresholding. *Z-REG-FDR*, an approximate version of *REG-FDR*, is also proposed which exhibits similar results with much improved computational speed. As *Z-REG-FDR* is very similar to *REG-FDR*, which is based on maximum likelihood estimation, *Z-REG-FDR* is conjectured to have near-optimality properties in estimation due to its use of an approximate MLE. This method is not only very fast compared to other grouped hypothesis testing methods, but it also does not require the full data to fit the model. In fact, using only the p-values for each gene-SNP pair is sufficient to conduct the gene-level hypothesis testing and control of the FDR.

1.3 Overview of the thesis

In Chapter 2 we develop *RankCover*, a new method to detect general association. The results of application of the method on both simulated and real datasets are presented. Our proposed methodologies to control FDR in a grouped hypothesis set up are described in Chapter 3. The advantages and limitations of our approaches are discussed in this chapter. In Chapter 4 we discuss a multi-tissue extension of our grouped hypothesis testing method.

CHAPTER 2: A PROCEDURE TO DETECT GENERAL ASSOCIATION

2.1 Motivation

Adapting ideas from spatial analysis, we propose *RankCover*, a method that quantifies the concentration of (x, y) values by measuring the area covered by laying disks of a fixed radius over each point in the scatter plot of the ranks of the two variables. In the presence of association, this area is expected to smaller than that under independence. Therefore, a left tailed test is appropriate in this case.

RankCover starts by computing ranks of the original x and y values, and we assume there are no tied values. The use of ranks considerably simplifies the problem, by placing the intervals between successive ranked values on a common scale. In addition, for ranked values, the null distribution depends only on the sample size n. Thus the only computation lies in computing the observed statistic, while the null distribution can be pre-computed and is applicable to any dataset of size n.

Diggle's $F(\delta)$ function as introduced in Diggle (1983) is the distribution function of the distance between a randomly chosen point in a region to the nearest observed point (x_k, y_k). To obtain an empirical estimate of the $F(\delta)$, the investigator conceptually lays disks of radius δ on each point (x_k, y_k) and calculates the proportion of the surrounding region covered by the union of the disks (Figure 2.1). If x and y are highly associated, the areas covered by the disks should be small, and therefore *RankCover* rejects only in the left tail of the statistic described below.

Different distance metrics can be used for this purpose and the shape of the disks

depend on the choice of the distance metric. For instance, Euclidean distance leads to circular equidistance contours, resulting in circular disks, while the disks are diamond-shaped for Manhattan distance (Figure 2.1).

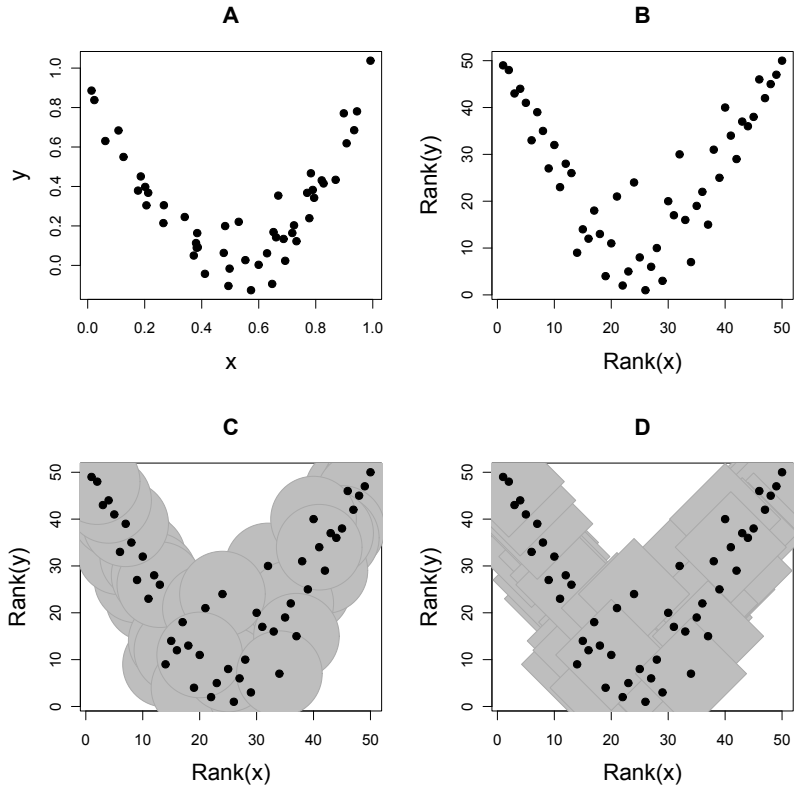

Figure 2.1: Illustration of *RankCover* for sample size $n = 50$: **A.** Scatter plot of the two variables. **B.** Scatter plot on the rank scale **C.** Disks laid on the scatter plot on rank scale using Euclidean distance **D.** Disks laid on the scatter plot on rank scale using Manhattan distance.

2.2 The test statistic

The empirical estimate of $F(\delta)$ can be obtained using the proportion of area covered by the discs. For a given sample $((x_1, y_1), ..., (x_n, y_n))$, $x_k \in \mathcal{X}, y_k \in \mathcal{Y}, k = 1, 2, ..., n$, let the total area covered by the union of the disks of radius δ be $A(\delta)$. The empirical estimate of F is given by

$$\hat{F}(\delta) = \frac{A(\delta)}{|\mathcal{X} \times \mathcal{Y}|} \tag{2.1}$$

Let (r_k, s_k) denote the ranks of the kth sample pair, $k = 1, 2, ..., n$. The corresponding version of $\hat{F}$ for ranks is given by

$$\hat{F}_R(\delta) = \frac{A_R(\delta)}{n^2} \tag{2.2}$$

where $A_R(\delta)$ is the area covered by union of disks placed at each of (r_k, s_k).

However, it is difficult to calculate the exact area covered by the union of disks due to the complex nature of possible intersections. Acknowledging the discrete nature of the ranks, we consider only the $n \times n$ grid of possible rank pairs, $\{1, 2, ..., n\} \times \{1, 2, ..., n\}$, and whether each of these values on the grid is covered by at least one disk.

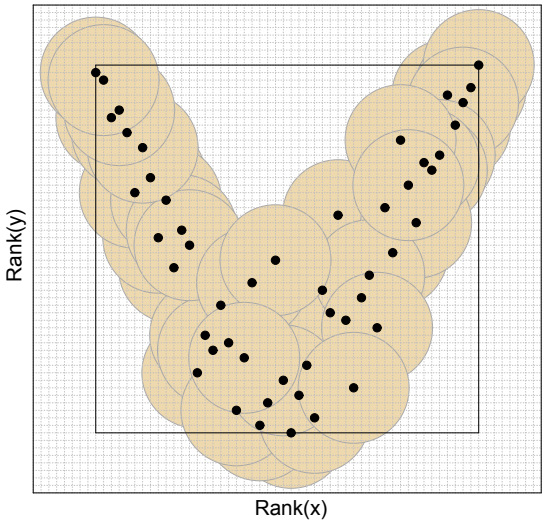

Figure 2.2: Showing the Grid based approach of *RankCover*

Definition 1. *Define* $d(i, j, x_k, y_k)$ = *distance between the point* (i, j) *on the grid and* (x_k, y_k)*;* $d_{ij} = \min_k d(i, j, x_k, y_k)$

Using this definition, a reasonable statistic for fixed δ is

$$\hat{F}_{RG}(\delta) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} I(d_{ij} \leq \delta), \tag{2.3}$$

where $I(.)$ is the indicator function. The grid-based empirical distribution function (EDF) for ranks $\hat{F}_{RG}(\delta)$ can be considered as an approximation to $\hat{F}_R(\delta)$.

The choice of disk size δ is an important consideration which has not been fully addressed in the spatial statistics literature. Diggle (1983) suggested computing the entire empirical distribution function (EDF) $\hat{F}(\delta)$ to develop a new summary statistic to compare against the null curve. However, this approach makes the procedure prohibitively computationally expensive, and we propose using a fixed $\delta = \sqrt{n}$ for Euclidean distance (Section 2.3), with slight modification under Manhattan distance. It is observed that there is very little difference in power to detect association between the method using the entire EDF and the statistic using a fixed $\delta = \sqrt{n}$ (Figure 2.3). In addition, we modify the statistic to account for edge effects of the grid, using an $(n + \lceil\delta\rceil) \times (n + \lceil\delta\rceil)$ grid extending beyond the range of the scatterplot. Here $\lceil\delta\rceil$ is the smallest integer greater than or equal to δ. Finally, our modified test statistic is

$$T(\delta) = \frac{1}{n^2} \sum_{i=1-\lceil\delta\rceil}^{n+\lceil\delta\rceil} \sum_{j=1-\lceil\delta\rceil}^{n+\lceil\delta\rceil} I(d_{ij} \leq \delta), \tag{2.4}$$

where the range of $\{i, j\}$ reflects the outer boundaries of a larger region to account for edge effects. Note that the same divisor n^2 is used allowing $T(\delta)$ to be greater than 1. $T(\delta)$ can be interpreted as the proportion or area covered by the disks as compared to the area of $\mathcal{R}_n$, the $n \times n$ region which is the range of the original scatter plot.

The null distribution of T depends entirely on n, so tables based on simulated null distibutions can be pre-computed for various sample sizes. The following lemma shows that the grid based statistic $T(\delta)$ is asymptotically equivalent to the corresponding

area-based statistic.

Lemma 1. *Let $T_A(\delta)$ be the area based test statistic corresponding to $T(\delta)$ with areas of the disks extending beyond the $n \times n$ square being taken into account. For $\delta = O(\sqrt{n})$, $|T_A(\delta) - T(\delta)| \xrightarrow{a.s.} 0$ as $n \to \infty$.*

Proof. For a single disk with radius δ, by the Gauss circle problem (Gauss 1986), the difference of its actual area and its lattice based approximation $N(\delta)$ is bounded by $2\sqrt{2}\pi\delta$.

Therefore, for $\delta = O(\sqrt{n})$, with probability 1, $|T_A(\delta) - T(\delta)| \leq \frac{2n\sqrt{2}\pi\delta}{n^2} = O(\frac{1}{\sqrt{n}}) \to 0$ as $n \to \infty$. . □

The implication of this lemma becomes obvious in (Section 2.6) when we discuss large sample properties of *RankCover*. In small samples, the two statistics $T_A(\delta)$ and $T(\delta)$ might be quite different. However, there is no reason to believe that one is inferior to the other in small samples since $T(\delta)$ actually computes similar disk coverage statistic for a different disk shape that looks like a polygon.

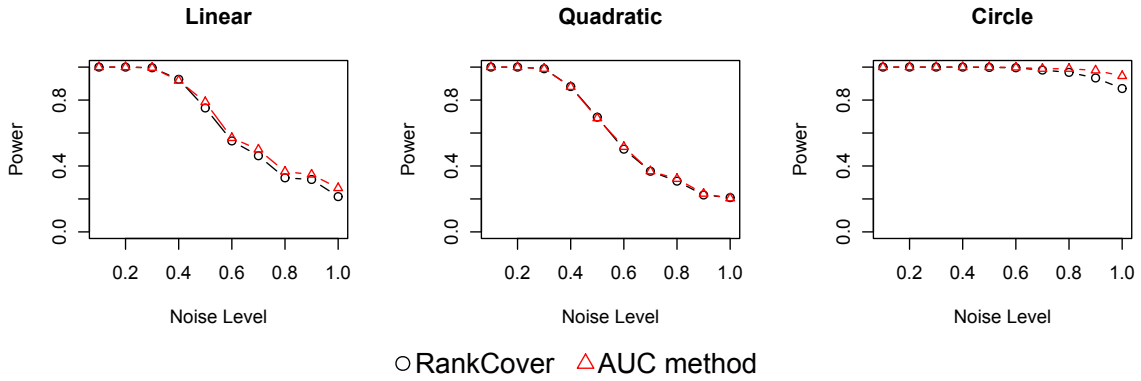

Figure 2.3: Showing the comparison of power of the method using the area under the EDF (AUC method) and that of the method using $\delta_{opt} = \sqrt{n}$

2.3 Choice of parameters and distance metric

The choice of the disc size δ is an important consideration. We have proposed the use of a single optimum choice of δ as opposed to the whole δ versus $\hat{F}(\delta)$ curve used by Diggle (1983). The argument for choosing $\delta_{opt} = \sqrt{n}$ for Euclidean distance and $\delta = \sqrt{\frac{\pi}{2}n}$ is somewhat heuristic, but based on empirical observations for several sample sizes.

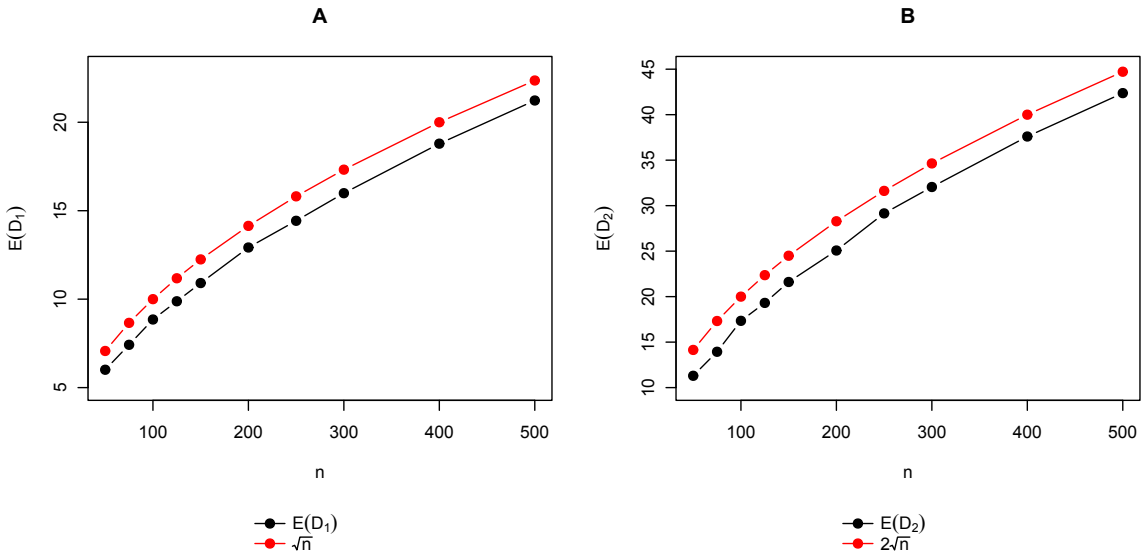

Figure 2.4: Showing the expected δ for which **A.** $T(D_1) = 1$ **B.** $\hat{F}_{RG}(D_1) = 1$

The external region beyond $\mathcal{R}_n$ is used to take care of the edge effects. However, it is the behavior of the disks inside $\mathcal{R}_n$ that primarily differentiates between null and alternative. While trying to find a disk size that will enhance this difference the most, it is reasonable to believe that increasing disk size will not provide much of information once $\mathcal{R}_n$ is completely covered. Since the computational cost increases with the increase of the disk size, one would like to stop increasing the disk size when it stops providing much information. Therefore, we try to find out the disk size for which $\mathcal{R}_n$ is completely covered.

It is a difficult problem to analytically determine the 'stopping' disk size. Further-

more, Hall et al. (1985) proved that for a Boolean process (Discussed in Section 2.6), the probability of coverage is 1 if the area of the disk a_n satisfies

$$a_n/n - log(n) - log(log(n)) \to \infty \text{ as } n \to \infty. \tag{2.5}$$

It is evident that for $\delta = n^\alpha$, this condition is satisfied if and only if $\alpha > \frac{1}{2}$. Even though this result does not have a direct implication in our case, it is suggestive of the order of the 'stopping' disk size. To further explore the stopping condition, we used simulated data and calculated the expectation of two variables D_1 and D_2 defined as below.

$$D_1 = \text{Smallest disk size for which the realized } T(D_1) > 1.$$
$$D_2 = \text{Smallest disk size for which the realized } \hat{F}_{RG}(D_1) > 1.$$

Figure 2.4 Shows that both $E(D_1)$ and $E(D_2)$ are probably of the order $\sqrt{n}$.

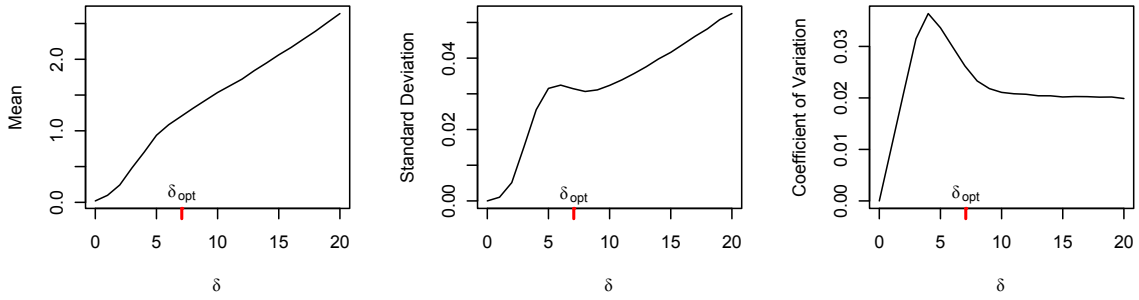

Figure 2.5: Showing the mean, sd and coefficient of variation of $T(\delta)$ for sample size 50 (Euclidean distance is used)

Next, we examine the expectation and standard deviation of $T(\delta)$ under the null for varying δ. These curves calculated based on 1000 simulations under the null are shown in Figure 2.5 and Figure 2.6 for Euclidean distance. There is a clear change of curvature in the expectation in the vicinity of $\delta = \sqrt{n}$, and the standard deviation exhibits a local maximum and minimum in the vicinity. We reason that the local

minimum of the standard deviation represents a good choice for δ. We also note that the point where the expectation curve changes the curvature is approximately the same point as the local minimum of the standard deviation, and the coefficient of variation is almost constant beyond this point. However, there is no closed form expression for this point of local minimum. From simulations under different sample sizes, we have established that such local minima occur near $\delta = \sqrt{n}$ for Euclidean distance, and propose it as our choice of δ_{opt}.

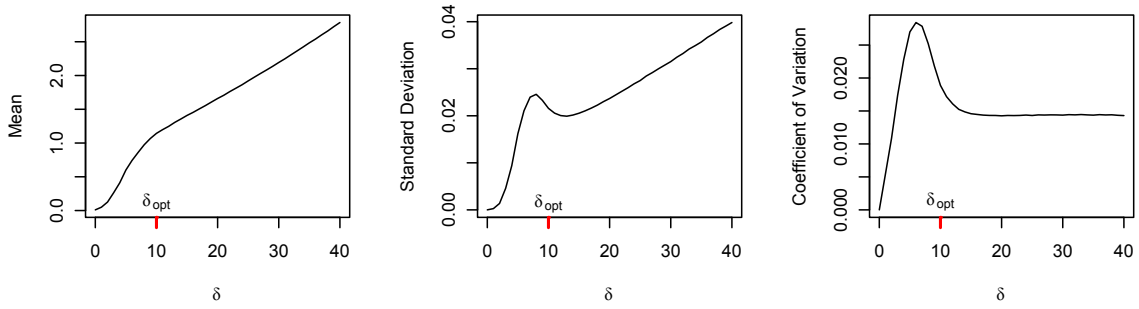

Figure 2.6: Showing the mean, sd and coefficient of variation of $T(\delta)$ for sample size 100 (Euclidean distance is used)

Thus, there is enough reasons to believe that the optimal δ should be of the order $\sqrt{n}$ even though the minimum of the standard deviation is not exactly at $\sqrt{n}$. Rather, the minimum can be better empirically modeled as $\sqrt{n} + \sqrt{n/5 - 10}$ for sample sizes up to 500 (Figure 2.7). However, all these heuristic arguments deal with the behavior of the test statistic under null. We have also compared its power against different alternatives for varying δ. Figure 2.8 shows the average p-value in $-log_{10}$ scale for different forms of association. Clearly there is no single δ for which the power is maximized. However, the power for $\delta = \sqrt{n}$ is close to the maximum power achieved in all the cases. Therefore we conclude that it is not possible to find out a disk size that is 'optimum' in the true sense, but $\delta_{opt} = \sqrt{n}$ can be considered as a reasonable choice for Euclidean distance.

Also, it is observed from simulations that the shape of these curves depends on δ

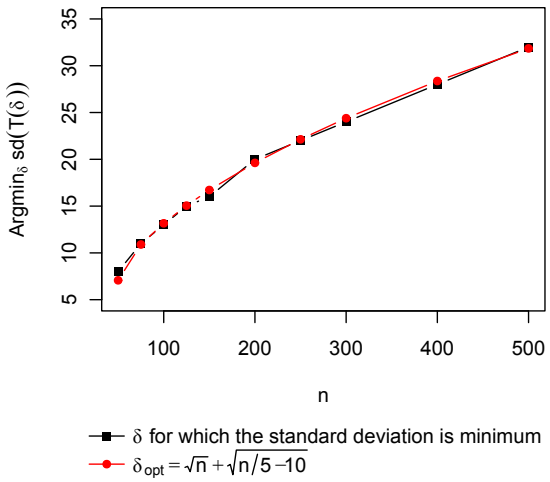

Figure 2.7: Showing the δ for which standard deviation of $T(\delta)$ is minimum for different sample sizes

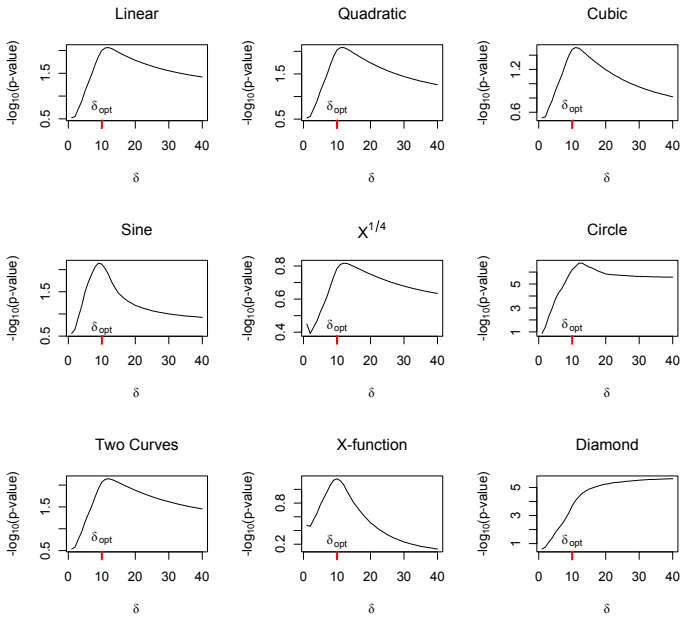

Figure 2.8: Showing the Average p-value using different disk sizes when testing against various forms of association

only through the area of the disk (also shown by Hall (1988) for Boolean process), and so we use $\delta_{opt} = \sqrt{\frac{\pi}{2}n}$ for the Manhattan distance. Using simulations, we have tested that such a choice of δ produces similar curves for Manhattan distance.

For the distance metric d, we consider here both Euclidean and Manhattan distances, for which simulations show similar performance (Section 2.7.1). However, the Manhattan distance has advantages in approximating tail areas since the rejection thresholds follow a sawtooth pattern (Figure 2.9), with jump points occurring at the values of n where $[\delta]$ changes. For large values of n, to reduce computation, one can perform direct simulation for the values of n at, and just prior to, the jump points, followed by linear interpolation for remaining values of n. Therefore we recommend its use and here present results using Manhattan distance.

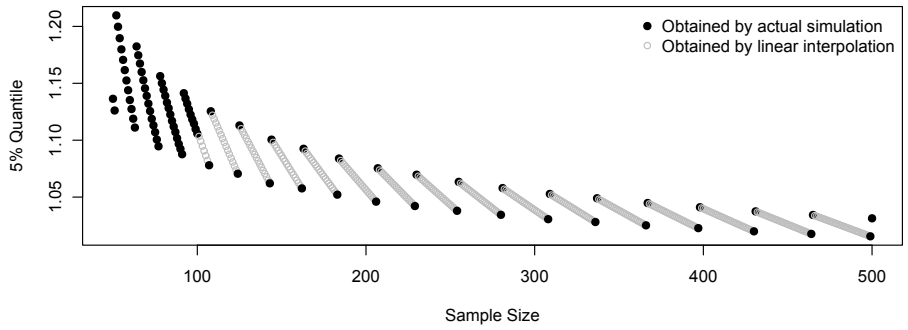

Figure 2.9: Showing the pre-computed thresholds for the *RankCover* method with Manhattan distance. 100000 simulations were used to calculate the thresholds in each case. Simulations were performed for $n = 20, ..., 100$. For large values of n, to reduce computation, tables were generated by (i) performing direct simulation for the values of n at, and just prior to, the jump points, followed by (ii) linear interpolation for remaining values of n.

2.4 Fast Computation of the test statistic

The crude way to compute the test statistic needs to calculate the distances of the n sample points from each of the $(n + \lceil\delta\rceil)^2$ points on the grid. Thus, the order of computation is n^3. We have proposed a method with complexity $O(n^2)$ (Zhou, Wright; personal communication, November 2014). The algorithm first calculates a $(2\lceil\delta\rceil + 1) \times (2\lceil\delta\rceil + 1)$ prototype matrix of 1's and 0's that represents the shape of the

disk. Then the prototype matrix is used to "punch" a hole at each of the sample points (Figure 2.10).

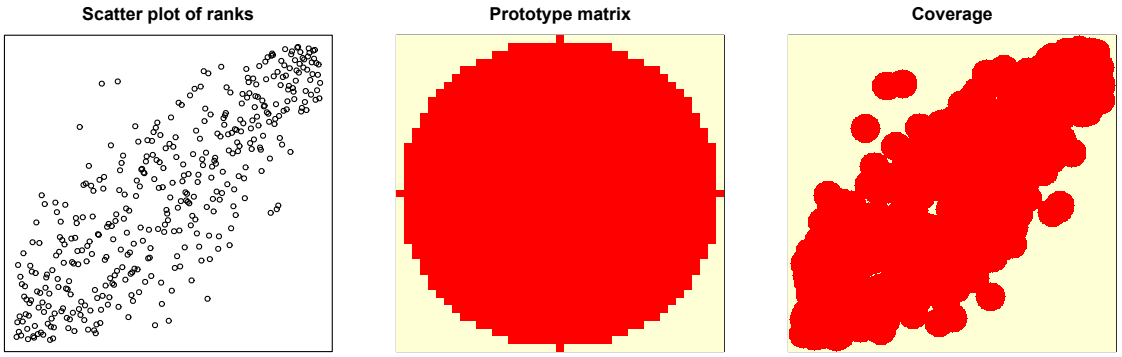

Figure 2.10: Showing the fast computation of *RankCover*

2.5 Exact expectation of the *RankCover* statistic for Manhattan distance

We exploit the desirable properties of Manhattan distance to obtain the exact value of $E(T_n(\delta))$. Let us define the random variables I_{ij}, $i = 1, 2, ..., n; j = 1, 2, .., n$, for each point (i, j) on the grid. I_{ij} is 1 if there is any sample point within the distance δ from (i, j) and 0 otherwise.

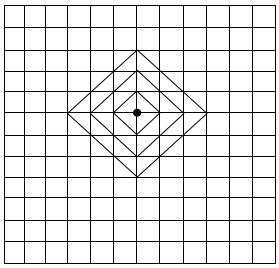

Figure 2.11: Schematic to illustrate calculation of $P(I_{ij} = 1)$ for $1 \leq i \leq n, 1 \leq j \leq n$.

Let us consider the case where the δ-ball lies completely within the $\mathcal{R}_n$. From Figure 2.11, clearly,

$P(I_{ij} = 1) = 1 - \frac{n-1}{n},\ 0 < \delta < 1$

$P(I_{ij} = 1) = 1 - \frac{n-3}{n}\frac{n-2}{n-1}\frac{n-3}{n-2},\ 1 \leq \delta < 2$

$P(I_{ij} = 1) = 1 - \frac{n-5}{n}\frac{n-4}{n-1}\frac{n-5}{n-2}\frac{n-4}{n-3}\frac{n-5}{n-4},\ 2 \leq \delta < 3$ and so on.

In general, if $[\delta] = k$,

$$P(I_{ij} = 1) = 1 - \frac{(n-2k-1)^{k+1}(n-2k)^k}{(n)_{(2k+1)}} \tag{2.6}$$

It becomes more complicated when a part of the δ-ball lies outside the $n \times n$ region. It is difficult to obtain a simplified formula like above, but similar counting procedure can be used to get the expression of the expectation.

Let $[\delta] = k$ and n are given. We need to find $p_{ij}(k, n) = P(I_{ij} = 1)$ for a given point (i, j) on the grid. Define

$$n_l(t, k, n) = min\{t - 1, k\}$$

$$n_r(t, k, n) = min\{n - t, k\}$$

and

$$n(i, k, n) = 1 + n_l(t, k, n) + n_r(t, k, n)$$

.

$n_l(t, k, n)$ is the number of points at the left of $(t, .)$ on the same horizontal line within the δ-ball as well as within the $n \times n$ region. $n_r(t, k, n)$ is the number of such points at the right and $n(t, k, n)$ is the number of such points on that horizontal line. Let $I(t, k, n)$ denote the index vector of the relative positions of the $n(t, k, n)$ points with respect to $(t, .)$. We assume that $I(t, k, n)$ consist of the sorted absolute values and call the rth element of it $I_r(t, k, n)$. For example, in Figure 2.11, for $\delta = 2$, $I(6, 2, 12) = (0, 1, 1, 2, 2)$.

Using simple arguments of geometric probability, clearly,

$$p_{ij}(k,n) = 1 - \prod_{r=1}^{n(i,k,n)} \frac{n-r+1-n(j,k-I_r(i,k,n),n)}{n-r+1} \tag{2.7}$$

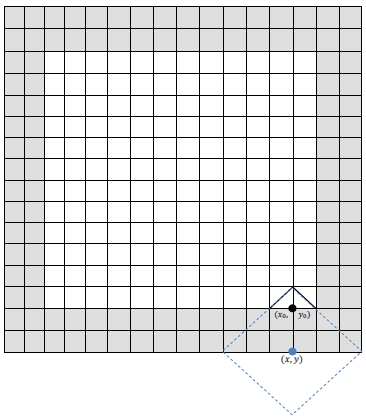

Figure 2.12: Showing the existence of (i_0, j_0) for a point (i, j) outside the $n \times n$ region

Equation 2.7 applies to any point (i, j) within the $n \times n$ region. For (i, j) outside the region, there exists a point (i_0, j_0) (See Figure 2.12) on the edge of the region such that

$$p_{ij}(k,n) = p_{i_0 j_0}(k_0, n)$$

.

Here

$$i_0 = I\{i < 1\} + nI\{i > n\} + iI\{1 \le i \le n\},$$

$$j_0 = I\{j < 1\} + nI\{j > n\} + jI\{1 \le j \le n\},$$

$$k_0 = k - |i - i_0| - |j - j_0|.$$

Equation 2.7 can then be used to obtain $p_{i_0 j_0}(k_0, n)$.

2.6 Large sample properties of *RankCover*

The computation of the *RankCover* statistic might be quite slow if the sample size is very large. For instance, with $n > 10000$, the Monte Carlo simulations to produce the null distribution of the test statistic becomes computationally expensive. The testing procedure will be much simpler and faster if the large sample theoretical distribution of *RankCover* can be determined. In the following sections we discuss the established large sample results pertaining to the theory of coverage process and *RankCover*'s relationship with them. Euclidean distance is considered as the distance metric, but the same arguments can be easily shown to apply for Manhattan distance too.

2.6.1 Coverage Process

The theory of coverage process is related to the idea of *RankCover*. In a simple set up, a coverage process can be thought of as a countable sequence of sets in an Euclidean space (Section 2.6). Suppose $\mathcal{P} = \{\xi_1, \xi_2, ..\}$ is a countable collection of points in $\mathbb{R}^k$ (which might be a stochastic point process (Karr 1991)), and $\{S_1, S_2, ...\}$ is a countable collection of non-empty sets (might be random sets). If $\xi_i + S_i$ denotes the set $\{\xi_i + x : x \in S_i\}$, then $\mathcal{C} = \{\xi_i + S_i : i = 1, 2, ...\}$ is a coverage process. The union of all sets in $\mathcal{C}$ is known as a 'germ-grain' model where the points ξ_i are referred to as 'germs' and the sets S_i as 'grains'. If $\mathcal{P}$ is a stationary Poisson process and S_i's are iid random sets independent of $\mathcal{P}$, then $\mathcal{C}$ is known as a 'Boolean' process.

In a simpler version of coverage process, which is relevant to our problem, the sets S_i are all equal to a fixed set S (in our case, the disks), and the point process $\{\xi_1, \xi_2, ...\}$ is assumed to be generated from a region R, which is known as the 'experiment space'. While $C = \cup_i(\xi_i + S_i)$ is called the total coverage, the vacancy within a subset R of $\mathbb{R}^k$ is defined as

$$V = V(R) = R \setminus C.$$

Note that the set R does not have to be same as the experimental space $\mathcal{R}$ although most of the coverage process literature deals with the vacancy $V(\mathcal{R})$ within $\mathcal{R}$. The proportion of vacancy within R is called the *porosity*, and is directly related to the way *RankCover* is formulated. The major difference is the point process in *RankCover* which is not a Poisson process due to the use of ranks.

Various researchers has found out moments and limiting distributions of vacancy under different conditions. Hall (1985) proved the aymptotic normality of vacancy for a Boolean process and provided the expressions for its mean and variance. Moran (1974) computed limiting distributions of coverage assuming that the points are generated from a normal distribution. Similar work has been done by Miles (1969), Ailam (1966), Hall (1984). However, most of the work in this area has assumed that the points are generated independently. In the presence of dependency, the derivation of these limit theorems becomes extremely complicated (Hall 1988). Little work has been done with dependent cases, and very specific situations are handled in the few attempts that have been made (Moran 1973). Those situations are not similar to *RankCover*.

We present a few early results with the conjecture that as n becomes large, the difference between *RankCover* and the case considered by Hall (1985) becomes negligible. We provide empirical evidence to support the conjecture that for very large n the two distributions to become similar.

2.6.2 Asymptotic Negligibility of the edge effect

Hall (1984) proved that the edge effects are asymptotically negligible in the sense that the distribution of vacancy under Boolean process remains the same even if edge effects are ignored. However, the way edge effects are defined by Hall (1984) are quite different from what we consider for *RankCover*. The coverage was considered within the experimental space $\mathcal{R}$, and the edge effects in that case were the way the probability of a point within $\mathcal{R}$ being covered changes when it is near the edge. For *RankCover*,

we consider the coverage beyond the experimental space $\mathcal{R}$. Therefore the result due to Hall (1984) does not directly apply. The following lemma proves that the edge effect for *RankCover* converges to zero as n becomes large.

Lemma 2. *For* $\delta = O(\sqrt{n})$, $|T(\delta) - \hat{F}_{RG}(\delta)| \xrightarrow{a.s.} 0$ *as* $n \to \infty$.

Proof. We consider Euclidean distance as the distance metric. Let $\delta = O(\sqrt{n})$ be the radius of the disks, and $k = [\delta]$. For any circular disk lying partially outside $\mathcal{R}_n$, there exists a rectangle within which the circular portion can be inscribed. Considering the area of such rectangle as an upper bound for the area of the portion of the circle, we obtain, with probability 1,

$|T(\delta) - \hat{F}_{RG}(\delta)| \leq 4\{2\delta^2 + 2\delta(\delta-1) + 2\delta(\delta-2) + ... + 2\delta(\delta-k)\}$

$= 8\delta\{(k+1)\delta - \frac{k(k-1)}{2}\} = O(\frac{1}{\sqrt{n}}) \to 0$ as $n \to \infty$. □

One should note that such convergence is clearly quite slow and the sample size needs to be very large in order for the edge effect to be negligible for practical purposes.

2.6.3 Asymptotics of coverage for Boolean process

Hall (1985) proved the aymptotic normality of vacancy V for Boolean process and provided the expressions for its mean and variance. The expression for the mean and variance of the proportion of coverage C follows directly from those. For $\delta = \sqrt{n}$, the expressions are

$$E(C) = 1 - exp(\pi) \tag{2.8}$$

$$\sigma^2 = nV(C) = \pi e^{-2\pi}(8\int_0^1 u\{e^{2\pi J_k(u)} - 1\}du - \pi) = \pi e^{-2\pi}(8 \times 0.997216 - \pi), \tag{2.9}$$

where $J_k(u) = \frac{1}{\pi}(\frac{\pi}{2} - sin^{-1}(u) - \frac{1}{2}sin(2sin^{-1}u))$.

It also follows that

$$\sqrt{n}(C - E(C)) \xrightarrow{d} N(0, \sigma^2). \tag{2.10}$$

However, these results do not directly apply to *RankCover*, and the difference might be substantial even for moderately large n (Figure 2.13).

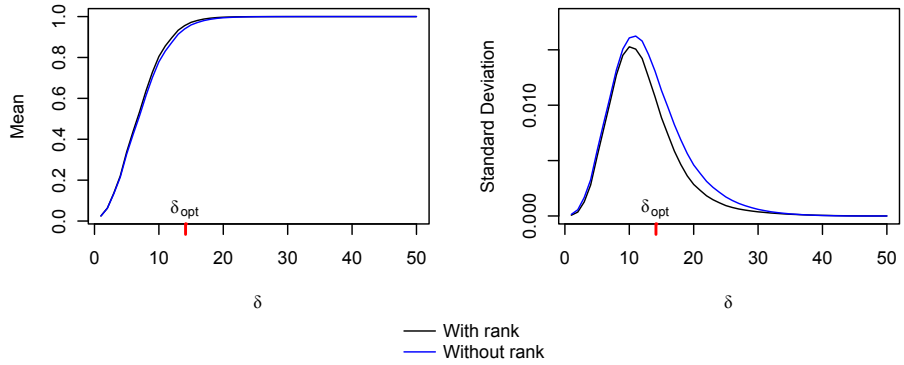

Figure 2.13: Showing the difference in mean and standard deviation between total coverage C for Boolean process and the *RankCover* statistic.

2.6.4 Applicability of the results to *RankCover*

If it can be shown that the difference between the total coverage as in Hall (1985) and the *RankCover* statistic becomes negligible as n becomes large, then Equation 2.8, Equation 2.9 and Equation 2.10 can be conveniently used for large n to test for general association.

Let us examine the difference between the joint distributions of $((x_1, y_1), ..., (x_n, y_n))$ under the null in both cases. If $((x_1, y_1), ..., (x_n, y_n))$ are independent samples from a bivariate discrete uniform distribution over $\{1, 2, ..., n\} \times \{1, 2, ..., n\}$, the joint density is

$$f_1((x_1, y_1), ..., (x_n, y_n)) = \frac{1}{n^{2n}}. \tag{2.11}$$

If $((x_1, y_1), ..., (x_n, y_n))$ are the ranks, the joint density becomes

$$f_2((x_1, y_1), ..., (x_n, y_n)) = \frac{1}{n!^2}. \tag{2.12}$$

The Hellinger distance between the two distributions is $H(f_1, f_2) = \sqrt{1 - \sqrt{\frac{n!^2}{n^{2n}}}} \to 1$ as $n \to \infty$. Therefore, the effect of rank does not wash away as n becomes large. However, the effect of rank on the test statistic might still be asymptotically negligible. But, it is difficult to prove or disprove it analytically.

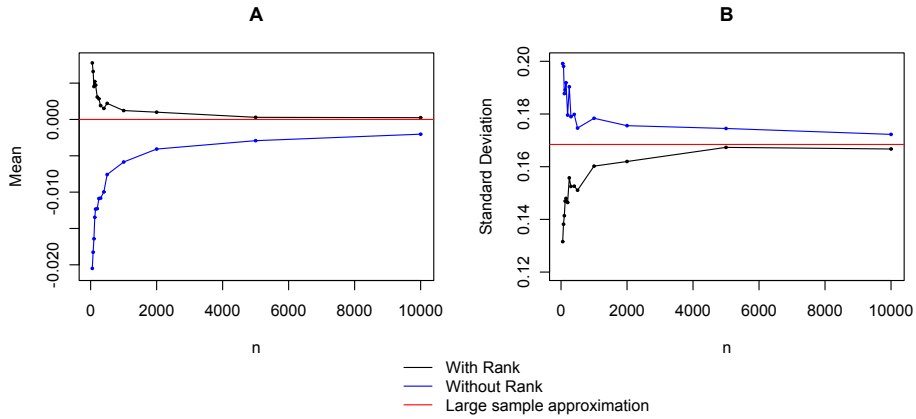

Figure 2.14: Showing the **A.** mean and **B.** standard deviation of $\sqrt{n}(C - E(C))$ for Boolean process and the corresponding statistic for $\hat{F}_{RG}(\delta)$.

To see the differences, we examined the behavior of coverage proportion C as in Hall (1985) and the *RankCover* test statistic for simulated datasets (Figure 2.14, Figure 2.15). Figure 2.14 indicates that the expectation and variance of C and $\hat{F}_{RG}(\delta)$ might be sufficiently close for very large n, but there is no conclusive proof. By Lemma 2, this implies that C and $T(\delta)$ might also be close asymptotically. However, it requires even larger sample size for them to be close enough (Figure 2.15). Based on Figure 2.14, we suggest that for sample sizes in the range 2000-10000, $\hat{F}_{RG}(\delta)$ can be used as the test statistic and the asymptotic results in Equation 2.10 hold approximately true. Based on our simulations, the type-I errors using Equation 2.10 for

$n = 2000, 5000$ and 10000 were $0.045, 0.054$ and 0.053.

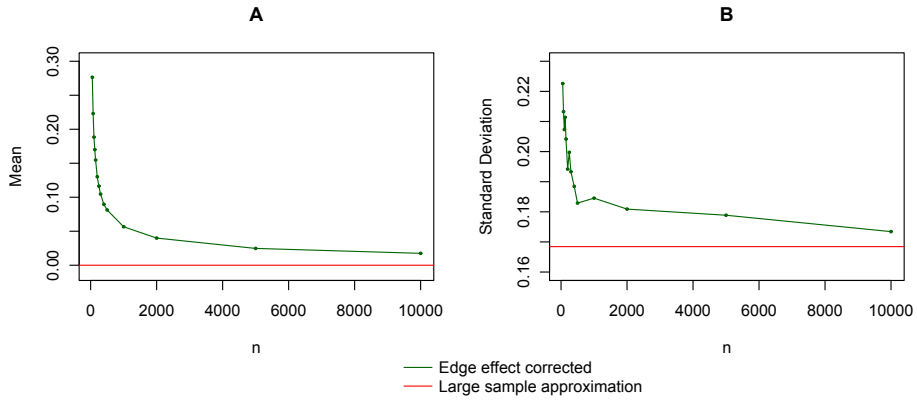

Figure 2.15: Showing the **A.** mean and **B.** standard deviation of $\sqrt{n}(C - E(C))$ for Boolean process and the corresponding statistic for $T(\delta)$.

2.7 Simulation Results

2.7.1 Comparison of different methods for simulated datasets

Following the simulation procedure used in Simon and Tibshirani (2014), we have simulated pairs of variables with several canonical dependency relationships (Figure 2.16) and with varying noise levels. In each scenario, the X values were simulated iid from a uniform distribution, while the noise distribution was Gaussian. However, the overall results were similar for other distributional forms.

The simulation results indicate that *RankCover* and dCor have some complementary characteristics, and so we additionally propose a hybrid statistic using results from *RankCover* and dCor. The hybrid method uses the minimum p-value from *RankCover* and rank-based dCor as a new statistic.

Figure 2.17 shows the power for the methods for various relationships, with varying noise levels, for sample size $n = 50$. Here the 'noise level' is a scale quantity appropriate to each relationship form, following Simon and Tibshirani (2014). It is evident that

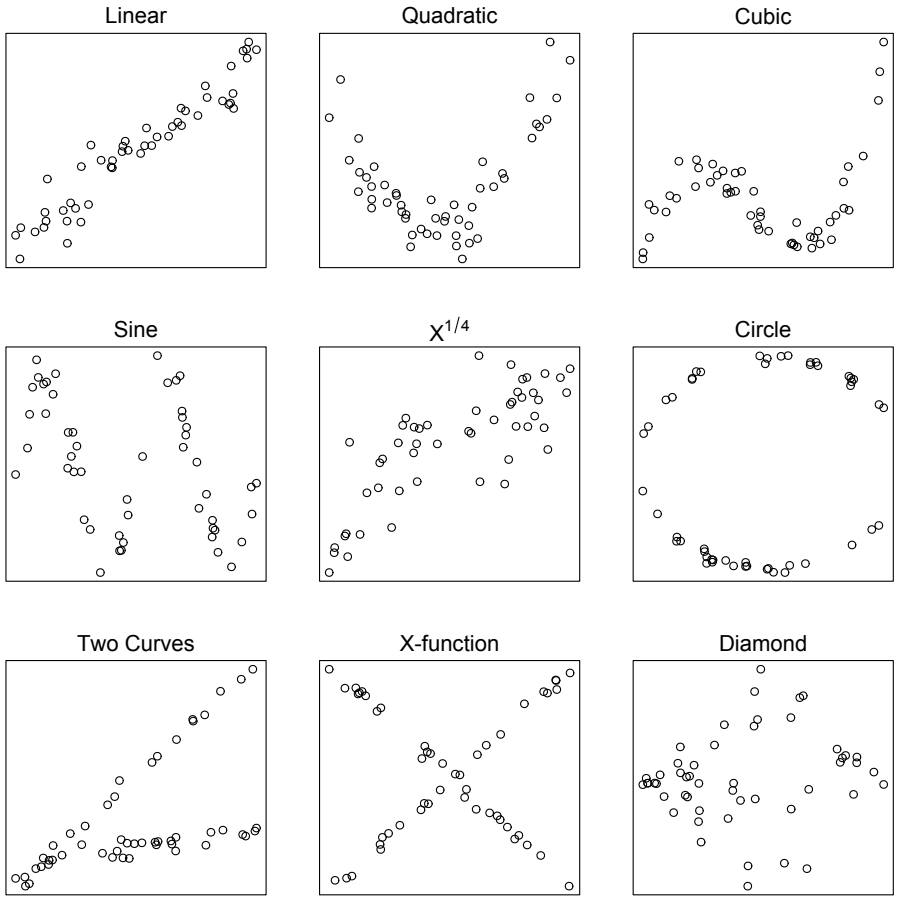

Figure 2.16: Showing the scatter plots for different relationships between the pair of variables (low noise level).

RankCover performs better than MIC in all the situations we have considered. It is found to be more powerful than dCor and HHG in several cases while these methods are found to be more powerful in other cases. Even when dCor or HHG is more powerful, *RankCover* still has reasonable power to identify the association. We have tested that these observations hold true for varying sample sizes, levels of noise, and functional forms for the originating X and noise distributions.

A careful look into the results indicate that dCor is more powerful than *RankCover* when the type of association is monotone. When the relationship is non-monotone, dCor is typically not as powerful. We attribute this behavior to the fact that dCor

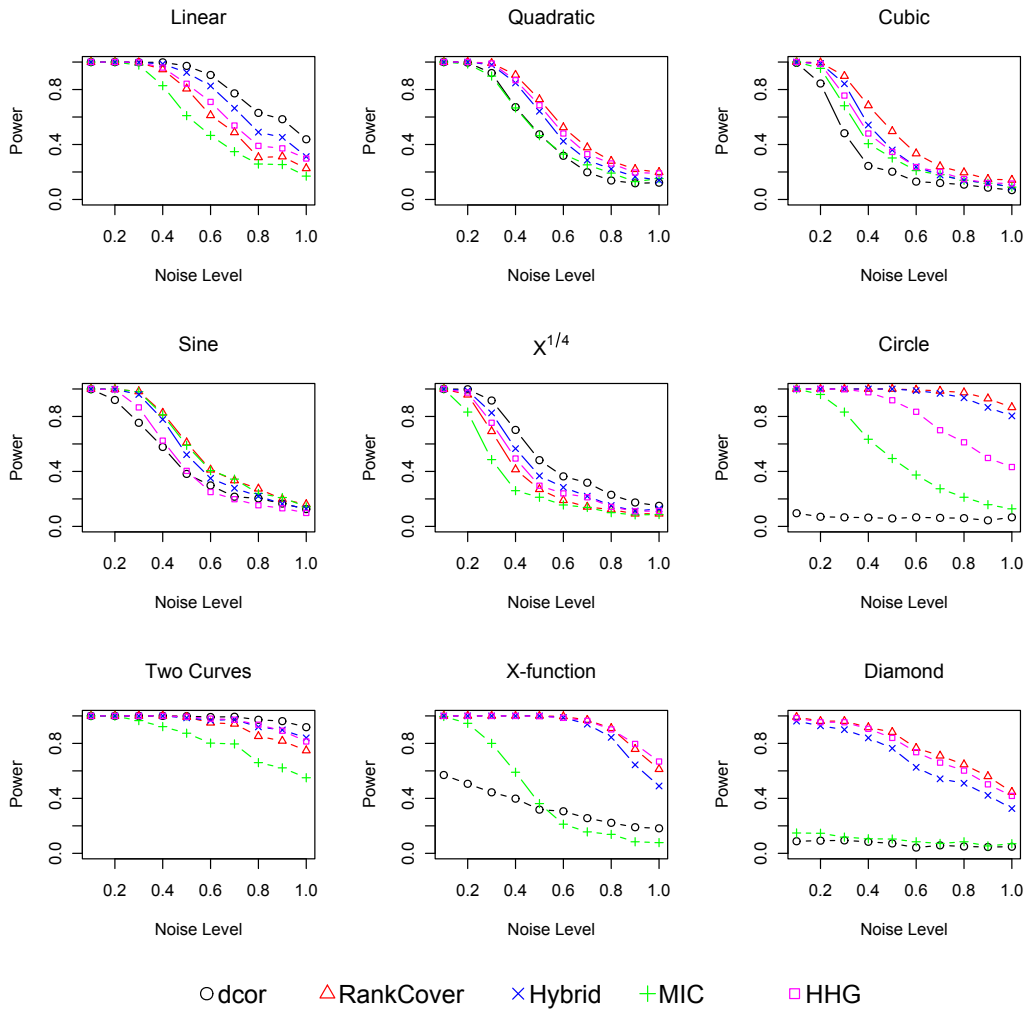

Figure 2.17: Showing the power of different methods (type-I $\alpha = 0.05$) against different relationships at varying noise levels (Manhattan distance), $n = 50$.

is less sensitive to non-monotone relationships for the reasons described earlier (Section 1.1.4). We have also observed that with monotone relationships, the Spearman's rank correlation is as powerful as dCor. Therefore, one might simply use Spearman's rank correlation if there is prior knowledge that the relationship is monotone. On the other hand, *RankCover* is more sensitive to local clustering of points rather than trends. Thus, it is powerful against even non-monotone relationships like cubic, circular or the "X" relationship.

These observations motivate the use of a hybrid method utilizing both *RankCover*

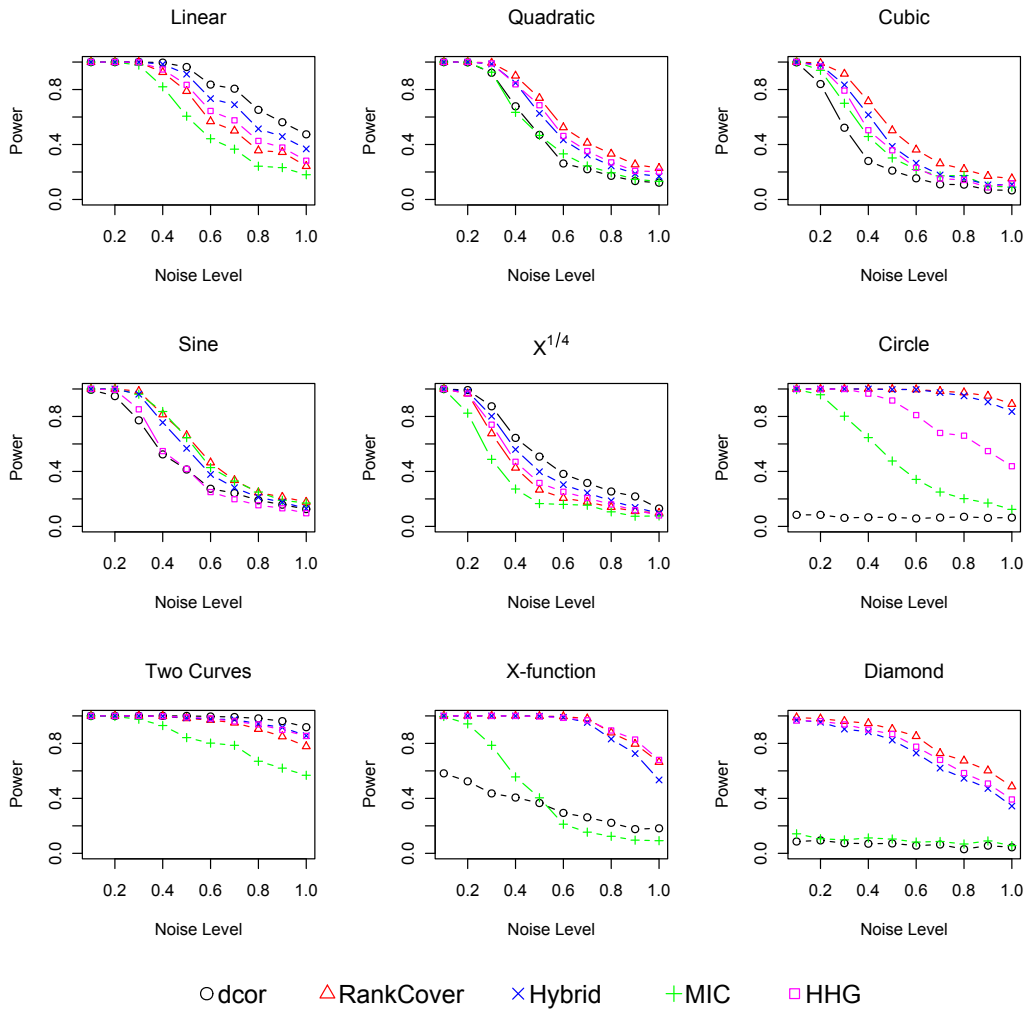

Figure 2.18: Showing the power of different methods (type-I $\alpha = 0.05$) against different relationships at varying noise levels (Euclidean distance), $n = 50$.

and dCor, as the two methods appear powerful in different situations. Formally, a new statistic is defined $s_{hybrid} = \min(p_{dCor}, p_{RankCover})$, where $p_{RankCover}$ is the p-value obtained by using *RankCover*, and p_{dCor} is that using dCor on $(rank(x), rank(y))$. The p-value for the hybrid method is $p_{hybrid} = P(S_{hybrid} \leq s_{hybrid})$. As with *RankCover*, the p-value can be obtained by using pre-computed simulations. The hybrid method, as expected, is always less powerful than the most powerful statistic for each scenario, but seems to be robust against all forms of association investigated.

The HHG method also appears to be relatively robust. However, the ability of

RankCover and the hybrid method to detect periodic relationships and non-functional relationships makes it very useful against such alternatives. The fact that *RankCover* is especially powerful against periodic relationships will be reinforced by the results in Section 2.8.3 and Section 2.8.4.

We summarize by emphasizing that *RankCover* and the hybrid method are powerful and robust in comparison to competing methods, and that these simulations cover a large range of relationships and noise levels. The broad conclusions are also not very sensitive to the marginal distributions of X and the error distributions.

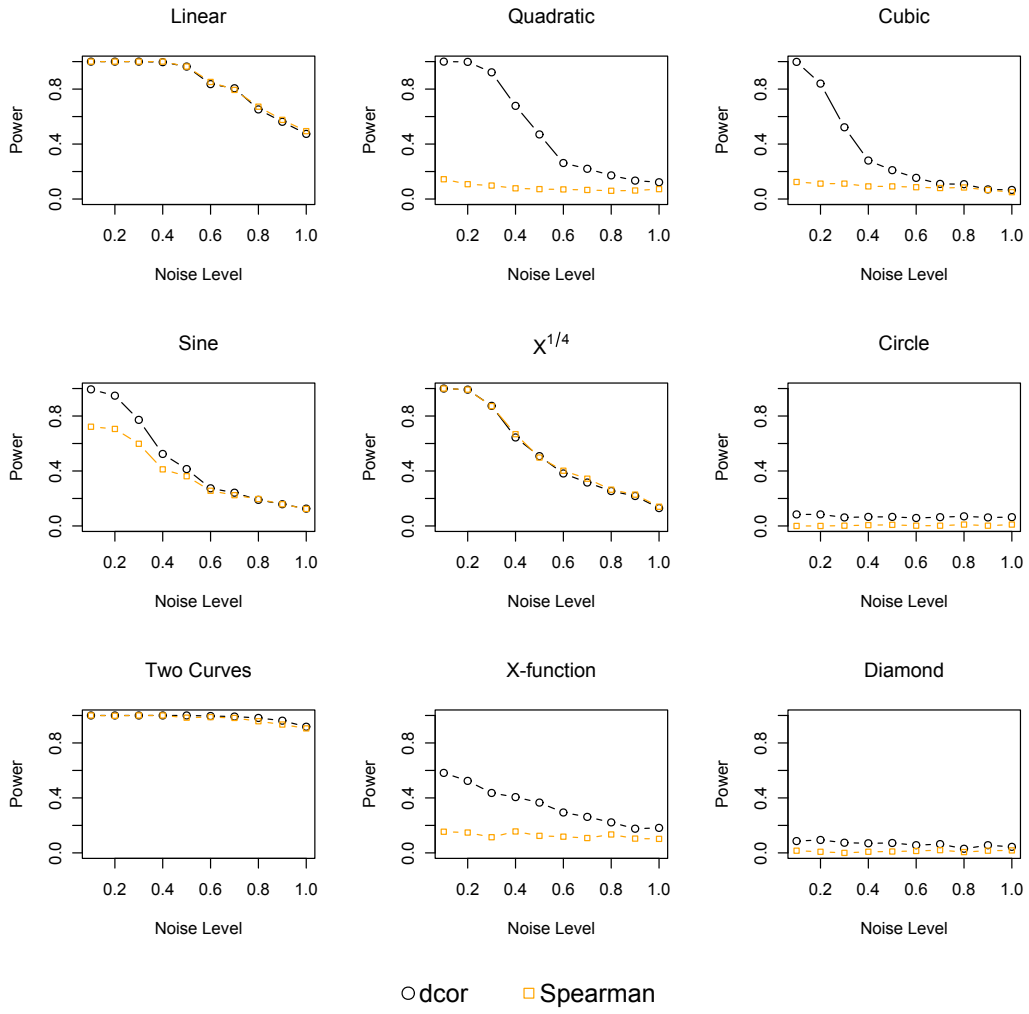

Figure 2.19: Showing the power comparison of dCor and Spearman's rank correlation

2.7.2 Comparison of dCor and Rank Correlation

Distance Correlation (dCor) seems to be the most powerful method among all the competing methods when the relationship is monotone (eg linear, $X^{1/4}$, Two curves). However, further simulations show that even Spearman's rank correlation is equally powerful in those cases (Figure 2.19). Therefore, if we have prior knowledge that the relationship is monotone, then we do not gain power by using the more recently developed methods anyway, and could use Spearman's rank correlation instead. We note that Spearman's rank correlation does not have much "generality" in the sense that it is not powerful against non-monotone alternatives. However, dCor has also been shown to have similar limitations.

2.8 Application on Real Data

In addition to simulated data, we illustrate all the approaches on several real datasets.

2.8.1 Example 1: Eckerle4 data

We show data from a study of circular interference transmittance (Eckerle 1979) from the NIST Statistical Reference Datasets for non-linear regression. The data were analyzed by Székely and Rizzo (2009) to illustrate dCor, and contain 35 observations on the predictor variable wavelength and the response variable transmittance.

Figure 2.20 shows the scatter plot of the predictor and the response along with the fitted curve (NIST StRD for non-linear regression) based on the model

$$y = \frac{\beta_1}{\beta_2} exp\{\frac{(x-\beta_3)^2}{2\beta_2^2}\} + \epsilon,$$

where $\beta_1, \beta_2 > 0, \beta_3 \in \mathbb{R}$ and ϵ is random Gaussian noise.

From the plot, it is evident that there is a very strong non-linear relationship between the two variables. For dCor, $p = 0.02072$, while MIC and HHG have p-values $< 10^{-5}$.

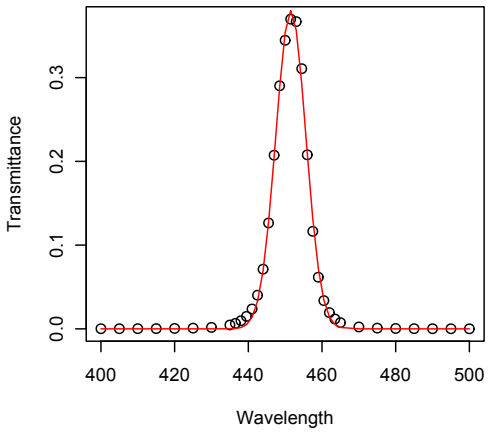

Figure 2.20: Showing the scatter plot and the fitted curve for the Eckerle4 dataset

The *RankCover* method and the hybrid method are also highly significant, with $p < 10^{-5}$.

2.8.2 Example 2: Aircraft data

We have explored the Saviotti aircraft data (Saviotti 1996) which was also analyzed by Székely and Rizzo (2009). We consider the wing span (m) vs. speed (km/h) ($n = 230$, Bowman and Azzalini (1997)). Figure 2.21 shows the scatter plot of the two variables, alongside non-parametric density estimate contours (log scale). It is clear from the plot that there is a non-linear relationship (Pearson's product moment correlation is a modest 0.0168, p-value= 0.8001), although the relationship is complicated and apparently not monotone.

All of the methods described here were significant at $\alpha = 0.05$. The p-values for dCor, MIC, and HHG were 0.00013, 0.00004, and $< 10^{-5}$, respectively. For *RankCover* the test was also significant with a $p = 0.0008$, and for the hybrid method $p = 0.0002$.

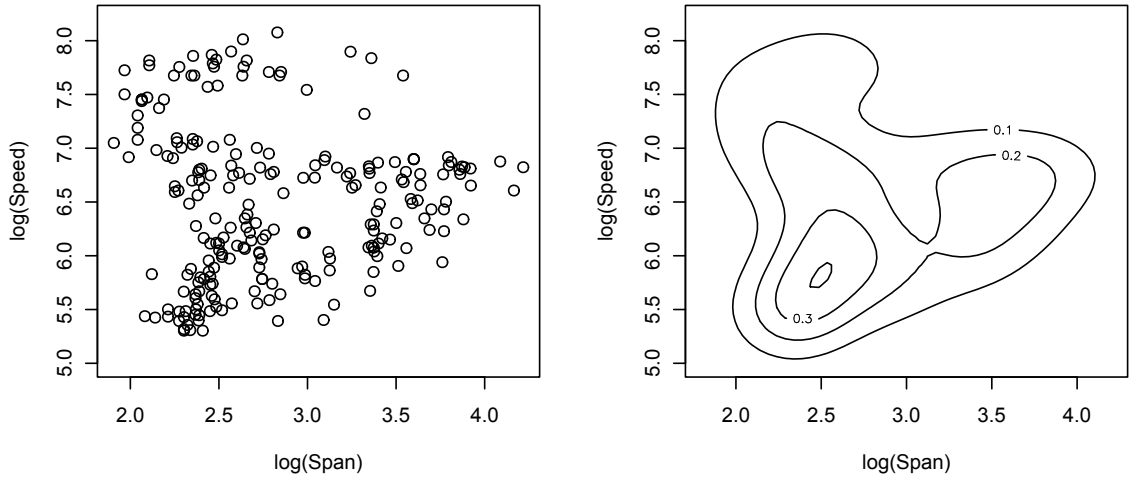

Figure 2.21: Showing the scatter plot and the density estimate contours for the aircraft speed and wing span

2.8.3 Example 3: ENSO data

The ENSO data (also taken from the NIST Statistical Reference Datasets for non-linear regression) consists of monthly average atmospheric pressure differences between Easter Island and Darwin, Australia (Kahaner et al. 1989), with 168 observations. There are 168 observations.The data form a time series, and has different cyclical components which were modeled (NIST StRD for non-linear regression) by the proposed model

$$y = \beta_1 + \beta_2 cos(\tfrac{2\pi x}{12}) + \beta_3 sin(\tfrac{2\pi x}{12}) + \beta_5 cos(\tfrac{2\pi x}{\beta_4}) + \beta_6 sin(\tfrac{2\pi x}{\beta_4}) + \beta_8 cos(\tfrac{2\pi x}{\beta_7}) + \beta_9 sin(\tfrac{2\pi x}{\beta_7}) + \epsilon,$$

where $\beta_1, \beta_2, ..., \beta_9 \in \mathbb{R}$ and ϵ is random Gaussian noise.

Figure 2.22 shows the scatter plot of the data along with the fitted curve. The cyclical fluctuations are evident, but no linear trend is observed. Thus, the Pearsonian correlation (0.0843) fails to capture the pattern. However a simple serial correlation with lag 1 (0.6102) reveals the association. With 100,000 simulations, the *RankCover* test is significant with p-value 0.00032. The hybrid test and MIC test are also significant

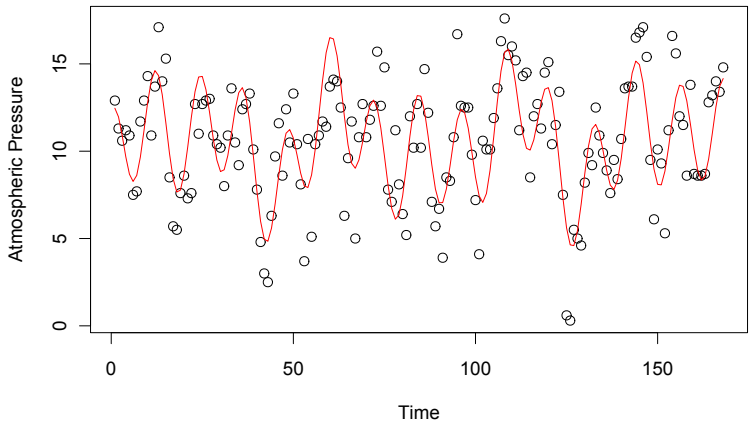

Figure 2.22: Showing the scatter plot and the fitted curve for the ENSO dataset

with p-values 0.00064 and 0.00027 respectively. However dCor and HHG fail to detect significant association (p-values 0.13521 and 0.07617, respectively).

2.8.4 Example 4: Yeast data

In this example, we analyze a yeast cell cycle gene expression dataset with 6223 genes Spellman et al. (1998). The experiment was designed to identify genes with activity varying throughout the cell cycle (Spellman et al. 1998), and thus transcript levels would be expected to oscillate. This data has been analyzed by many researchers, including Reshef et al. (2011), who used it to verifying the ability of MIC to detect oscillating patterns. We have run dCor, MIC, HHG, *RankCover* and the hybrid methods of test on the data and used the Benjamini-Hochberg method to control the false discovery rate.

We have listed the genes identified by different methods after controlling the false discovery rate (FDR) at the 5% level and compared them with the list of genes identified by Spellman et al. (1998). Of all the genes identified by Spellman et al. (1998), *RankCover* found 16% to be significant, while dCor, MIC and HHG found only 6%, 2% and 8% respectively. The hybrid method could identify 12% of those genes. Instead

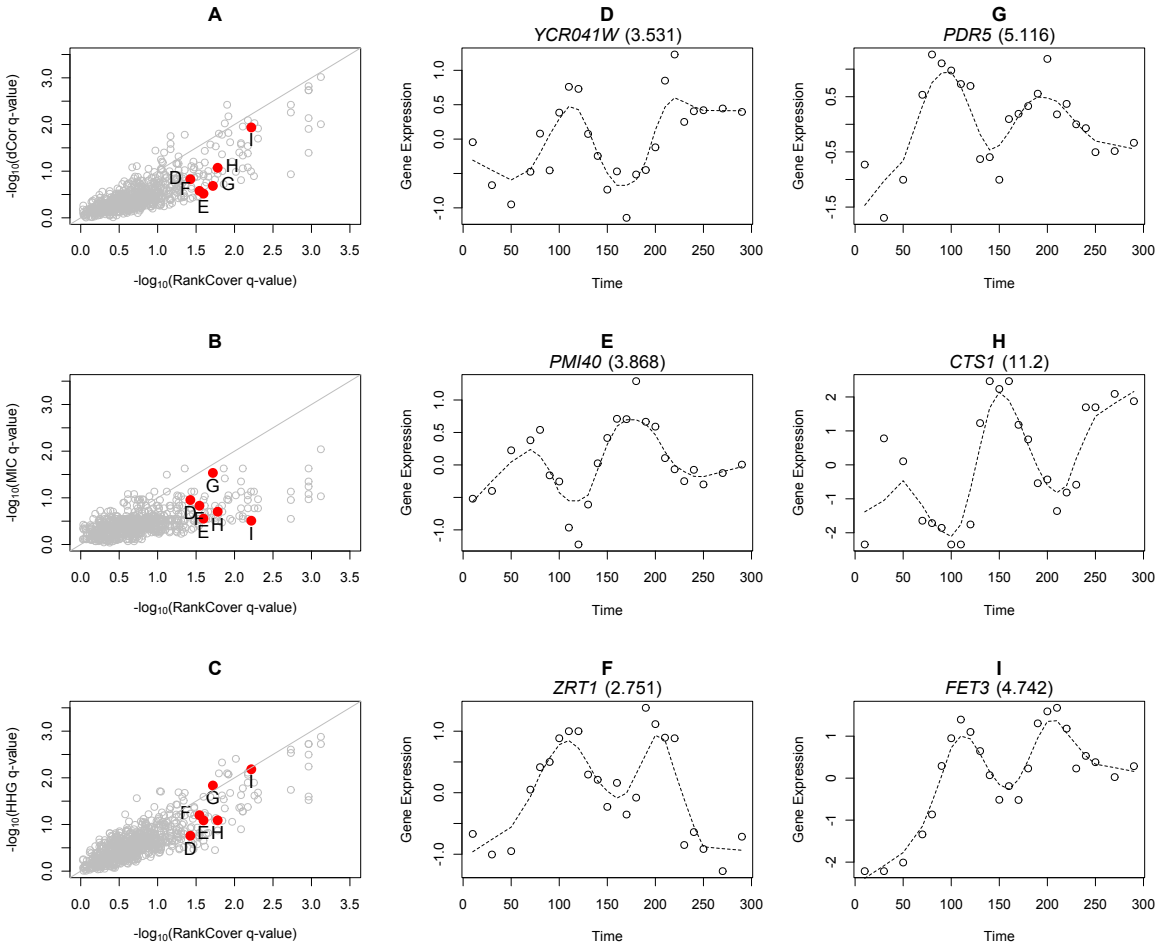

Figure 2.23: **A.** The plot comparing the FDR adjusted q-values of the test using *RankCover* and that using dCor for the genes in Spellman's list in a log scale. It is evident that most of the genes in Spellman's list have a smaller q-value when the *RankCover* test is used. **B.** A similar plot comparing the q-values of *RankCover* and MIC. **C.** A similar plot comparing the q-values of *RankCover* and HHG. **D-I.** Examples of genes in the Spellman's list that were identified by *RankCover*, but not by at least one of dCor, MIC or HHG. The values in parentheses are the Spellman scores for the genes.

controlling the FDR at 25%, the figures for HHG, dCor, MIC, *RankCover* and the hybrid method become 39%, 23%, 18%, 57% and 47% respectively.

For these data, *RankCover* was clearly successful at identifying oscillating patterns expected for the experiment. This is also clear from Figure 2.23 (panel A, B and C) which compares the FDR adjusted q-values of our *RankCover* test with those of dCor,

MIC and HHG on a logarithmic scale. Most of the genes in Spellman's list which were identified by dCor, MIC or HHG were also identified by *RankCover*, but *RankCover* identified more genes than the other methods. Figure 2.23 (panels D-I) shows some of the genes that were found significant by *RankCover* at 5% level, but not found significant by at least one of the other three methods. *PDR5* was found significant by MIC, HHG and *RankCover*, but not by dCor. On the other hand MIC could not identify *FET3*, which was identified by dCor, HHG and *RankCover*. The other four genes shown in Figure 2.23 were found significant by *RankCover* but not by dCor, MIC or HHG. Note that all of the six genes were found to be significant by the hybrid method.

2.9 Method to test the association of two variables after adjusting the effect of a third variable

The ideas of partial and multiple correlation coefficients do not easily generalize to the case of general association. Little work has been done in this area. Kendall (1942) discussed partial rank correlations and Moran (1951) proposed some methods to quantify partial and multiple rank correlations. However, the distribution of the statistics are difficult to obtain even in large samples (Maghsoodloo 1975). *RankCover* easily lends itself to the generalization to a multiple correlation analogue by computing the proportion of coverage in higher dimensions. The approach can have some usefulness in the theory of model selection, but an analogue of the partial correlation would be the more useful and interesting quantity.

The partial correlation coefficient is used to quantify and test the association of two variables after adjusting for other variables. However it applies only to linear associations. In the linear case, the correlation of two variables x and y for a fixed value of a third variable z does not depend on the fixed value. However, that may

not be true in the general case, which makes the situation more complicated (Speed 2011). The early works on this subject have either used Cochran-Mantel-haenszel type contingency table approach (Birch 1965), or are similar to rank correlation (Kendall 1942, Moran 1951). In both cases, the measures are expected to suffer for non-monotone relationships. Lehmann (1977) and Hubert (1985) discussed association and partial association in a more general set up, that also, is powerful only against monotone relationships. Recently, Qiuheng et al. (2014) proposed a method Partial Maximal Information Coefficient (PMIC) that attempts to fit a curve and the compute the MIC of the residuals. However, the model to be fitted is chosen separately on a case by case basis using other methodologies, and this defeats the idea of general association. Szekely et al. (2014) defined Partial Distance Correlation (pdCor) by introducing a Hilbert space and also proposed a method to test if the pdCor is significantly different from 0. Since dCor has been shown to suffer from lack of power to detect non-monotone relationships, pdCor is expected to have similar problems.

Using *RankCover*, we propose a general test of association after controlling the effect of a third variable. It can be generalized to more variables.

Our method consists of calculating the test statistic $T(\delta)$ for a number of strata and take the average of them. The strata are formed by different ranges of values of the third variable that is believed to be controlling the two variables of interest. For a fixed stratum size, s, we sort our observations in order of the values of the third variable and classify the first s observations to the first stratum, the next s observations to the next stratum and so on. In order to do the hypothesis test, we permute the ranks of x and y within each stratum.

The choice of s is vital. If s is too large, ie the number of strata is very small, the procedure will not be able to control the type-I error since there will be some association within each stratum between the two variables due to the effect of the third variable. On the other hand, a very small value of s will lead to loss of power. A value of s

which controls the type-I error at desired level and maximizes the power should be the optimal one.

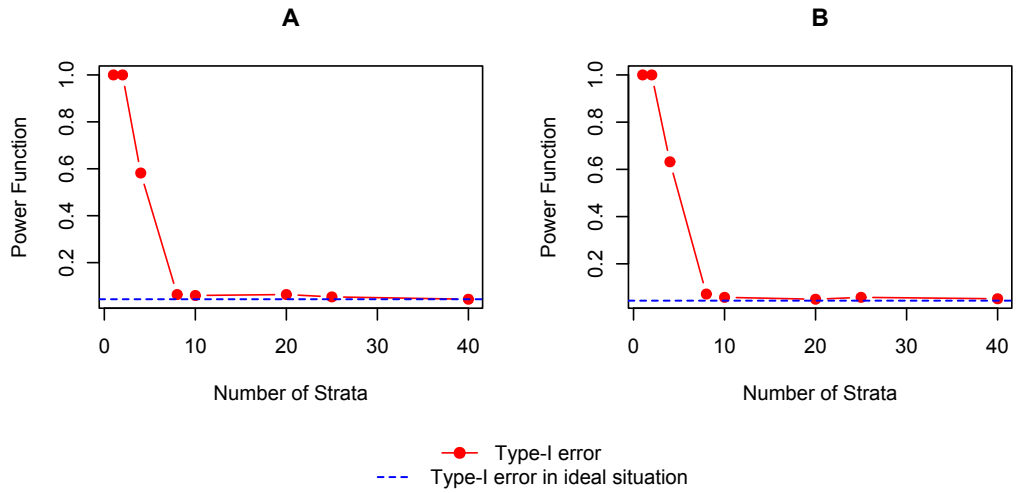

Figure 2.24: Showing the effect of number of strata on the type-I error of stratified approach. The horizontal line is the type-I error of the *RankCover* test in the ideal situation where one knows the exact form of x-z and y-z dependence. **A.** x-z and y-z are linear **B.** x-z is linear and y-z is quadratic.

Here we present results for simulated data with a sample size 200. We considered six different cases for the marginal relationships between x, y and z:

1. x-y, x-z and y-z are linear, all the slopes have the same sign.

2. x-y is quadratic, x-z and y-z are linear.

3. x-y is circular, x-z and y-z are linear.

4. x-y is circular, x-z is linear and y-z is quadratic.

5. x-y is $X^{\frac{1}{4}}$, x-z is linear and y-z is quadratic.

6. x-z and y-z are linear with positive slopes, x-y is linear with a negative slope.

In order to test how the type-I error is controlled as s is decreased, we used the cases where x and y are conditionally independent given z, and (i) x-z and y-z are linear or (ii) x-z is linear and y-z is quadratic.

These examples cover different situations such as (a) the association x and y is enhanced by their relationship with z (1,2,3 above), (b) the association may not be enhanced, but is of a different shape (4,5 above), (c) the association is masked by the effect of z (6 above), (d) there is no association between x and y, but spurious association is introduced by the effect of z.

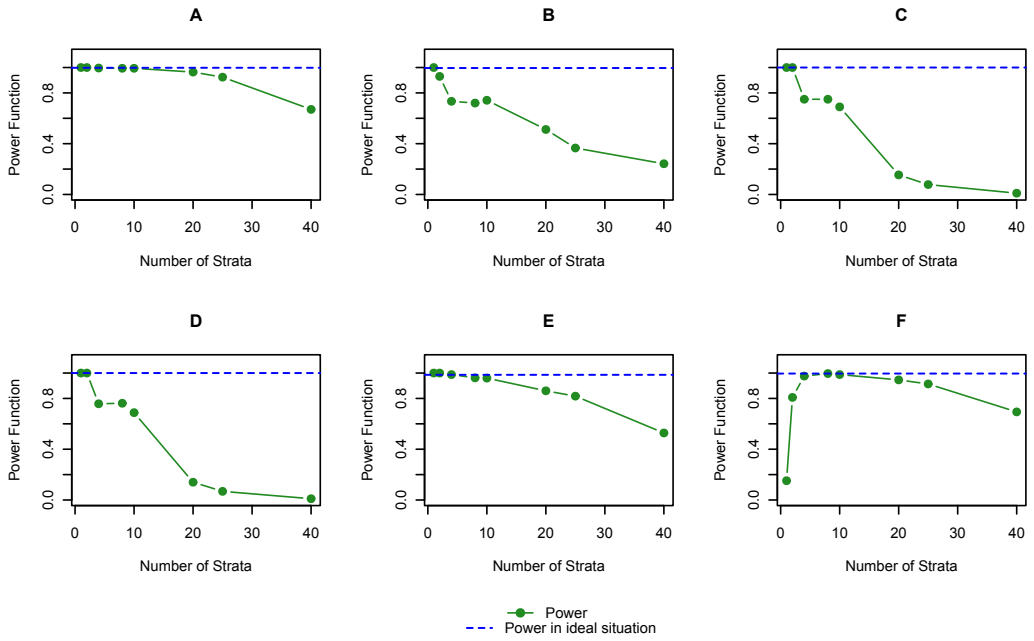

Figure 2.25: Showing the effect of number of strata on the power of stratified approach. The horizontal line is the power of the *RankCover* test in the ideal situation where one knows the exact form of x-z and y-z dependence. **A.** x-y, x-z and y-z are linear, all the slopes have the same sign **B.** x-y is quadratic, x-z and y-z are linear **C.** x-y is circular, x-z and y-z are linear **D.** x-y is circular, x-z is linear and y-z is quadratic **E.** x-y is $X^{\frac{1}{4}}$, x-z is linear and y-z is quadratic **F.** x-z and y-z are linear with positive slopes, x-y is linear with a negative slope.

Figure 2.24 shows the type-I error of the test against the number of strata. Figure 2.25 shows the power of the test for different situations. The power (or type-I error) for the ideal situation where one knows the exact form of x-z and y-z dependence is also presented. It is obvious from the figure that the power of the test decreases with the increase in the number of strata. However, if z masks the association of x and y, then the power increases initially and decreases when stratum size becomes very

small. The power can drop drastically as compared to the ideal situation especially when the x-y relationship is non-linear. Fortunately, the type-I error is controlled with a few strata (in these cases, 10). The choice of optimal number of strata for various situations requires further studies on this topic.

2.10 Discussion and future work

Our *RankCover* testing procedure serves as a simple and powerful method to test for general association between a pair of variables. The method is applicable to the problem of testing general association irrespective of the marginal distributions of the (continuous) variables. Use of the rank scale also allows a pre-computed null distribution for the statistic, avoiding the need for actual permutation. This, along with the introduction of the idea of using a single disk size, makes the procedure computationally feasible. The testing procedure has been shown to be powerful in simulated datasets even with a small sample size. A variety of real datasets, ranging from studies of cell cycle effects in gene expression to studies involving circular interference transmittance show that the approach provides useful and interpretable results.

Although dCor is theoretically motivated by consideration of characteristic functions, in practice it suffers for non-monotone relationships. Our *RankCover* procedure is generally powerful and robust, and is more powerful than MIC, dCor and HHG for a number of scenarios. *RankCover* may be especially useful to detect oscillating relationships, keeping in mind that such relationships need not be periodic and the amplitudes may vary. A hybrid of *RankCover* and dCor is proposed, which is shown to be highly robust against many forms of associations.

With the rapid rise of large datasets in today's scientific community, *RankCover* provides a useful tool to detect general association. The approach is both sensitive and relatively powerful, even with small samples, against various and general forms of

association.

We have demonstrated that when the sample size is very large, the large sample distributions of coverage for Boolean coverage process can be used as the null distribution of *RankCover* without edge effect correction, thus avoiding the need for permutation.. However, we have not been able to provide an analytical proof and further research is required. Also, central limit theorems related to the coverage process for ranks might be pursued independently.

We have also proposed a partial *RankCover* technique that is shown under different situations to control the type-I error and at the same time have reasonable power to detect the association after removing the effect of a confounder. However, the choice of the stratum size is critical to strike this balance. Also, the procedure to form the strata for more than one covariate is unclear, unless the sample size is sufficient to allow for stratification by multiple variables.. Hence, in our future work, it might be interesting to find a way to determine optimum stratum size for a given dataset and try to define the test statistic in a definitive way for more than one covariates.

CHAPTER 3: CONTROL OF FALSE DISCOVERIES IN GROUPED HYPOTHESIS TESTING FOR EQTL DATA

Expression quantitative trait loci (eQTL) analysis aims to detect the loci that affect the expression of one or more genes. The gene expression is considered as the quantitative trait potentially associated with the genotypes at different sites in the genome that are usually various single nucleotide polymorphisms (SNPs). As mentioned in Chapter 1, even though there has been substantial literature on both eQTL mapping and grouped hypothesis testing, consideration of the natural grouping in the eQTL data is comparatively unexplored. Analysis of gene-level eQTLs and specifying causal SNPs is an important biological problem. Testing whether there is any eQTL in an entire gene after controlling FDR for multiple genes may be interesting for various reasons. In the following sections, we discuss the structure of the eQTL data and how the grouped nature can be accounted for using a random effects model. We consider only the case of a cis-eQTL, i.e. when the variant affecting the gene expression is in the immediate neighborhood of the gene.

3.1 Structure of the eQTL data and the hypotheses

The eQTL data is usually in the form of an expression matrix consisting of a number of genes (say N) along with a genotype matrix which has genotypes of the same samples for several SNPs. Suppose that the number of samples is n and let the expression matrix be $Y_{N\times n}$. We can consider the genotype matrix $X^{(i)}_{m_i\times n}$, $i = 1, 2, ..., N$, corresponding to each gene by picking up the SNPs that are local to the gene. The genotype matrices are often adjusted for covariates, and thus can be considered to be continuous.

Let H_{0ij} denote the gene-SNP level null hypothesis that there is no eQTL at the jth SNP local to the ith gene, $j = 1, 2, ..., m_i, i = 1, 2..., N$. Therefore there are $\sum_{i=1}^{N} m_i$ gene-SNP level tests. These tests can be grouped into N groups corresponding to the N genes with m_i tests in the ith group. Define H_{0i} to be the gene level null hypothesis for the ith gene that there is no eQTL in the ith gene. Therefore the gene level hypothesis can be written as

$$H_{0i} = \cap_{j=1}^{m_i} H_{0ij}, \tag{3.1}$$

i.e. the gene level null requires that all m_i hypotheses be null.

3.2 The empirical Bayes set up

We adopt an empirical Bayes approach for controlling the FDR. Empirical Bayes approaches have been used in many genetic applications in recent times (Efron and Tibshirani 2002, Ferkingstad et al. 2008). The merit of using an empirical Bayes approach using the local false discovery rate (lfdr) instead of p-value based FDR controlling approaches has been discussed in Efron et al. (2001a) and Kendziorski et al. (2003). Let us define the lfdr corresponding to the gene level and gene-SNP level hypotheses respectively as

$$\lambda_i(Y_i, X^{(i)}) = P(H_{0i}|Y_i, X^{(i)}),\ i = 1, 2, ..., N, \tag{3.2}$$

and

$$\lambda_{ij}(Y_i, X_j^{(i)}) = P(H_{0ij}|Y_i, X_j^{(i)}),\ j = 1, 2, ..., m_i,\ i = 1, 2..., N, \tag{3.3}$$

where Y_i is the ith row of Y and $X_j^{(i)}$ is the jth row of $X^{(i)}$.

If we can obtain the lfdr λ_i for each of the gene level hypothesis, we can control the

FDR at target level α for gene-level testing using the following adaptive thresholding procedure which appears in Newton et al. (2004), Sun and Cai (2007), Cai and Sun (2009), Li et al. (2013).

1. Enumerate the index $i_1, i_2, ..., i_N$ of the genes such that $\lambda_{i_1} \leq \lambda_{i_2} \leq ... \leq \lambda_{i_N}$.

2. Reject hypotheses $H_{0i_1}, ..., H_{0i_L}$ where L is the largest integer such that

$$\frac{1}{L}\sum_{l=1}^{L} \lambda_{i_l} \leq \alpha.$$

Sun and Cai (2007) and subsequently Cai and Sun (2009) showed that the adaptive thresholding procedure is valid in the sense that it controls the FDR at target level α for an 'oracle' procedure where the true parameters of the model are assumed to be known. It is asymptotically valid for a 'data-driven' procedure when the parameters are consistently estimated from the data. Li et al. (2013) proved its validity under further relaxed conditions. The proof makes use of the following theorem (Averaging Theorem, Efron and Tibshirani (2002)).

Theorem 1. *Let $lfdr(z) = P(H_0|z)$ denote the lfdr for observed data z. Then, for a rejection region $\mathcal{R}$, the FDR will be given by*

$$FDR(\mathcal{R}) = P(H_0|Z \in \mathcal{R}) = E(lfdr(Z)|Z \in \mathcal{R})$$

A similar procedure can be used to control the FDR for gene-SNP level tests. In the next section, we suggest a model which enables us to calculate the lfdr's.

3.3 The Random Effects model and testing procedure for Group-level FDR control (*REG-FDR*)

Our *REG-FDR* is a model to obtain the gene-level lfdr's that can be subsequently used to test the gene level hypotheses after controlling the FDR using the adaptive

thresholding method. The model is based on the following assumptions.

1. For any gene i, under the gene level alternative hypothesis H_{0i}^c, there exists a single causal SNP that influences its expression. (A1)

2. Each of the m_i SNPs has equal probability to be the causal SNP. (A2)

We will use the above assumptions throughout even though the assumption (A2) can be relaxed if required. One might use some other probability distribution over the SNPs instead of the uniform distribution if there is prior knowledge about the distribution. Assumption (A1) may not always be valid, however such an assumption is not uncommon (Kendziorski et al. 2006, Gelfond et al. 2007, Ardlie et al. 2015). Under these assumptions, the gene level lfdr for the ith gene has the following form.

$$\lambda_i(Y_i, X^{(i)}) = P(H_{0i}|Y_i, X^{(i)}) = \frac{\pi_0 f_0(Y_i)}{\pi_0 f_0(Y_i) + (1-\pi_0)\frac{1}{m_i}\sum_{j=1}^{m_i} f_1(Y_i|X_j^{(i)}, \beta)}, \tag{3.4}$$

where $\pi_0 = P(H_{0i})$, $f_0(Y_i)$ is the density of Y_i under the null and $f_1(Y_i|X_j^{(i)}, \beta)$ is the conditional density under the alternative given that the jth SNP is causal. The marginal density $p(X^{(i)})$ is cancelled from numerator and denominator. Importantly, this cancellation allows us to bypass the modeling of the dependence structure of the SNPs which might have been difficult to estimate.

We assume that $f_0(.)$ is the density of $N_n(0, I_n)$ and $f_1(.|X_j^{(i)}, \beta)$ is the density of $N_n(\beta X_j^{(i)}, (1-\beta^2)I_n)$, and β is the correlation between Y_i and $X_j^{(i)}$. The choice of this density ensures that the unconditional variance of Y_i is free of β. To take care of the variability across the genes, we assume β to be a random effect such that $\sqrt{n-3}\ tanh^{-1}(\beta)$ has a $N(0, \sigma^2)$ distribution. Since β is a correlation coefficient, the Fisher transformation is used to ensure that the variance does not depend on the mean.

Similarly, the gene-SNP level lfdr for the jth SNP local to the ith gene is given by

$$\lambda_{ij}(Y_i, X_j^{(i)}) = P(H_{0ij}|Y_i, X_j^{(i)}) = \frac{\tilde{\pi_0} f_{0ij}(Y_i|X^{(i)})}{\tilde{\pi_0} f_{0ij}(Y_i|X^{(i)}) + (1-\tilde{\pi_0}) f_1(Y_i|X_j^{(i)})}, \tag{3.5}$$

where $\tilde{\pi_0} = P(H_{0ij})$, $f_{0ij}(.)$ is the density under H_{0ij}. Under the assumption that $X_j^{(i)}$'s for varying j's are related by an AR(1) structure with serial correlation ρ, it can be shown that

$$f_{0ij}(Y_i|X^{(i)}) = \theta f_0(Y_i) + (1-\theta) \sum_{k \neq j} f_{2jk}(Y_i|X_j^{(i)}, \beta, \rho), \tag{3.6}$$

where $\theta = P(H_{0i}|H_{0ij}) = \frac{\pi_0}{\pi_0 + \frac{m_i - 1}{m_i}(1-\pi_0)}$, and $f_{2jk}(.|X_j^{(i)}, \beta, \rho)$ is the probability density of $N_n(\beta\rho^{|k-j|}X_j^{(i)}, (1-\beta^2\rho^{2|k-j|})I_n)$. However, this assumption is not necessary for the estimation of π_0, σ and the gene level lfdr.

We can estimate the parameters π_0 and σ using a maximum likelihood approach and plug the estimates into Equation 3.4 or Equation 3.5. This enables us to use the adaptive thresholding procedure for carrying out the tests with proper control of the FDR. Note that we cannot bypass the modeling of the dependence structure of the SNPs in order to obtain the λ_{ij}'s. However, simulations show that when the dependence is not very strong, $f_0(.)$ can be used as an approximation of $f_{0ij}(.)$.

3.4 An EM algorithm to estimate *REG-FDR* parameters

The log-likelihood for *REG-FDR* is

$$L(\pi_0, \sigma|X, Y) = log(p(X)) + \sum_{i=1}^{N} log[\pi_0 f_0(Y_i) + (1-\pi_0)\frac{1}{m_i}\sum_{j=1}^{m_i} f_1(Y_i|X_j^{(i)}, \sigma)]$$

where $p(X)$ is the marginal density of X that we avoid modelling, but assume to be

free of π_0 and σ. We introduce the following unobserved variables.

$\delta_i = 1$ or 0 according as the ith gene has an eQTL or not, $i = 1, 2, ..., N$.

$S_{ij} = 1$ or 0 according as the jth SNP local to the ith gene is causal or not, $j = 1, 2, ..., m_i$.

Given the data (X, Y), δ_i follows $Bernoulli(\pi_0)$. Given the data and $\delta_i = 1$, $(S_{i1}, S_{i2}, ..., S_{im_i})$ follows a $Multinomial(1; 1/m_i, 1/m_i, ..., 1/m_i)$ distribution.

Now the complete log-likelihood becomes

$$
\begin{aligned}
&L_c(\pi_0, \sigma | X, Y, \delta, S) \\
&= \log(p(X)) + \textstyle\sum_{i=1}^{N} \log[(\pi_0 f_0(Y_i))^{(1-\delta_i)}((1-\pi_0)\frac{1}{m_i}\prod_{j=1}^{m_i} f_1(Y_i|X_j^{(i)}, \sigma)^{S_{ij}})^{\delta_i}] \\
&= \log(p(X)) + \textstyle\sum_{i=1}^{N} [(1-\delta_i)\log(\pi_0) + \delta_i \log(1-\pi_0)] + \sum_{i=1}^{N}\sum_{j=1}^{m} S_{ij}\delta_i log[f_1(Y_i|X_j^{(i)}, \sigma)]
\end{aligned}
$$

The M-step gives

$$\hat{\pi_0} = \frac{1}{N}\sum_{i=1}^{N}(1-\delta_i)$$

and

$$\hat{\sigma} = \underset{\sigma}{ArgMax} \sum_{i=1}^{N}\sum_{j=1}^{m} S_{ij}\delta_i \log[f_1(Y_i|X_j^{(i)}, \sigma)]$$

In the kth iteration, the E-step replaces δ_i by $E(\delta_i|X, Y, \hat{\pi}_0^{(k-1)}, \hat{\sigma}^{(k-1)})$ and $S_{ij}\delta_i$ by $E(S_{ij}\delta_i|X, Y, \hat{\pi}_0^{(k-1)}, \hat{\sigma}^{(k-1)})$. These are given by

$$E(\delta_i|X, Y, \hat{\pi}_0^{(k-1)}, \hat{\sigma}^{(k-1)}) = \frac{\hat{\pi}_0^{(k-1)} f_0(Y_i)}{\hat{\pi_0}^{(k-1)} f_0(Y_i) + (1-\hat{\pi_0}^{(k-1)})\frac{1}{m_i}\sum_{j=1}^{m_i} f_1(Y_i|X_j^{(i)}, \hat{\sigma}^{(k-1)})}$$

and

$$E(S_{ij}\delta_i|X, Y, \hat{\pi}_0^{(k-1)}, \hat{\sigma}^{(k-1)}) = E(\delta_i|X, Y, \hat{\pi}_0^{(k-1)}, \hat{\sigma}^{(k-1)}) \times \frac{f_1(Y_i|X_j^{(i)})}{\sum_{t=1}^{m_i} f_1(Y_i|X_t^{(i)}, \hat{\sigma}^{(k-1)})}$$

The updating continues until $|L(\hat{\pi}_0^{(k+1)}, \hat{\sigma}^{(k+1)}|X, Y) - L(\hat{\pi}_0^{(k)}, \hat{\sigma}^{(k)}|X, Y)|$ becomes sufficiently small.

3.5 The *Z-REG-FDR* model

One computational challenge with the REG-FDR model is that the density $f_1(Y_i|X_j^{(i)})$ doesn't have a closed form expression. It can be expressed as the following integral.

$$f_1(Y_i|X_j^{(i)}) = \int_{-1}^{1} f_1(Y_i|X_j^{(i)}, \beta) \frac{\sqrt{n-3}}{\sqrt{2\pi}\sigma(1-\beta^2)} e^{-\frac{n-3}{2\sigma^2}\{tanh^{-1}(\beta)\}^2} \tag{3.7}$$

The maximum likelihood estimation becomes computationally burdensome if the integral is evaluated using numerical quadrature. We propose an alternative model entitled *Z-REG-FDR* which avoids this problem. In this approach, we consider the Fisher transformed and scaled z-statistics as our data. Thus, for each gene i, we have a vector of z-statistics

$$z^{(i)} = (z_1^{(i)}, z_2^{(i)}, ..., z_{m_i}^{(i)}),\ i = 1, 2, ..., N,$$

where $z_j^{(i)} = \sqrt{n-3}\ tanh^{-1}(r_j^{(i)})$, $r_j^{(i)}$ being the sample correlation of Y_i and $X_j^{(i)}$.

The Fisher transformation and scaling ensures that $z^{(i)}$ is approximately normal and variance of each component is 1 under both null and alternative. Under the null, the mean of $z^{(i)}$ is zero.

We treat the z_i's as if they are independent across different genes. This assumption is realistic since very few genes share common SNPs. We keep our assumptions (A1) and (A2) of having only one causal SNP under the alternative which can be any one of the m_i SNPs with equal probability. Let the kth SNP be the causal one. Then, we assume the following.

1. The distribution of $(z_1^{(i)}, ..., z_{k-1}^{(i)}, z_{k+1}^{(i)}, ..., z_{m_i}^{(i)})$ given $z_k^{(i)}$ under the alternative is same as that under the null. (A3)

In particular, this assumption is true if the components of $z^{(i)}$ have a Markov dependence structure with the serial correlation being the same under null and alternative, which is true in the special case that the successive marker correlations are zero. In general, this assumption is obviously violated, but as shown in Section 3.6, the overall procedure appears to work well in many circumstances.

Under the above assumptions, we can write the joint distribution of the random vector $z^{(i)} = (z_1^{(i)}, z_2^{(i)}, ..., z_{m_i}^{(i)})$ as

$$f_1(z_1^{(i)}, z_2^{(i)}, ..., z_{m_i}^{(i)}) = p_1(z_k^{(i)}) f_{0|k}(z_1^{(i)}, ..., z_{k-1}^{(i)}, z_{k+1}^{(i)}, ..., z_{m_i}^{(i)}) \tag{3.8}$$

under the alternative, and

$$f_0(z_1^{(i)}, z_2^{(i)}, ..., z_{m_i}^{(i)}) = p_0(z_k^{(i)}) f_{0|k}(z_1^{(i)}, ..., z_{k-1}^{(i)}, z_{k+1}^{(i)}, ..., z_{m_i}^{(i)}) \tag{3.9}$$

under the null.

We assume $p_0(.)$ to be the density of $N(0,1)$ and $p_1(.)$ to be the density of $N(\mu,1)$ where μ is assumed to be random with a $N(0,\sigma^2)$ distribution. We do not assume anything about the form of $f_{0|k}$ except that it is multivariate normal and does not involve the other parameters, in this case π_0 and σ.

The gene level lfdr for this model reduces to

$$P(H_{0i}|z^{(i)}) = \frac{1}{1 + \frac{1-\pi_0}{\pi_0} \frac{1}{m_i} \sum_{k=1}^{m_i} \frac{p_1(z_k^{(i)})}{p_0(z_k^{(i)})}}, \ i = 1, 2, ..., N. \tag{3.10}$$

We have not modeled a part of the full likelihood $\prod_{i=1}^{N} (\pi_0 f_0(z^{(i)}) + (1-\pi_0) f_1(z^{(i)}))$. Instead we maximize $\prod_{i=1}^{N} \frac{\pi_0 f_0(z^{(i)}) + (1-\pi_0) f_1(z^{(i)})}{f_0(z^{(i)})}$. This is equivalent to the maximum likelihood estimation under the assumption that $f_{0|k}$ does not involve the parameters π_0 and σ. Note that we need to estimate only the parameters π_0 and σ to obtain the gene level lfdr using Equation 3.10.

Table 3.1 shows the results for simulated datasets (1000 simulations) where z's are directly simulated from an autoregressive structure. The estimates are accurate to within about 15% when the true σ is at least 2. The control of the FDR is also satisfactory.

True π_0	True σ	True ρ	Mean $\hat{\pi_0}$	Mean $\hat{\sigma}$	SE($\hat{\pi_0}$)	SE($\hat{\sigma}$)	Realized FDR(5%)	Realized FDR(10%)
0.20	1	0.10	0.2030	0.9964	0.1841	0.0823	0.0954	0.1236
0.20	2	0.10	0.1865	1.9660	0.0469	0.0374	0.0576	0.1136
0.20	5	0.10	0.1977	4.9383	0.0094	0.0306	0.0507	0.1014
0.20	1	0.50	0.1932	0.9919	0.1613	0.0757	0.0922	0.1252
0.20	2	0.50	0.1873	1.9663	0.0417	0.0352	0.0565	0.1121
0.20	5	0.50	0.1977	4.9383	0.0092	0.0303	0.0508	0.1013
0.20	1	0.80	0.1857	0.9875	0.1308	0.0664	0.0882	0.1245
0.20	2	0.80	0.1894	1.9673	0.0325	0.0317	0.0545	0.1090
0.20	5	0.80	0.1979	4.9388	0.0085	0.0292	0.0507	0.1012

Table 3.1: Showing summary of the simulation studies with directly simulated z from an AR(1) model with correlation ρ

When the required assumptions are not satisfied, this method can still be used as an approximate maximum likelihood approach. For instance, when the $X_j^{(i)}$'s are related by an AR(1) structure, it can be shown that the correlation between the z-statistics depends on the effect size, i.e. the correlation between Y_i and the causal SNP, hence violating the assumption (A3). The following lemma shows the extent to which the conditional distribution $f_{0|k}$ might depend on the effect size for any correlation structure among normally distributed SNPs. We use a trivariate normal distribution for illustration, as it is rich enough for demonstration while still analytically tractable.

Lemma 3. *Suppose* (X_1, X_2, X_3) *are jointly normal with mean* $(0, 0, 0)$ *and covariance matrix*

$$\begin{pmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_3 \\ \rho_2 & \rho_3 & 1 \end{pmatrix}.$$

Let $Y = \beta X_1 + \epsilon$, where $\epsilon \sim N(0, 1-\beta^2)$, and r_1, r_2, r_3 denote the sample product moment correlation coefficient of Y with X_1, X_2 and X_3 respectively for a sample of size n. The asymptotic correlations between these sample correlations are given by

$$Cor(r_1, r_2) = \rho_{12} = \frac{\rho_1(2 - \beta^2 - \beta^2\rho_1^2)}{2n(1 - \beta^2\rho_1^2)}$$

and

$$Cor(r_2, r_3) = \rho_{23} = \frac{2\rho 3 + \beta^2(\rho_1^2 + \rho_2^2)(\beta^2\rho_1\rho_2 - 2\rho 3) + \beta^2\rho_1\rho_2(\rho_3^2 - 1)}{2n(1 - \beta^2\rho_1^2)(1 - \beta^2\rho_2^2)},$$

ρ_{13} having the same form as ρ_{12}.

Proof. For the ith sample, let us define

$$\mathbf{Z}_i = (X_{1i}, X_{2i}, X_{3i}, Y_i, X_{1i}^2, X_{2i}^2, X_{3i}^2, Y_i^2, X_{1i}Y_i, X_{2i}Y_i, X_{3i}Y_i).$$

Clearly, $E(\mathbf{Z}_i) = \mu = (0, 0, 0, 0, 1, 1, 1, 1, \rho_1, \rho_2, \rho_3)$, and suppose $V(\mathbf{Z}_i) = \Sigma = (\sigma_{ij})_{11\times 11}$.

Define the functions g_1, g_2 and g_3, all $\mathbb{R}^{11} \to \mathbb{R}$, as

$$g_1(x_1, x_2, ..., x_{11}) = \frac{x_9 - x_1x_4}{\sqrt{(x_5 - x_1^2)(x_8 - x_4^2)}},$$

$$g_2(x_1, x_2, ..., x_{11}) = \frac{x_{10} - x_2x_4}{\sqrt{(x_6 - x_2^2)(x_8 - x_4^2)}},$$

$$g_3(x_1, x_2, ..., x_{11}) = \frac{x_{11} - x_3x_4}{\sqrt{(x_7 - x_3^2)(x_8 - x_4^2)}}.$$

Then, $r_1 = g_1(\bar{\mathbf{Z}})$, $r_2 = g_2(\bar{\mathbf{Z}})$ and $r_3 = g_3(\bar{\mathbf{Z}})$.

By the delta method,

$$\sqrt{n}(r_1 - \beta, r_2 - \beta\rho_1, r_3 - \beta\rho_2) \xrightarrow{d} N(\mathbf{0}, \Gamma),$$

where $\Gamma_{ij} = \sum_{k=1}^{11}\sum_{l=1}^{11} \sigma_{kl}\frac{\partial g_i}{\partial \mu_k}\frac{\partial g_j}{\partial \mu_l}; i = 1,2,3; j = 1,2,3.$

Now,

$$\frac{\partial g_1}{\partial \mu_1} = \frac{\partial g_1}{\partial \mu_2} = \frac{\partial g_1}{\partial \mu_3} = \frac{\partial g_1}{\partial \mu_4} = \frac{\partial g_1}{\partial \mu_6} = \frac{\partial g_1}{\partial \mu_7} = \frac{\partial g_1}{\partial \mu_{10}} = \frac{\partial g_1}{\partial \mu_{11}} = 0,$$

$$\frac{\partial g_1}{\partial \mu_5} = \frac{\partial g_1}{\partial \mu_8} = -\frac{1}{2}\beta,\ \frac{\partial g_1}{\partial \mu_9} = 1.$$

$$\frac{\partial g_2}{\partial \mu_1} = \frac{\partial g_2}{\partial \mu_2} = \frac{\partial g_2}{\partial \mu_3} = \frac{\partial g_2}{\partial \mu_4} = \frac{\partial g_2}{\partial \mu_5} = \frac{\partial g_2}{\partial \mu_7} = \frac{\partial g_2}{\partial \mu_9} = \frac{\partial g_2}{\partial \mu_{11}} = 0,$$

$$\frac{\partial g_2}{\partial \mu_6} = \frac{\partial g_2}{\partial \mu_8} = -\frac{1}{2}\beta\rho_1,\ \frac{\partial g_2}{\partial \mu_{10}} = 1.$$

$$\frac{\partial g_3}{\partial \mu_1} = \frac{\partial g_3}{\partial \mu_2} = \frac{\partial g_3}{\partial \mu_3} = \frac{\partial g_3}{\partial \mu_4} = \frac{\partial g_3}{\partial \mu_5} = \frac{\partial g_3}{\partial \mu_6} = \frac{\partial g_3}{\partial \mu_9} = \frac{\partial g_3}{\partial \mu_{10}} = 0,$$

$$\frac{\partial g_3}{\partial \mu_7} = \frac{\partial g_3}{\partial \mu_8} = -\frac{1}{2}\beta\rho_2,\ \frac{\partial g_3}{\partial \mu_{11}} = 1.$$

Since the partial derivative matrix is very sparse, we don't need to calculate all the terms of the matrix Σ. The ones that are needed are calculated below.

$\sigma_{5,6} = E(X_1^2X_2^2) - 1 = 2\rho_1^2 + 1 - 1 = 2\rho_1^2$

$\sigma_{5,8} = E(X_1^2Y^2) - 1 = 2\beta^2 + 1 - 1 = 2\beta^2$

$\sigma_{5,10} = E(X_1^2X_2Y) - \beta\rho_1 = 3\beta\rho_1 - \beta\rho_1 = 2\beta\rho_1$

$\sigma_{8,6} = E(X_2^2Y^2) - 1 = 2\beta^2\rho_1^2 + 1 - 1 = 2\beta^2\rho_1^2$

$\sigma_{8,8} = E(Y^4) - 1 = 2$

$\sigma_{8,10} = E(X_2Y^3) - \beta\rho_1 = 3\beta\rho_1 - \beta\rho_1 = 2\beta\rho_1$

$\sigma_{9,6} = E(X_1X_2^2Y) - \beta = 2\beta\rho_1^2 + \beta - \beta = 2\beta\rho_1^2$

$\sigma_{9,8} = E(X_1Y^3) - \beta = 3\beta - \beta = 2\beta$

$\sigma_{9,10} = E(X_1X_2Y^2) - \beta^2\rho_1 = 2\beta^2\rho_1 + \rho_1 - \beta^2\rho_1 = \rho_1(1+\beta^2)$

$\sigma_{6,7} = E(X_2^2X_3^2) - 1 = 2\rho_3^2 + 1 - 1 = 2\rho_3^2$

$\sigma_{6,11} = E(X_2^2X_3Y) - \beta\rho_2 = 2\beta\rho_1\rho_3 + \beta\rho_2 - \beta\rho_2 = 2\beta\rho_1\rho_3$

$\sigma_{8,7} = E(X_3^2Y^2) - 1 = 2\beta^2\rho_2^2 + 1 - 1 = 2\beta^2\rho_2^2$

$\sigma_{8,11} = E(X_3Y^3) - \beta\rho_2 = 3\beta\rho_2 - \beta\rho_2 = 2\beta\rho_2$

$\sigma_{10,7} = E(X_2X_3^2Y) - \beta\rho_2 = 2\beta\rho_2\rho_3 + \beta\rho_2 - \beta\rho_2 = 2\beta\rho_2\rho_3$

$\sigma_{10,11} = E(X_2X_3Y^2) - \beta^2\rho_1\rho_2 = \rho_3 + 2\beta^2\rho_1\rho_2 - \beta^2\rho_1\rho_2 = \rho_3 + \beta^2\rho_1\rho_2$

Combining, we get,

$$Cov(\sqrt{n}(r_1 - \beta), \sqrt{n}(r_2 - \beta\rho_1)) = \frac{\rho_1}{2}(1 - \beta^2)(2 - \beta^2 - \beta^2\rho_1^2),$$

$$Cov(\sqrt{n}(r_2 - \beta\rho_1), \sqrt{n}(r_3 - \beta\rho_2)) = 2\rho 3 + \beta^2(\rho_1^2 + \rho_2^2)(\beta^2\rho_1\rho_2 - 2\rho 3) + \beta^2\rho_1\rho_2(\rho_3^2 - 1).$$

Also,

$$Var(\sqrt{n}(r_1 - \beta)) = (1-\beta^2)^2,\ Var(\sqrt{n}(r_2 - \beta\rho_1)) = (1-\beta^2\rho_1^2)^2,\ Var(\sqrt{n}(r_3 - \beta\rho_2)) = (1-\beta^2\rho_2^2)^2.$$

Hence,

$$Cor(r_1, r_2) = \rho_{12} = \frac{\rho_1(2 - \beta^2 - \beta^2\rho_1^2)}{2n(1 - \beta^2\rho_1^2)}$$

and

$$Cor(r_2, r_3) = \rho_{23} = \frac{2\rho 3 + \beta^2(\rho_1^2 + \rho_2^2)(\beta^2\rho_1\rho_2 - 2\rho 3) + \beta^2\rho_1\rho_2(\rho_3^2 - 1)}{2n(1 - \beta^2\rho_1^2)(1 - \beta^2\rho_2^2)}.$$

□

Corollary 3.1. *Let z_1, z_2 and z_3 be the Fisher transformed unscaled z-statistics corresponding to r_1, r_2 and r_3. Then,*

$$\sqrt{n-3}\begin{pmatrix} z_1 - tanh^{-1}(\beta) \\ z_2 - tanh^{-1}(\beta\rho_1) \\ z_3 - tanh^{-1}(\beta\rho_2) \end{pmatrix} \xrightarrow{d} N(\mathbf{0}, \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho 23 \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix}),$$

where

$$\rho_{12} = \frac{\rho_1(2 - \beta^2 - \beta^2\rho_1^2)}{2(1 - \beta^2\rho_1^2)}$$

and

$$\rho_{23} = \frac{2\rho 3 + \beta^2(\rho_1^2 + \rho_2^2)(\beta^2\rho_1\rho_2 - 2\rho 3) + \beta^2\rho_1\rho_2(\rho_3^2 - 1)}{2(1 - \beta^2\rho_1^2)(1 - \beta^2\rho_2^2)},$$

ρ_{13} having the same form as ρ_{12}.

Corollary 3.2. *The covariance of the z-statistics converge to the covariance matrix for the case $\beta = 0$ as $|\rho_1| \to 1$ and $|\rho_2| \to 1$, or $|\rho_1| \to 0$ and $|\rho_2| \to 0$. This is also true for the conditional mean $E(z_2, z_3|z_1)$.*

The proof of Corollary 3.1 and Corollary 3.2 follows directly from Lemma 3. Clearly, similar results apply to more than three variables. Corollary 3.2 immediately implies that the conditional distribution of $(z_2, z_3|z_1)$ is approximately free of β when the correlations ρ_1 and ρ_2 are very large or very small. So, if the data has a block structure where there is very high correlation among SNPs within a block and there is very small correlation across blocks, then assumption (A3) may hold approximately, in a manner that supports the use of Z-REG-FDR.

To understand the difference between null and alternative of the conditional covariance matrices and mean vectors, we calculated the large sample means and covariance matrices under the two cases using Corollary 3.1. The dependence structure among the SNPs is (i) assumed to be an AR(1) structure with serial correlation 0.9, (ii) obtained from a real SNP matrix (Lonsdale et al. 2013).

For case (i), Figure 3.1 shows the plot of the elements of the conditional covariance matrix under the null and that under the alternative for different effect sizes. The maximum difference in the conditional mean is also reported for each case. Figure 3.2 shows the same plot for case (ii). The fact that the differences are small, especially for the real SNP matrix, is an encouraging sign in favor of *Z-REG-FDR*.

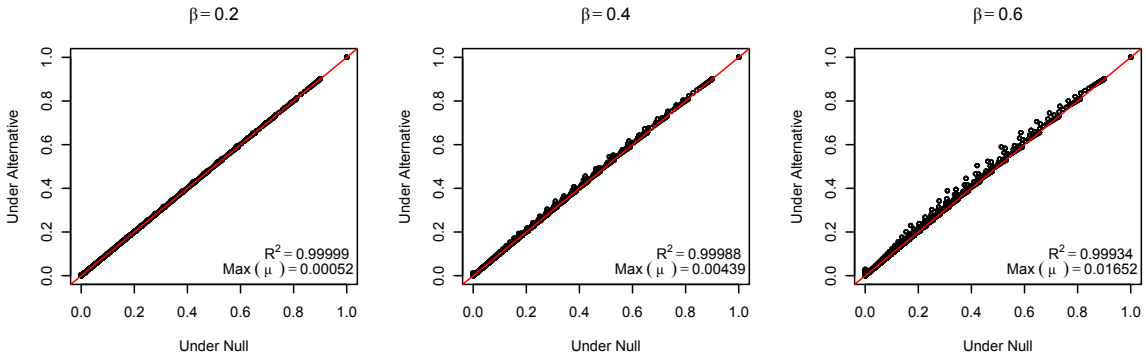

Figure 3.1: Comparing the elements of conditional covariance matrix of Z under the null and those under the alternative. The R^2 as well as the maximum difference in the conditional means are reported. The correlation structure of the SNPs is assumed to be AR(1). β is the effect size.

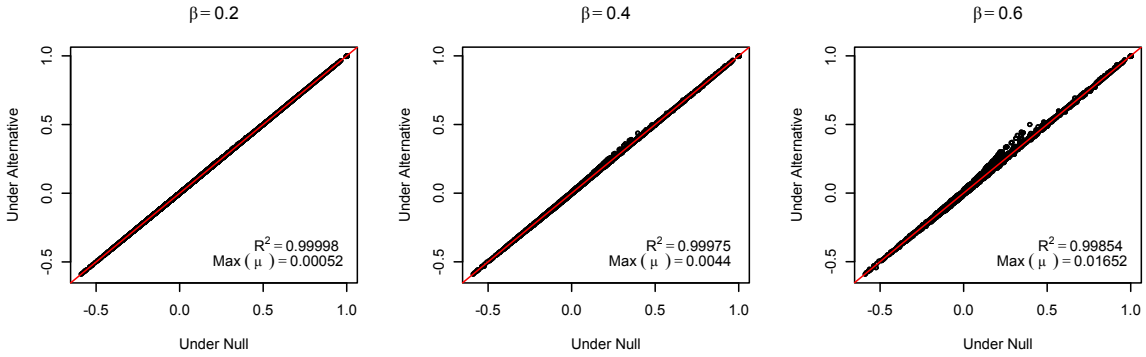

Figure 3.2: Comparing the elements of conditional covariance matrix of Z under the null and those under the alternative. The R^2 as well as the maximum difference in the conditional means are reported. The correlation structure of the SNPs is obtained from a real data. β is the effect size.

3.6 Results of *Z-REG-FDR* as an approximate maximum likelihood estimation

To study the accuracy of the estimation when the method is only an approximate maximum likelihood estimation, we have simulated data which uses the covariate adjusted genotype matrix of a real dataset from the GTEx project (Lonsdale et al. 2013). The genotype matrix corresponding to the tissue 'heart', which had 83 samples, is selected for analysis. For computational purposes, 10,000 genes were chosen randomly

from 28991 genes. Use of genotype matrices from real data ensures that we are not enforcing assumption (A3) while simulating, and our choice of $f_{0|k}$ for the simulation is realistic. We simulate the Y_i's (1000 simulations) using the following scheme.

1. For each gene, decide whether it has an eQTL using a Bernoulli(π_0) distribution.
2. Pick a causal SNP using a discrete uniform distribution over the m_i SNPs. Let it be the kth SNP.
3. If the gene has an eQTL, simulate Y_i from $N(\beta X_k^{(i)}, 1)$ with $\sqrt{n-3}\ tanh^{-1}(\beta)$ simulated from a $N(0, \sigma^2)$ distribution. If the gene doesn't have an eQTL, simulate Y_i from $N(0, 1)$.

True π_0	True σ	Mean $\hat{\pi_0}$	Mean $\hat{\sigma}$	SE($\hat{\pi_0}$)	SE($\hat{\sigma}$)	Realized FDR(5%)	Realized FDR(10%)
0.10	1	0.1665	1.0771	0.0829	0.0479	0.0415	0.0659
0.10	2	0.0871	2.0443	0.0234	0.0234	0.0616	0.0964
0.10	5	0.0994	5.1088	0.0073	0.0221	0.0509	0.0974
0.20	1	0.2599	1.0802	0.0846	0.0534	0.0512	0.0903
0.20	2	0.1864	2.0437	0.0237	0.0263	0.0568	0.1106
0.20	5	0.1986	5.1075	0.0080	0.0275	0.0518	0.1017

Table 3.2: Showing summary of the simulation studies using the SNP matrix from real data

Table Table 3.2 shows the results for this data which indicates that the estimates are still accurate and control of FDR is satisfactory unless σ is very small. We often observe large effect sizes for eQTL data, so that σ is not expected to be very small. Therefore, the *Z-REG-FDR* has valid applications for eQTL data. When the SNP correlation structure is assumed to be AR(1), the results are slightly anti-conservative even for large σ (Table 3.3). This indicates that the accuracy of the *Z-REG-FDR* method depends on the actual correlation structure among the SNPs even though we avoid modeling such correlation structure. However, the results from Table 3.2 support the validity of the method for real data.

True π_0	True σ	True ρ	Mean $\hat{\pi_0}$	Mean $\hat{\sigma}$	SE($\hat{\pi_0}$)	SE($\hat{\sigma}$)	Realized FDR(5%)	Realized FDR(10%)
0.20	1	0.10	0.1955	1.0358	0.1559	0.0644	0.0982	0.1333
0.20	2	0.10	0.1521	1.9561	0.0465	0.0356	0.0757	0.1392
0.20	5	0.10	0.1918	4.9370	0.0092	0.0306	0.0534	0.1053
0.20	1	0.50	0.2169	1.0392	0.1431	0.0665	0.0838	0.1232
0.20	2	0.50	0.1608	1.9613	0.0412	0.0337	0.0700	0.1318
0.20	5	0.50	0.1924	4.9383	0.0089	0.0298	0.0532	0.1049
0.20	1	0.80	0.2375	1.0367	0.1221	0.0657	0.0706	0.1078
0.20	2	0.80	0.1808	1.9742	0.0325	0.0306	0.0590	0.1156
0.20	5	0.80	0.1948	4.9448	0.0084	0.0287	0.0523	0.1034

Table 3.3: Showing summary of the simulation studies where the SNP matrix has an AR(1) structure with correlation ρ

Figure 3.3 shows the plot of *REG-FDR* estimates against the *Z-REG-FDR* estimates for simulated datasets (500 simulations) using the above scheme. It is clear from the plot that the two methods agree with each other to a large extent (having correlations 0.9064 and 0.9522 for π_0 and σ respectively) and largely falling near the unit line, which implies that the approximate maximum likelihood method in *Z-REG-FDR* is quite effective in controlling the FDR with a much improved computation speed. A comparison of the estimated lfdr and estimated FDR of the two methods is also shown (Figure 3.4).

Based on these simulations, the *Z-REG-FDR* estimate of π_0 has a relative efficiency of 0.81 when compared with the corresponding estimate of *REG-FDR*. The relative efficiency of the σ estimate of *Z-REG-FDR* is 0.96. Figure 3.5 shows the histogram of correlations between the estimated FDR based on the true values of the parameters and that based on *REG-FDR* or *Z-REG-FDR*. Clearly, the correlations are very high, especially for *REG-FDR*. The higher correlation in case of *REG-FDR* is believed to be partly due to the higher efficiency of the parameter estimates and partly due to the fact that it uses the 'correct' expression for the lfdr. However, we have seen from simulations that in a few cases, *REG-FDR* estimates are much worse as compared to *Z-REG-FDR*. This may be due to convergence issues as the likelihood surfaces can

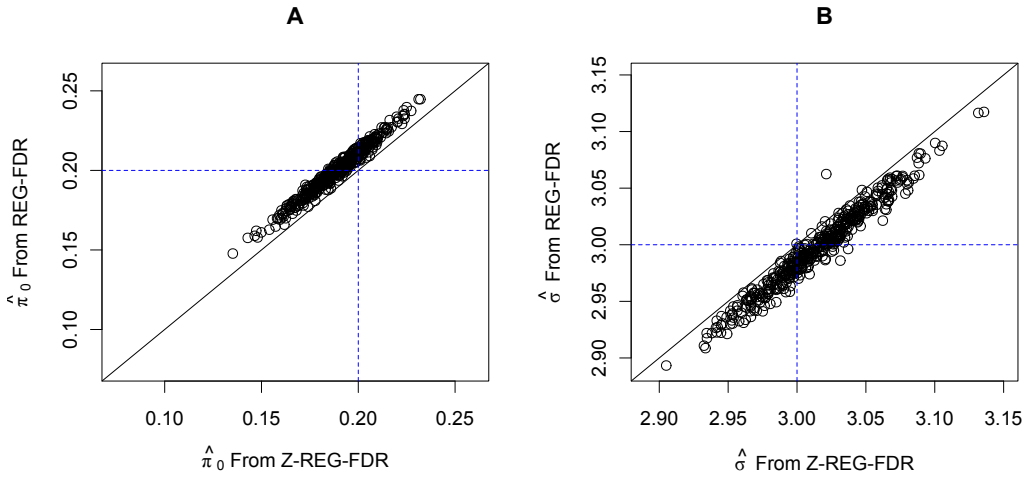

Figure 3.3: Showing the comparison of the estimates using REG-FDR and Z-REG-FDR. Except a small number of cases, the two estimates agree with each other. The blue lines show the true values of the parameters.

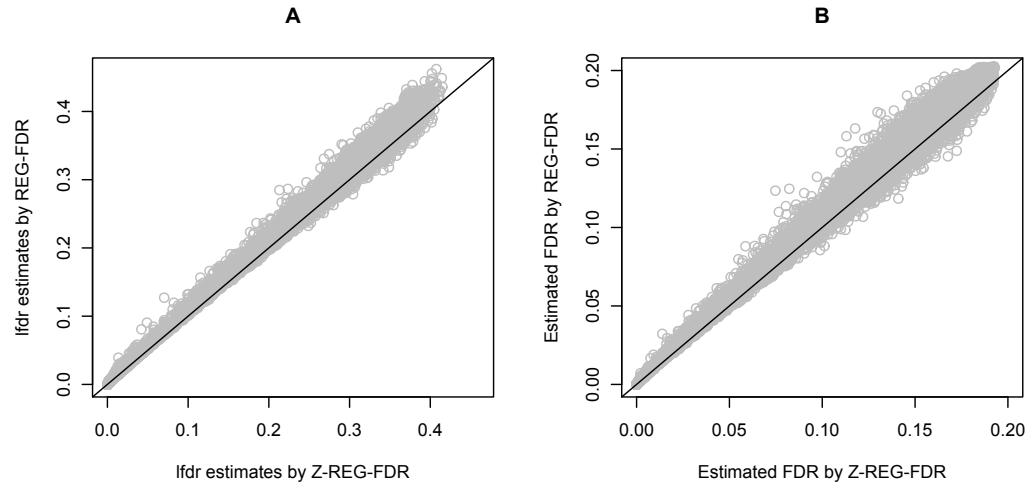

Figure 3.4: Showing the **A.** estimated lfdr and **B.** estimated FDR for *REG-FDR* and *Z-REG-FDR*.

sometimes be very flat.

It is a standard result that the expected log-likelihood is maximized at the true value of the parameter under standard regularity conditions (Cox and Hinkley 1979). Since *REG-FDR* is the true maximum likelihood method for the proposed model, it is expected to satisfy this property. However, *Z-REG-FDR* is only an approximate

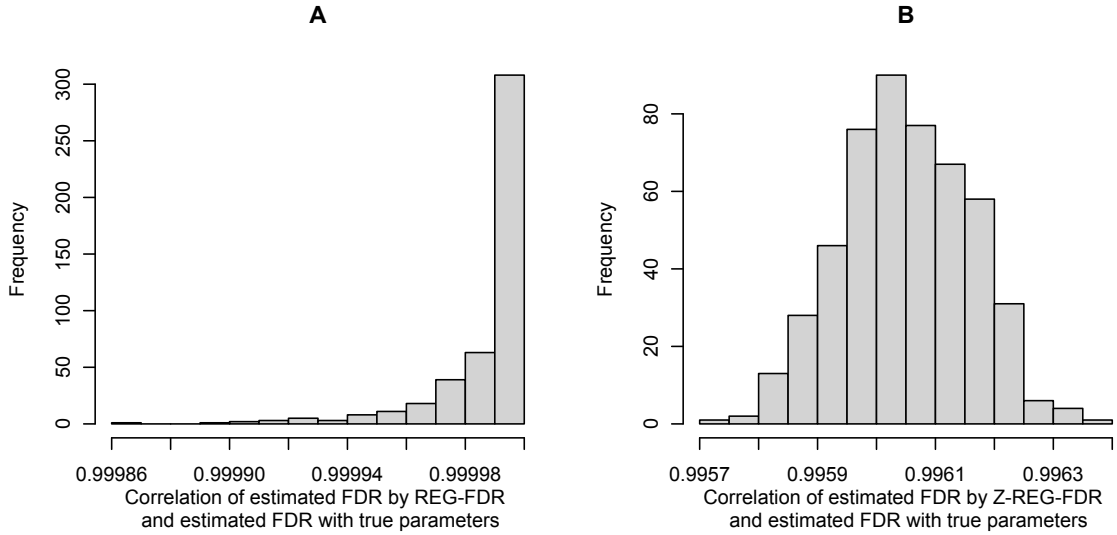

Figure 3.5: Showing the histograms of correlations between the estimated FDR based on the true values of the parameters and that based on **A.** *REG-FDR* **B.** *Z-REG-FDR*.

maximum likelihood method and may not have the property. We explored several combinations of the true parameters and observed that the pseudo-log-likelihood of *Z-REG-FDR* peaks near the true value. It is a difficult task to analytically compute the expected pseudo-log-likelihood, and so Monte-Carlo integration was used. Figure 3.6 shows the expected pseudo-log-likelihood surface of *Z-REG-FDR* for $\pi_0 = 0.2$ and $\sigma = 3$. A contour plot is also confirms the fact the surface peaks near the true values of the parameters.

3.7 Comparison with other methods

It is possible to use other methodologies to control FDR in grouped hypothesis testing problem for eQTL data. A conservative approach might be to obtain the Bonferroni adjusted p-values for each gene where the p-value for each gene-SNP pair is computed based on usual t-test or z-test, and using an FDR controlling approach (eg Benjamini and Hochberg 1995, Storey 2002, Strimmer 2008) with those p-values. Ardlie et al. (2015) used a permutation based approach for GTEx data. The method uses the

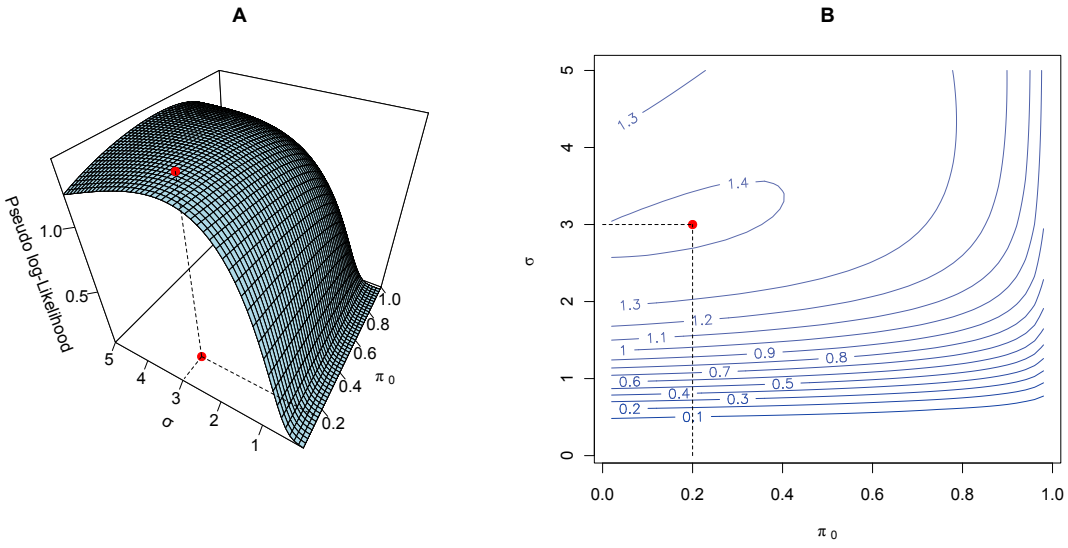

Figure 3.6: Showing **A.** the surface plot and **B.** the contour plot of expected pseudo-log-likelihood surface for the *Z-REG-FDR* method. True π_0 and σ are 0.2 and 3 respectively.

smallest gene-SNP p-value for a gene as the test statistic and computes its distribution by permuting the expression values. Such a distribution can be used to obtain p-values for each gene that can subsequently be used to control the FDR.

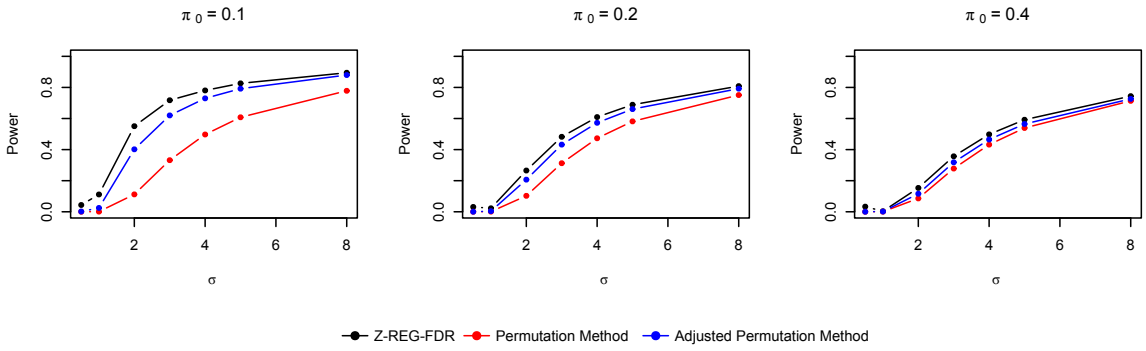

Figure 3.7: Showing the power curves of different methods for varying combinations of the true parameter values.

The Bonferroni method is usually very conservative and hence less powerful. Even the permutation method can suffer from lack of power to detect the genes having an eQTL since it uses an extreme value statistic (not based on likelihood). Our model, on

the other hand, utilizes more data through its likelihood. We have carried out some simulation studies to compare the performance of the methods in terms of their power. The simulations were done using the simulation scheme described in Section 3.6. As expected, the Bonferroni method turned out to have very low power. The permutation method, along with Storey's q-value method (Storey 2002), appeared to be conservative and less powerful as compared to *Z-REG-FDR* (Figure 3.7). We also applied an adjusted version of Storey's method that controls the FDR at some target level $\alpha > 0.05$ such that the realized FDR is 0.05. Note that this method is applied just for the comparison purpose and is inapplicable in real data scenarios as it requires the knowledge of the true states of the hypotheses. The method still appears to be less powerful than *Z-REG-FDR*.

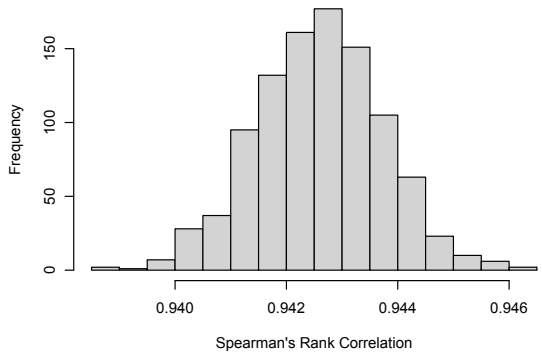

Figure 3.8: Showing the histogram of correlations between estimated FDR using the permutation method and that using *Z-REG-FDR*.

The estimated FDR (Strimmer 2008) using the permutation method and our *Z-REG-FDR* method tend to be highly correlated (Figure 3.8). The correlation of the estimated FDR using the true parameter values and that using the permutation method is also high, but not as high as *REG-FDR* or *Z-REG-FDR*. The correlations are much lower for the Bonferroni method (Figure 3.9).

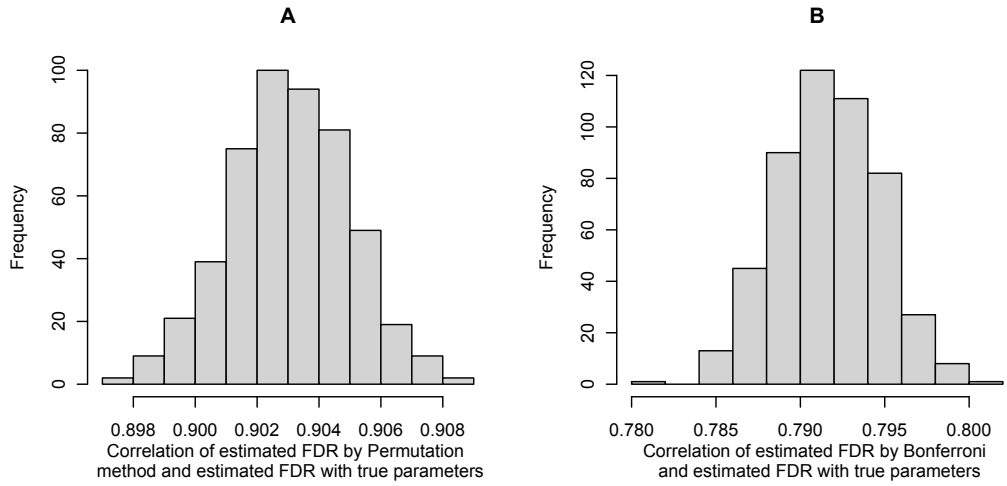

Figure 3.9: Showing the histogram of correlations between estimated FDR using the true parameter values and that using permutation method or Bonferroni method.

3.8 Advantage of *Z-REG-FDR* over other methods

The major advantage of *Z-REG-FDR* seems to be its computational efficiency. While other methods can take days to complete the analysis of a real eQTL dataset, *Z-REG-FDR* can do the same in a few minutes. For instance, it takes about two minutes to fit the model and find significant genes by *Z-REG-FDR* for a data with 4.5 million SNPs grouped into 10000 genes. *REG-FDR* takes about a day and the permutation method (for $10,000$ permutations) takes about 6 hours to analyze the same data. Since there are thousands of simultaneous tests, even $10,000$ permutations may not be enough to detect significance properly. While the Bonferroni method is very fast, it has little power to detect the genes having eQTL.

Z-REG-FDR has other advantages too. One important feature of the method is that it does not require the access to the full data. In fact, the symmetry of the distributions involved in the *Z-REG-FDR* pseudo-likelihood ensures that only the gene-SNP level p-values (or equivalently the absolute z-values) are sufficient to fit the model. Not only do we not model the correlation structure of the SNPs, we do not even need to have

access to that data. This might be very useful since in many genetic applications, data are found in the form of summary measures.

Also, *Z-REG-FDR* apparently does not suffer from the convergence issues that sometimes affect the estimation for *REG-FDR*. Therefore, it can be considered as a slightly less efficient, but reliable method. *Z-REG-FDR* can be slightly anti-conservative depending on the true values of the parameters. Various simulations show that if σ is large enough, which is often the case for eQTL data, the control of FDR is satisfactory. The fact that assumption (A3) is not satisfied does not affect the FDR control too much. Therefore that assumption can be thought of as a means to reduce computation burden, rather than a necessary assumption for the model. In the next section, we will demonstrate empirical evidence that the method remains valid even for more than one causal SNPs under certain conditions.

3.9 Effect of more than one causal SNPs

One concern about our model is that it may have limited applicability for large cis-windows since it uses the assumption of only one causal SNP. We have explored through simulation the effect of more than one causal SNPs on the control of the FDR. We observed that under certain conditions, even in the presence of two causal SNPs, *Z-REG-FDR* is only very slightly anti-conservative.

True π_0	True σ	Mean $\hat{\pi_0}$	Mean $\hat{\sigma}$	SE($\hat{\pi_0}$)	SE($\hat{\sigma}$)	Realized FDR(5%)	Realized FDR(10%)
0.10	1	0.2178	1.1354	0.0800	0.0508	0.0320	0.0533
0.10	2	0.0942	2.1099	0.0237	0.0264	0.0566	0.0945
0.10	5	0.0884	5.1313	0.0070	0.0218	0.0574	0.0999
0.20	1	0.3039	1.1353	0.0764	0.0550	0.0439	0.0780
0.20	2	0.1926	2.1071	0.0241	0.0294	0.0545	0.1066
0.20	5	0.1885	5.1269	0.0077	0.0278	0.0549	0.1075

Table 3.4: Showing summary of the simulation studies for two causal SNPs

Table 3.4 shows the results for simulated dataset. Under the alternative hypothe-

sis, the expressions are simulated using one primary causal SNP for which the Fisher transformed effect size follows a normal distribution with standard deviation σ, and there might exist (with probability 1/2) a secondary causal SNP which has an effect size that is smaller in magnitude and has the same sign as the primary effect size. Note that it is not possible to have the secondary effect size to be unconstrained and at the same time maintain the desired variance of Y. It can be shown that the simulation using the above mentioned conditions is always feasible (For more details see Appendix C). Table 3.4 demonstrates that if the secondary effect size is not very large and has the same direction, then *Z-REG-FDR* achieves reasonable control of the FDR.

3.10 Analysis of real data

In this section, we will present the results of application of *Z-REG-FDR* on some real datasets. The data were taken from the GTEx pilot study (Lonsdale et al. 2013). *Z-REG-FDR*, along with the Bonferroni method and the permutation method, was applied on the eQTL data for nine tissues separately. For more details about the datasets and the data pre-processing, see Appendix C.

Tissue	$\hat{\pi}_0$	$\hat{\sigma}$	Number of significant genes by *Z-REG-FDR*	Number of significant genes by Bonferroni	Number of significant genes by Permutation
Adipose	0.4536	2.6525	2857	1338	3578
Artery	0.4032	2.8706	3806	1851	3944
Heart	0.4591	2.4807	2443	1094	3591
Lung	0.4249	2.9145	3688	1787	3769
Muscle	0.4481	2.8733	3340	1629	3188
Nerve	0.3562	2.6904	4032	1791	4739
Skin	0.3999	2.6156	3320	1451	3820
Thyroid	0.3511	2.9399	4794	2269	4875
Blood	0.4817	3.2248	3718	2078	3535

Table 3.5: Showing *Z-REG-FDR* parameter estimates and summary of the findings for the GTEx datasets

The results of the analysis are summarized in Table 3.5. Clearly the methods agree with each other to some extent in terms of number of discoveries. The *Z-REG-FDR* method has much higher number of discoveries compared to the Bonferroni method, but in most cases has fewer discoveries compared to the permutation method that controls the FDR using the Benjamini-Hochberg method.

3.11 Inverse Average Method

There are a number of methods that can be used to estimate the lfdr at the gene-SNP level. Therefore it will be useful if those gene-SNP level lfdr's within a gene can be combined in some way to obtain the gene level lfdr. Given the set up in Section 3.3, Equation 3.4 and Equation 3.5 indicate that the inverse average or harmonic mean of the λ_{ij}'s can be equal to the gene level lfdr λ_i except for the difference in the priors (π_0 and $\tilde{\pi_0}$) and the difference between $f_0(.)$ and $f_{0ij}(.)$. In fact, it can be shown that the inverse average of λ_{ij}'s is an upper bound for the gene level lfdr λ_i. To show this, consider

$$\begin{aligned}
\lambda_{ij} &= \frac{f(H_{0ij}, Y_i, X_j^{(i)})}{f(H_{0ij}, Y_i, X_j^{(i)}) + f(H_{0ij}^c, Y_i, X_j^{(i)})} \\
&\geq \frac{f(H_{0i}, Y_i, X_j^{(i)})}{f(H_{0i}, Y_i, X_j^{(i)}) + f(H_{0ij}^c, Y_i, X_j^{(i)})} \\
&= \frac{\pi_0 f_0(Y_i)}{\pi_0 f_0(Y_i) + (1-\tilde{\pi_0}) f_1(Y_i|X_j^{(i)})} \\
&\geq \frac{\pi_0 f_0(Y_i)}{\pi_0 f_0(Y_i) + (1-\pi_0) f_1(Y_i|X_j^{(i)})}
\end{aligned}$$

The first inequality follows since $H_{0i} \subseteq H_{0ij}$, and the second inequality follows from the fact $\tilde{\pi_0} \geq \pi_0$. Therefore, using Equation 3.4, we obtain

Inequality 1. $\frac{1}{\frac{1}{m_i}\sum_{j=1}^{m_i}\frac{1}{\lambda_{ij}}} \geq \lambda_i$.

However, Figure 3.10 shows that such upper bound might not be sharp enough for feeding it into the adaptive thresholding procedure for controlling FDR. The difference is believed to be due the difference in the priors which, under the set up in Section 3.3, are related by the equation

$$\tilde{\pi_0} = \pi_0 + \frac{m_i - 1}{m_i}(1 - \pi_0) \tag{3.11}$$

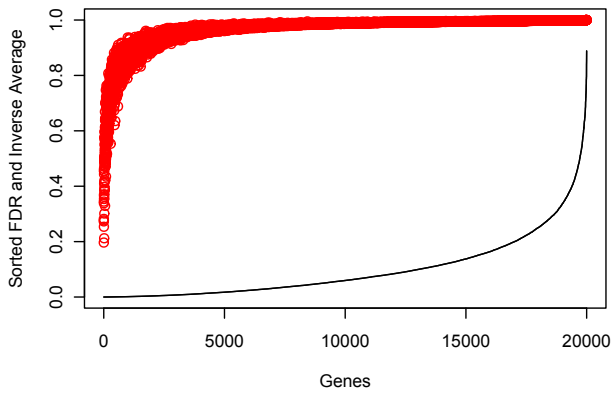

Figure 3.10: Showing the sharpness of the Inverse Average bound using a simulated data. The black line shows the sorted true gene lfdr's and the red dots are the inverse average of the corresponding gene-SNP level lfdr's. The simulation procedure used is similar to the scheme described in Section 3.5.

We propose an adjustment factor to adjust the inverse average which addresses the difference between $\tilde{\pi_0}$ and π_0. The proposal is to use the adjusted inverse average given by

$$\frac{1}{\frac{AF}{m_i}\sum_{j=1}^{m_i}\frac{1}{\lambda_{ij}} + (1 - AF)}$$

where AF is the adjustment factor defined as

$$AF = \frac{\tilde{\pi_0}(1 - \pi_0)}{\pi_0(1 - \tilde{\pi_0})} \tag{3.12}$$

When such adjustment is used, the performance of the inverse average improves greatly in terms of the sharpness of the bound (See Figure 3.11). The realized FDR while controlling FDR at 5% level is 0.046. The results from the simulations under various conditions show that once adjusted for the differences in the priors, the difference between $f_0(.)$ and $f_{0ij}(.)$ does not have much effect.

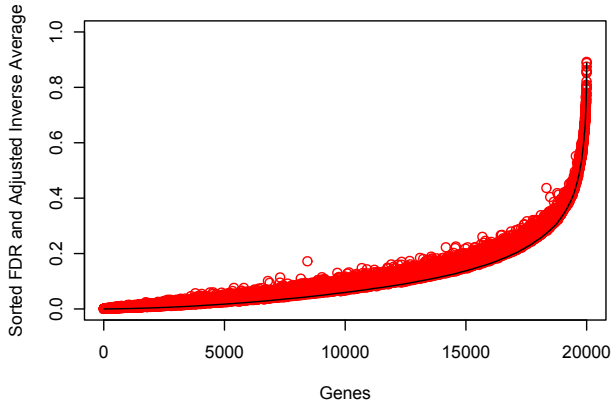

Figure 3.11: Showing the sharpness of the Inverse Average bound after adjustment. The black line shows the sorted true gene lfdr's and the red dots are the adjusted inverse average of the corresponding gene-SNP level lfdr's.

However, the use of the inverse average method for eQTL data has a serious difficulty. It is difficult to obtain the lfdr's λ_{ij} for the gene-SNP level hypotheses that tests whether the SNP is causal for the gene. The SNPs that are in linkage disequilibrium with the causal SNP will also show high association with the expression of the transcript and that will lead to under-estimation of the lfdr's in general. Therefore, even after the adjustment, the inverse average may not be an upper bound for the gene level lfdr.

Even though there is no guarantee in the real data scenario for the inverse average to be an upper bound for the gene level lfdr, it might still be useful in some cases. For instance, consider a hypothetical situation where the SNPs are divided into several blocks with the correlations within block being very high and those across blocks near

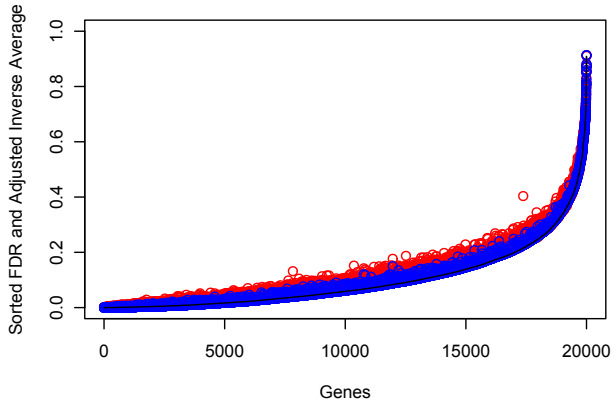

Figure 3.12: Showing the sharpness of the Inverse Average bound for a blocked data structure. The black line shows the sorted true gene lfdr's and the red dots are the adjusted inverse average for the hypothesis of causality. The blue dots are the adjusted inverse average for the gene-SNP level significance test.

zero. Simulations show that in such a situation, the adjusted inverse average serves as an approximate upper bound of the gene level lfdr (Figure 3.12).

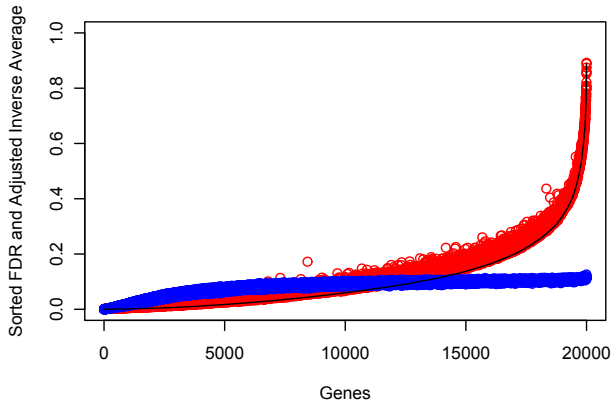

Figure 3.13: Showing the sharpness of the Inverse Average bound for a window type data structure. The black line shows the sorted true gene lfdr's and the red dots are the adjusted inverse average for the hypothesis of causality. The blue dots are the adjusted inverse average for the gene-SNP level significance test.

Figure 3.13 illustrates that if the data is such that all the SNPs within a window around the causal SNP have significantly small gene-SNP level lfdr, then the adjusted

inverse average is an upper bound when the lfdr are small, but is not an upper bound for larger values. Therefore, in such a scenario, the method may be useful if the target FDR level is small. However, one needs to look at the sorted inverse average values carefully and decide from its shape whether or not to use them as approximate gene level lfdrs. For more details of the simulation schemes, see Appendix C.

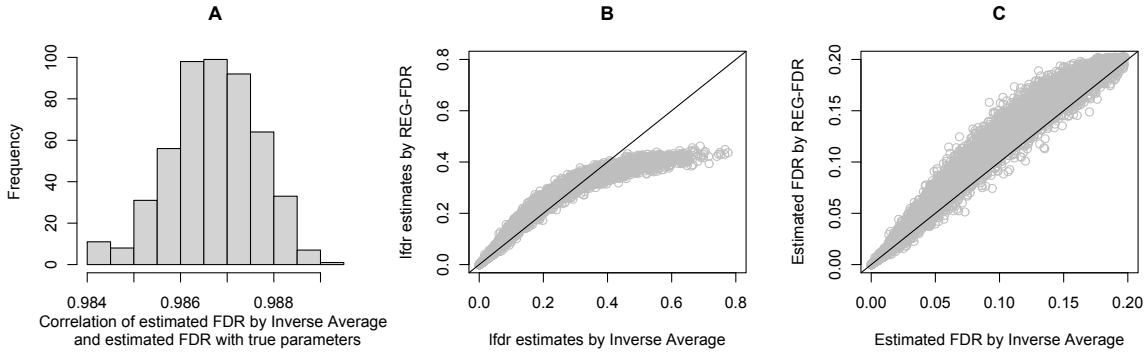

Figure 3.14: Showing the comparison of inverse average method with *REG-FDR*

Figure 3.14 shows the behavior of the inverse averages for a real SNP data while the expressions are simulated using our model. The comparison with *REG-FDR* shows that the inverse average method might be useful in this case. For the particular example, the realized FDR using inverse average method was 0.0652 implying that it is somewhat anti-conservative, but can still be used by setting the target FDR level slightly lower. However, one needs to know or estimate both the gene level and gene-SNP level priors in order to calculate the adjustment factor.

3.12 Discussion and future work

The *REG-FDR* method is the true maximum likelihood approach for the assumed model. However, the *Z-REG-FDR* method is computationally much efficient, and as shown above, produces estimates very similar to the *REG-FDR* procedure. Therefore, it enjoys the desirable properties of the maximum likelihood estimator with an improved

computation speed. It was also shown that the *Z-REG-FDR* method remains valid even when the assumptions are not fully true. Its performance under simulations and real data seem satisfactory in terms of controlling the FDR and at the same time achieving higher power as compared to some other methods.

Our model may be useful to analyse other similar data types. For instance, a similar version can be proposed for genome-wide association stuides (GWAS). However, GWAS data seldom have a true σ as large as what we observed in eQTL data. When the true σ is small, both *REG-FDR* and *Z-REG-FDR* seem to be inefficient. In particular, the estimation using both methods perform poorly when σ is smaller than 1. Therefore, even though this model may be applicable to many types of data, it is not advisable to use it unless the true σ is expected to be large.

The method might be slightly anti-conservative in some situations and future research is needed to understand the bias of the estimators so that the procedure can be adjusted to take care of such anti-conservativeness. Also, the current research focuses only on cis-eQTL. Analysis of trans-eQTL is an interesting statistical problem that is beyond the scope of the model in its current form. Further studies are required to modify the model in order to apply it to trans-eQTL problem.

The inverse average method is a simple tool to combine individual level lfdr to obtain the group level lfdr. However, in order to gurantee that the inverse average method will work, one needs to know the individual level lfdr for the hypothesis of causality, which is difficult to model statistically. Regardless, the inverse average has been shown to have the potential to work under different circumstances. However, it is inferior to the *Z-REG-FDR* in the sense that it is expected to be more anti-conservative.

Future research is required to verify if there are situations other than the eQTL data where a simple adjusted inverse average can be applied. One such hypothetical situation is described below.

Consider a similar set up, but not necessarily related to eQTL data. Suppose

we have an n-dimensional y-vector for each of N groups. Corresponding to a group i, we have a matrix $X^{(i)}$ with n columns and m_i rows. Let us call the m_i rows $X_1^{(i)}, X_2^{(i)}, ..., X_{m_i}^{(i)}$. We won't model the correlation structure of the $X_j^{(i)}$'s and assume that Y_i may have a causal relationship with at most one of the $X_j^{(i)}$'s. Suppose each $X_j^{(i)}$ consists of two parts given by

$$X_j^{(i)} = w_j^{(i)} + e_j^{(i)}.$$

While $w_j^{(i)}$'s might be considered either fixed or random, $e_j^{(i)}$'s are random errors and they have a correlation structure that generates the correlation structure of the $X_j^{(i)}$'s. $w_j^{(i)}$'s are assumed to be independent of the $e_j^{(i)}$'s and independent among themselves. The causal relationship of Y_i with the causal $X_j^{(i)}$ is given by

$$Y_i = \beta w_{causal}^{(i)} + \epsilon_i.$$

Under this situation, the sample correlations $r_j^{(i)} = Cor(Y_i, X_j^{(i)})$ will have a spike only at the causal location even though the $X_j^{(i)}$'s are correlated. Note however that the $r_j^{(i)}$'s are not independent, the alternativeness is not "transferred" to the other $r_j^{(i)}$'s.

If the lfdr's are known for each location, one can combine them using inverse average to obtain the lfdr for each group. Verifying the existence of such a problem in practice and finding methods to estimate the priors (to calculate the adjustment factor) requires more research on this topic.

CHAPTER 4: MULTI-TISSUE EXTENSION OF *Z-REG-FDR*

Recently, some studies have been using eQTL data for multiple tissues simultaneously (Li et al. 2013, Petretto et al. 2010, Flutre et al. 2013). Such use of multi-tissue data is expected to provide better results by borrowing strength across tissues. It can be shown that (Li, Nobel; personal communication, September 2014) use of more data facilitates the inference in the expected sense as follows.

Lemma 4. *Let z_1 be a set of data to test the null hypothesis H_0 and z_2 is an additional set of data. Then*

$$E(P(H_0|Z_1)|H_0) \leq E(P(H_0|Z_1, Z_2)|H_0).$$

The same result is true for H_1 since there is no specific role of the null hypothesis in Lemma 4. Li et al. (2013) also provided empirical evidence that the power to detect eQTL increases when more tissues are used. However, there have not been multi-tissue eQTL studies to test hypotheses at the gene level. In the following sections, we will propose an extension of the *Z-REG-FDR* model to a multi-tissue set up.

4.1 Data, notations and basic assumptions

We assume that the data are collected on the exact same SNPs for each tissue. However, the sample sizes may be different for each tissue. There may or may not be shared samples across tissues. Most methods are incapable of accommodating different sample sizes in different tissues, but that is not a problem here due to the use of variance stabilized z-statistics. Suppose $z^{(i)}_{k.} = (z^{(i)}_{k1}, z^{(i)}_{k2}, ..., z^{(i)}_{km_i})$ is the Fisher transformed and

scaled z-vector for the ith gene in the kth tissue, $i = 1, 2, ..., N$, $k = 1, 2, ..., K$. Also define $z_{.j}^{(i)} = (z_{1j}^{(i)}, z_{2j}^{(i)}, ..., z_{Kj}^{(i)})$ as the z-vector across the tissues for the jth SNP local to the ith gene, $j = 1, 2, ..., m_i$, $i = 1, 2, ..., N$.

With such a matrix of z-values for every gene and no missing values, we make the following assumptions.

1. For any gene with an eQTL, there is exactly one causal SNP. (A4)
2. Given that there is a causal SNP, it might be any of the m_i SNPs with probability $1/m_i$. (A5)
3. The causal SNP is the same in all the tissues, however it might be 'active' in some tissues and 'inactive' in some others. (The probability structure of such a causal SNP being 'active' will be discussed in the next section.) (A6)

The assumption of the same causal SNP for each tissue may not always be true, but it is often assumed that a particular SNP may act as the causal one for multiple tissues (Ardlie et al. 2015).

4.2 Further assumptions and the *MT-Z-REG-FDR* model

The configuration of the 'activity' status at the causal locus will be a vector of 0 and 1's. Suppose, for ith gene, the configuration vector is $C^{(i)} = (C_1^{(i)}, C_2^{(i)}, ..., C_K^{(i)})$, $C^{(i)} \in \{0, 1\}^K$. Note that $C^{(i)} = (0, 0, ..., 0)$ refers to the case that there is no eQTL in the ith gene.

Clearly there are 2^K possible configurations. Let us call them $c(0), c(1), ..., c(2^K - 1)$ for some order of the configurations and let the corresponding prior probabilities be π_r, $r = 0, 1, ..., 2^K - 1$, with $r = 0$ specifically corresponding to the case $c(0) = (0, 0, ..., 0)$.

We can model the π_r's in different ways.

1. We can assume nothing about the π_r's, ie $\pi_r \in (0, 1)$ with $\sum_r \pi_r = 1$

2. We can assume that the gene is null with a certain probability π_0, and if it is

alternative, then at the causal SNP any tissue's activity status follows a Bernoulli distribution independent from other tissues. The parameter of the Bernoulli distribution might be allowed to vary across the tissues if there is any biological reason to believe that some tissue might be more likely to have an active causal SNP as compared to other tissues.

There might be possible other models. Without any prior knowledge about these probabilities, we proceed with the unstructured case.

Now we make further assumptions as we did in case of univariate *Z-REG-FDR* model. Suppose the tth SNP is the causal SNP for the ith gene.

1. We assume that the conditional distribution of $(z_{.1}^{(i)}, ..., z_{.t-1}^{(i)}, z_{.t+1}^{(i)}, ..., z_{.m_i}^{(i)})$ given $z_{.t}^{(i)}$ under the alternative is same as that under the null and does not depend on the configuration at the causal SNP. (A7)

2. There exists a correlation structure among the $z_{.j}^{(i)}$'s, $j = 1, 2, ..., m_i$, due to commonalities among tissues arising from the underlying sampling process, for example, shared samples among tissues. We assume that such correlation structure, reflected by the covariance matrix Δ, also does not depend on the configuration at that SNP. (A8)

3. We assume that $E(z_{.j}) = \mathbf{0}$ for non-causal SNPs and $E(z_{.t}^{(i)}) = c.\mu$, c being the configuration vector, where $\mu = (\mu_1, \mu_2, ..., \mu_K)$ is the vector of random effects and is assumed to be following a $N_K(\mathbf{0}, \Sigma)$ distribution. $x.y$ denotes the Hadamard (entrywise) product of x and y. The covariance matrix Σ refelects the biological commonalities among the tissues. (A9)

4.3 The likelihood

Given the above set up and unstructured $\pi = (\pi_0, \pi_1, ..., \pi_{2^K-1})$, we have $2^K + K^2 - 1$ free parameters to estimate. The diagonals of the matrix Δ are all 1 due to the variance stabilizing transformation, and the sum of the components of π is 1. The likelihood is of the following form.

$$L(\pi, \Delta, \Sigma|z) = \prod_{i=1}^{N} \{\pi_0 + \sum_{r>0} \frac{\pi_r}{m_i} \sum_j \frac{p_1^r(z_{.j}^{(i)})}{p_0(z_{.j}^{(i)})}\} f_0(z_{.1}^{(i)}, z_{.2}^{(i)}, ..., z_{.m_i}^{(i)}) \tag{4.1}$$

where $p_1^r(.)$ is the density of $N_K(c(r).\mu, \Delta)$ and $p_0(.)$ is the density of $N_K(0, \Delta)$, Δ being the assumed covariance matrix of $z_{.j}^{(i)}$'s given the effect sizes. The μ vector is distributed as $N(\mathbf{0}, \Sigma)$. It can be easily shown that at the causal SNP t, $z_{.t}^{(i)}$ marginally follows as multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\Delta + \Sigma.cc^T$. In particular, for the case where there is no eQTL in any of the tissues for the ith gene, $z_{.j}^{(i)}$ follows $N_K(\mathbf{0}, \Delta)$ for all j.

$f_0(z_{.1}^{(i)}, z_{.2}^{(i)}, ..., z_{.m_i}^{(i)})$ is the density of $(z_{.1}^{(i)}, z_{.2}^{(i)}, ..., z_{.m_i}^{(i)})$ under the null. Unlike the univariate case, $f_0(z_{.1}^{(i)}, z_{.2}^{(i)}, ..., z_{.m_i}^{(i)})$ is not independent of all the parameters we want to estimate as it involves Δ. Therefore, dividing the likelihood by that value and maximizing the ratio will not be equivalent to maximum likelihood estimation even if all the assumptions are true. On the other hand, it is not completely specified by the model since we avoid modeling the correlation structure among the SNPs. To overcome this difficulty, we assume that $f_0(z_{.1}^{(i)}, z_{.2}^{(i)}, ..., z_{.m_i}^{(i)})$ is the product of m_i independent $N_K(\mathbf{0}, \Delta)$ densities. Note that this is not a 'real' assumption, but a 'computational trick' required for the maximization of the likelihood. The rationale for the approach reflects the belief that the correlation structure among the SNPs do not contain much information about the parameters we estimate, and even with this assumption, most of the information about (π, Δ, Σ) is preserved.

4.4 Application on simulated datasets

We have applied the method on a simulated dataset with two tissues, $20,000$ genes, and the following choice of the true parameters:

$$\pi = c(0.2, 0.25, 0.4, 0.15),$$

$$\Delta = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} 4.25 & 3.5 \\ 3.5 & 5 \end{pmatrix}.$$

The average of the estimates for 500 simulations are

$$\pi = c(0.1236, 0.3447, 0.4972, 0.0344),$$

$$\Delta = \begin{pmatrix} 1 & 0.0089 \\ 0.0089 & 1 \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} 4.3374 & 2.2636 \\ 2.2636 & 5.1564 \end{pmatrix}.$$

Clearly, the estimates of Δ and the diagonals of Σ are quite accurate, whereas the off-diagonal of Σ and π are not accurately estimated. Our observation is that the smaller components of π are usually under-estimated and the larger components are over-estimated. Also, the biological correlations among tissues as reflected by the off-diagonals of Σ are usually under-estimated.

4.5 Application on real datasets

We also applied the method to a real dataset from GTEx (Lonsdale et al. 2013). We used the data for the two tissues adipose and thyroid, and applied the *MT-Z-REG-FDR* method. The estimates of the parameters are:

$$\pi = c(0.0123, 0.2855, 0.1503, 0.5519),$$

$$\Delta = \begin{pmatrix} 1 & 0.3211 \\ 0.3211 & 1 \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} 5.4274 & 3.8199 \\ 3.8199 & 6.9346 \end{pmatrix}.$$

Our estimates of the diagonal entries of the Σ matrix appear to be larger as compared to the estimates by the MT-eQTL model of Li et al. (2013). This is justified due to the fact that the Σ matrix applies to only the causal SNP in our model, while it applies to all the gene-SNP pairs in the model by Li et al. (2013). The estimate of the off-diagonal element of Σ is reported to be slightly smaller compared to their estimate. The estimate of $\pi_{\mathbf{0}}$ is much smaller as compared to their estimate which is expected since a null gene requires all the corresponding gene-SNP pairs to be null.

4.6 Discussion and future work

Our *MT-Z-REG-FDR* model shows potential to be useful in gene level multi-tissue eQTL studies. However, until now, we have applied the method to a very limited set up. It requires further research to explore its performance in different situations. The procedure was applied for only two tissues. It remains to be seen how it performs with larger number of tissues.

The estimates of the prior configuration probabilities π appear to be biased. It may

be possible to adjust for such bias, but that requires further research on this topic.

APPENDIX A: TECHNICAL DETAILS FOR CHAPTER 2

A.1 Details of the analysis of simulated data

This section explains the details of the analysis of simulated data in Section 3.1. We have used Manhattan distance throughout all the analyses due to the ease of tail area computation (Section 2.3). The *RankCover* procedure with Manhattan distance appears to give similar results to that with Euclidean distance.

The sample size is 50 and we used 1000 simulations under the null for *RankCover* and MIC. For dCor and HHG, 1000 permutations are used. The power curves are obtained based on 500 simulations. The independent variable x is simulated as $U(0,1)$. The dependent variable y is calculated using the equation

$$y = f(x) + \nu \times error, \tag{4.2}$$

where ν is the noise scale parameter and increases from 0.1 to 1 as in Figure 4. The error distribution was chosen to be normal. However, as in Simon and Tibshirani (2014), the variance of the error distribution was considered differently for different forms of relationship. Section A.2 shows how the results are similar with other distributions also. The details of the forms of the function $f(.)$ and the error distributions are as below.

- Linear: $f(x) = x$, error distribution is $N(0,1)$
- Quadratic: $f(x) = 4(x - 1/2)^2$, error distribution is $N(0,1)$
- Cubic: $f(x) = 128(x - 1/3)^3 - 48(x - 1/3)^2 - 12(x - 1/3)$, error distribution is $N(0,100)$
- Sine: $f(x) = sin(4\pi x)$, error distribution is $N(0,4)$

- $X^{1/4}$: $f(x) = x^{1/4}$, error distribution is $N(0, 1)$
- Circle: $f(x) = (2r-1)\sqrt{1-(2x-1)^2}$, error distribution is $N(0, 1/16)$, where r is a Bernoulli(1/2) variable
- Two curves: $f(x) = 2rx + (1-r)\sqrt{x}/2$, error distribution is $N(0, 1/4)$, where r is a Bernoulli(1/2) variable
- X-function: $f(x) = rx + (1-r)(1-x)$, error distribution is $N(0, 1/25)$, where r is a Bernoulli(1/2) variable
- Diamond: $f(x) = r_1 I(x < 0.5) + r_2 I(x \geq 0.5)$, error distribution is $N(0, 1/100)$, where r_1 is a $U(0.5-x, 0.5+x)$ variable and r_2 is a $U(x-0.5, 1.5-x)$ variable

A.2 Details of Simulation results for different marginal distributions of the variables

We have carried out the simulation analysis for different marginal distributions of x and different error distributions. Three distributions of different shapes are used for the marginal distribution of X: uniform, truncated normal (a normal distribution with mean 1/2 and variance 1/12 truncated between 0 and 1)and a U-shaped beta (beta$(1/2, 1/2)$). The choices for the error distributions are normal, U(0,1) and beta$(1/2, 1/2)$ with appropriate shift of origin and scale so that the mean and variance of the error distributions are 0 and 1 respectively.

The results of these nine cases show that *RankCover* has reasonable power in all these cases. Table 4.1 shows a summary of all the cases. The mean power over all the noise levels are shown for each case. Since the power curves rarely cross each other, the mean power (which is approximately proportional to area under the power curve) appears to be a good indicator of performance.

Table 4.1: Showing the mean power of the different methods for the nine cases. eg. Beta-Normal refers to the case where marginal of x is beta and error distribution is normal

	Linear	Quadratic	Cubic	Sine	$X^{1/4}$	Circle	2-Curves	X-function	Diamond
					Beta-Beta				
dCor	0.90	0.48	0.67	0.47	0.67	0.11	1.00	0.20	0.09
RankCover	0.97	0.94	0.91	0.63	0.85	1.00	1.00	0.95	0.84
Hybrid	0.95	0.90	0.87	0.56	0.79	1.00	1.00	0.93	0.76
MIC	0.88	0.50	0.55	0.43	0.69	0.71	0.96	0.50	0.14
HHG	0.94	0.72	0.74	0.47	0.77	0.97	1.00	0.89	0.76
					Beta-Normal				
dCor	0.90	0.52	0.69	0.51	0.69	0.10	1.00	0.19	0.09
RankCover	0.75	0.72	0.75	0.48	0.54	1.00	0.97	0.94	0.81
Hybrid	0.86	0.67	0.74	0.49	0.62	1.00	0.99	0.91	0.74
MIC	0.70	0.61	0.61	0.51	0.46	0.73	0.93	0.51	0.15
HHG	0.81	0.65	0.70	0.46	0.59	0.98	0.99	0.89	0.77
					Beta-Uniform				
dCor	0.89	0.49	0.67	0.46	0.66	0.11	1.00	0.20	0.09
RankCover	0.85	0.80	0.81	0.50	0.63	1.00	0.99	0.94	0.83
Hybrid	0.86	0.74	0.77	0.47	0.62	1.00	0.99	0.91	0.75
MIC	0.74	0.51	0.56	0.44	0.47	0.71	0.93	0.50	0.15
HHG	0.82	0.62	0.69	0.41	0.60	0.97	0.99	0.88	0.76
					Normal-Beta				
dCor	0.71	0.42	0.38	0.34	0.40	0.05	0.94	0.46	0.05
RankCover	0.86	0.76	0.68	0.81	0.58	0.87	0.90	0.89	0.63
Hybrid	0.81	0.68	0.59	0.72	0.51	0.82	0.90	0.84	0.49
MIC	0.69	0.32	0.44	0.47	0.42	0.33	0.73	0.33	0.08
HHG	0.77	0.64	0.55	0.50	0.49	0.57	0.90	0.92	0.64
					Normal-Normal				
dCor	0.73	0.44	0.40	0.35	0.42	0.05	0.94	0.47	0.04
RankCover	0.56	0.50	0.41	0.61	0.30	0.85	0.85	0.88	0.63
Hybrid	0.65	0.45	0.40	0.55	0.36	0.79	0.88	0.84	0.50
MIC	0.48	0.38	0.37	0.58	0.25	0.33	0.69	0.35	0.08
HHG	0.58	0.53	0.43	0.48	0.32	0.57	0.89	0.93	0.63
					Normal-Uniform				
dCor	0.70	0.41	0.37	0.34	0.40	0.06	0.93	0.47	0.05
RankCover	0.65	0.57	0.49	0.69	0.36	0.84	0.85	0.87	0.62
Hybrid	0.65	0.50	0.44	0.60	0.36	0.78	0.87	0.83	0.50
MIC	0.50	0.33	0.35	0.50	0.26	0.33	0.67	0.33	0.08
HHG	0.59	0.53	0.42	0.45	0.33	0.55	0.88	0.91	0.63
					Uniform-Beta				
dCor	0.82	0.47	0.32	0.44	0.51	0.06	0.98	0.32	0.07
RankCover	0.93	0.88	0.74	0.76	0.71	0.98	0.98	0.94	0.77

continued to next page...

	Linear	Quadratic	Cubic	Sine	$X^{1/4}$	Circle	2-Curves	X-function	Diamond
Hybrid	0.90	0.81	0.65	0.67	0.64	0.97	0.97	0.91	0.68
MIC	0.79	0.42	0.38	0.46	0.55	0.50	0.87	0.42	0.10
HHG	0.87	0.68	0.52	0.48	0.61	0.78	0.97	0.93	0.73
Uniform-Normal									
dCor	0.81	0.46	0.27	0.43	0.51	0.12	0.97	0.26	0.04
RankCover	0.66	0.62	0.51	0.59	0.39	0.98	0.94	0.93	0.78
Hybrid	0.76	0.57	0.45	0.54	0.46	0.96	0.96	0.90	0.68
MIC	0.59	0.51	0.42	0.58	0.33	0.50	0.82	0.44	0.09
HHG	0.71	0.60	0.44	0.46	0.43	0.80	0.96	0.94	0.76
Uniform-Uniform									
dCor	0.80	0.46	0.32	0.44	0.50	0.07	0.98	0.33	0.06
RankCover	0.75	0.70	0.57	0.64	0.47	0.97	0.95	0.92	0.78
Hybrid	0.76	0.63	0.50	0.58	0.46	0.95	0.96	0.89	0.68
MIC	0.61	0.43	0.36	0.48	0.35	0.49	0.82	0.42	0.10
HHG	0.71	0.59	0.43	0.44	0.43	0.76	0.96	0.92	0.74

A.3 Details of real data analyses

While analyzing real data, some ties may be present even if the variables under study are continuous. Whenever we found ties during the data analysis, we randomly broke the ties many (100) times, and considered the average *RankCover* as our test statistic.

A.3.1 Example 1: Eckerle4 data

100,000 simulations were used for *RankCover* and MIC. 100,000 permutations were used for dCor and HHG. The estimates of $\beta_1, \beta_2, \beta_3$ obtained from NIST website are used for plotting the fitted curve in Figure 5. Source of data: NIST StRD for non-linear regression.

A.3.2 Example 2: Aircraft data

100,000 simulations were used for *RankCover* and MIC. 100,000 permutations were used for dCor and HHG. Source of data: `sm` Package in R (Bowman and Azzalini 2013).

A.3.3 Example 3: ENSO data

100,000 simulations were used for *RankCover* and MIC. 100,000 permutations were used for dCor and HHG. The estimates of $\beta_1, \beta_2, ..., \beta_9$ obtained from NIST website are used for plotting the fitted curve in Figure 7. Source of data: NIST StRD for non-linear regression.

A.3.4 Example 4: Yeast data

100,000 simulations were used for *RankCover* and MIC. 100,000 permutations were used for dCor and HHG. The data was pre-processed before analysis as follows. The data contained several missing observations. Since the sample size is small (24), we removed all the genes that had more than 3 missing observations. All other missing observations were imputed using KNN imputation (Troyanskaya et al. 2001). Then quantile normalization was used to normalize the data. Unlike Reshef et al. (2011), we didn't remove any of the time points and didn't use any interpolation to find expression values for intermediate timepoints. Source of data: Comprehensive Identification of Cell Cycle regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization.

APPENDIX B: TABLES OF THRESHOLDS OF *RANKCOVER*

Table 4.2: Showing the p-th quantiles of the *RankCover* statistic

Sample Sizes	p=0.1	p=0.05	p=0.025	p=0.01	p=0.001	p=0.0001
20	1.31000	1.28500	1.26000	1.23250	1.16500	1.10500
21	1.27211	1.24717	1.22449	1.19501	1.12925	1.07936
22	1.23554	1.21281	1.19215	1.16529	1.10331	1.06405
23	1.39698	1.37240	1.35161	1.32325	1.26087	1.20227
24	1.36111	1.33854	1.31771	1.29167	1.23090	1.16493
25	1.32960	1.30720	1.28640	1.26240	1.20640	1.14720
26	1.30030	1.27959	1.25888	1.23373	1.17899	1.12574
27	1.27298	1.25240	1.23320	1.20850	1.15501	1.10151
28	1.24745	1.22832	1.20918	1.18622	1.13520	1.08291
29	1.22473	1.20452	1.18668	1.16290	1.11415	1.07134
30	1.20222	1.18333	1.16444	1.14222	1.09222	1.04889
31	1.18106	1.16233	1.14464	1.12279	1.07700	1.03018
32	1.30469	1.28613	1.26855	1.24609	1.19922	1.15625
33	1.28375	1.26538	1.24885	1.22865	1.18182	1.14509
34	1.26384	1.24567	1.22924	1.20934	1.16263	1.11938
35	1.24490	1.22776	1.21061	1.19102	1.14286	1.10286
36	1.22685	1.20988	1.19367	1.17361	1.13040	1.09259
37	1.21110	1.19430	1.17823	1.15997	1.11395	1.08400
38	1.19453	1.17798	1.16274	1.14474	1.10319	1.06856
39	1.17883	1.16239	1.14727	1.12821	1.08613	1.05523
40	1.16375	1.14813	1.13375	1.11563	1.07437	1.04375
41	1.26413	1.24866	1.23379	1.21594	1.17668	1.14456
42	1.24943	1.23413	1.21995	1.20181	1.16213	1.12132
43	1.23580	1.22012	1.20606	1.18875	1.15035	1.10871
44	1.22159	1.20713	1.19318	1.17717	1.13998	1.10795
45	1.20889	1.19407	1.18074	1.16395	1.12444	1.09284
46	1.19660	1.18195	1.16824	1.15217	1.11531	1.08932
47	1.18470	1.17021	1.15708	1.14079	1.10457	1.07062
48	1.17231	1.15842	1.14497	1.12934	1.09115	1.06510
49	1.16160	1.14744	1.13369	1.11828	1.08330	1.05373
50	1.15080	1.13640	1.12400	1.10760	1.07360	1.04520
51	1.13995	1.12611	1.11342	1.09765	1.06267	1.03114
52	1.22337	1.20969	1.19749	1.18158	1.14756	1.11501
53	1.21289	1.19972	1.18726	1.17230	1.13884	1.11463
54	1.20302	1.18964	1.17764	1.16324	1.12929	1.10391
55	1.19273	1.17983	1.16793	1.15372	1.12066	1.09421
56	1.18367	1.17060	1.15912	1.14445	1.11129	1.07175
57	1.17421	1.16159	1.15020	1.13573	1.10249	1.07572
58	1.16498	1.15250	1.14090	1.12634	1.09304	1.06421

continued to next page...

Table 4.2: Showing the p-th quantiles of the *RankCover* statistic

Sample Sizes	p=0.1	p=0.05	p=0.025	p=0.01	p=0.001	p=0.0001
59	1.15628	1.14392	1.13243	1.11807	1.08475	1.05918
60	1.14778	1.13528	1.12361	1.11000	1.07805	1.05333
61	1.13948	1.12739	1.11610	1.10266	1.06987	1.04139
62	1.13137	1.11889	1.10744	1.09417	1.06322	1.03668
63	1.12371	1.11111	1.10028	1.08743	1.05694	1.03452
64	1.19434	1.18237	1.17188	1.15845	1.12524	1.10181
65	1.18627	1.17467	1.16426	1.15101	1.12260	1.09870
66	1.17906	1.16736	1.15657	1.14371	1.11524	1.08655
67	1.17153	1.15995	1.14970	1.13678	1.11049	1.08599
68	1.16436	1.15268	1.14208	1.12954	1.10208	1.07656
69	1.15690	1.14556	1.13506	1.12329	1.09494	1.07498
70	1.15041	1.13898	1.12878	1.11612	1.08898	1.06510
71	1.14323	1.13212	1.12200	1.10930	1.08173	1.05237
72	1.13657	1.12558	1.11555	1.10359	1.07485	1.05112
73	1.12986	1.11878	1.10884	1.09758	1.06943	1.04447
74	1.12381	1.11304	1.10299	1.09112	1.06410	1.03853
75	1.11769	1.10702	1.09707	1.08533	1.05778	1.04000
76	1.11165	1.10059	1.09107	1.07877	1.05315	1.03116
77	1.10525	1.09462	1.08501	1.07320	1.04638	1.02749
78	1.16650	1.15631	1.14678	1.13560	1.11012	1.08695
79	1.16055	1.15014	1.14084	1.12931	1.10383	1.08348
80	1.15453	1.14437	1.13484	1.12344	1.09922	1.07344
81	1.14906	1.13900	1.12986	1.11888	1.09282	1.07194
82	1.14307	1.13311	1.12433	1.11288	1.08864	1.06856
83	1.13805	1.12818	1.11932	1.10814	1.08434	1.06474
84	1.13265	1.12259	1.11338	1.10247	1.07851	1.05782
85	1.12720	1.11696	1.10754	1.09689	1.07170	1.05190
86	1.12196	1.11195	1.10289	1.09248	1.06963	1.04070
87	1.11692	1.10715	1.09856	1.08733	1.06223	1.04082
88	1.11170	1.10176	1.09310	1.08226	1.05850	1.03719
89	1.10668	1.09670	1.08787	1.07650	1.05328	1.03055
90	1.10198	1.09235	1.08346	1.07284	1.04901	1.02914
91	1.09733	1.08767	1.07910	1.06883	1.04516	1.02210
92	1.15064	1.14130	1.13268	1.12228	1.09983	1.07999
93	1.14591	1.13666	1.12880	1.11840	1.09481	1.07619
94	1.14113	1.13207	1.12381	1.11374	1.09110	1.07209
95	1.13651	1.12720	1.11889	1.10903	1.08632	1.06825
96	1.13184	1.12250	1.11404	1.10406	1.08203	1.06272
97	1.12754	1.11829	1.11021	1.09980	1.07716	1.05707
98	1.12349	1.11454	1.10641	1.09652	1.07434	1.05269
99	1.11887	1.10978	1.10183	1.09193	1.07091	1.05387
100	1.11470	1.10580	1.09740	1.08760	1.06480	1.04480

continued to next page...

Table 4.2: Showing the p-th quantiles of the *RankCover* statistic

Sample Sizes	p=0.1	p=0.05	p=0.025	p=0.01	p=0.001	p=0.0001
101	1.11229	1.10332	1.09536	1.08528	1.06346	1.04473
102	1.10803	1.09910	1.09122	1.08117	1.05942	1.04081
103	1.10377	1.09488	1.08707	1.07706	1.05538	1.03689
104	1.09951	1.09066	1.08292	1.07295	1.05134	1.03298
105	1.09525	1.08644	1.07878	1.06884	1.04730	1.02906
106	1.09099	1.08222	1.07463	1.06473	1.04326	1.02514
107	1.08673	1.07800	1.07048	1.06062	1.03922	1.02122
108	1.13400	1.12543	1.11806	1.10897	1.08813	1.07073
109	1.13053	1.12200	1.11464	1.10560	1.08484	1.06766
110	1.12706	1.11857	1.11123	1.10223	1.08154	1.06459
111	1.12358	1.11514	1.10782	1.09886	1.07824	1.06152
112	1.12011	1.11171	1.10441	1.09550	1.07495	1.05845
113	1.11664	1.10828	1.10100	1.09213	1.07165	1.05537
114	1.11316	1.10485	1.09759	1.08876	1.06835	1.05230
115	1.10969	1.10143	1.09418	1.08539	1.06505	1.04923
116	1.10622	1.09800	1.09077	1.08203	1.06176	1.04616
117	1.10274	1.09457	1.08735	1.07866	1.05846	1.04309
118	1.09927	1.09114	1.08394	1.07529	1.05516	1.04002
119	1.09580	1.08771	1.08053	1.07192	1.05187	1.03695
120	1.09233	1.08428	1.07712	1.06856	1.04857	1.03388
121	1.08885	1.08085	1.07371	1.06519	1.04527	1.03081
122	1.08538	1.07742	1.07030	1.06182	1.04197	1.02773
123	1.08191	1.07399	1.06689	1.05845	1.03868	1.02466
124	1.07843	1.07056	1.06348	1.05509	1.03538	1.02159
125	1.12051	1.11296	1.10630	1.09811	1.07776	1.05946
126	1.11767	1.11013	1.10349	1.09531	1.07514	1.05713
127	1.11482	1.10731	1.10068	1.09251	1.07252	1.05480
128	1.11198	1.10448	1.09786	1.08971	1.06990	1.05247
129	1.10913	1.10166	1.09505	1.08690	1.06728	1.05014
130	1.10628	1.09883	1.09223	1.08410	1.06466	1.04782
131	1.10344	1.09601	1.08942	1.08130	1.06204	1.04549
132	1.10059	1.09318	1.08661	1.07850	1.05942	1.04316
133	1.09775	1.09036	1.08379	1.07570	1.05681	1.04083
134	1.09490	1.08753	1.08098	1.07290	1.05419	1.03851
135	1.09206	1.08471	1.07816	1.07009	1.05157	1.03618
136	1.08921	1.08188	1.07535	1.06729	1.04895	1.03385
137	1.08637	1.07906	1.07254	1.06449	1.04633	1.03152
138	1.08352	1.07623	1.06972	1.06169	1.04371	1.02919
139	1.08068	1.07341	1.06691	1.05889	1.04109	1.02687
140	1.07783	1.07058	1.06409	1.05608	1.03847	1.02454
141	1.07499	1.06776	1.06128	1.05328	1.03585	1.02221
142	1.07214	1.06493	1.05846	1.05048	1.03323	1.01988

continued to next page...

Table 4.2: Showing the p-th quantiles of the *RankCover* statistic

Sample Sizes	p=0.1	p=0.05	p=0.025	p=0.01	p=0.001	p=0.0001
143	1.06929	1.06211	1.05565	1.04768	1.03061	1.01756
144	1.10745	1.10050	1.09418	1.08656	1.06964	1.05314
145	1.10505	1.09812	1.09183	1.08422	1.06736	1.05108
146	1.10266	1.09574	1.08948	1.08188	1.06508	1.04901
147	1.10026	1.09336	1.08712	1.07954	1.06281	1.04694
148	1.09787	1.09098	1.08477	1.07720	1.06053	1.04487
149	1.09548	1.08860	1.08242	1.07486	1.05825	1.04281
150	1.09308	1.08622	1.08006	1.07252	1.05598	1.04074
151	1.09069	1.08384	1.07771	1.07018	1.05370	1.03867
152	1.08830	1.08146	1.07536	1.06784	1.05142	1.03660
153	1.08590	1.07908	1.07300	1.06550	1.04915	1.03454
154	1.08351	1.07670	1.07065	1.06316	1.04687	1.03247
155	1.08111	1.07432	1.06830	1.06081	1.04459	1.03040
156	1.07872	1.07193	1.06594	1.05847	1.04232	1.02833
157	1.07633	1.06955	1.06359	1.05613	1.04004	1.02627
158	1.07393	1.06717	1.06123	1.05379	1.03776	1.02420
159	1.07154	1.06479	1.05888	1.05145	1.03548	1.02213
160	1.06915	1.06241	1.05653	1.04911	1.03321	1.02006
161	1.06675	1.06003	1.05417	1.04677	1.03093	1.01800
162	1.06436	1.05765	1.05182	1.04443	1.02865	1.01593
163	1.09895	1.09248	1.08657	1.07949	1.06421	1.05123
164	1.09692	1.09046	1.08456	1.07749	1.06225	1.04937
165	1.09488	1.08844	1.08255	1.07549	1.06030	1.04752
166	1.09285	1.08643	1.08054	1.07349	1.05834	1.04566
167	1.09081	1.08441	1.07853	1.07149	1.05638	1.04381
168	1.08878	1.08239	1.07652	1.06949	1.05443	1.04196
169	1.08674	1.08037	1.07451	1.06749	1.05247	1.04010
170	1.08471	1.07836	1.07250	1.06549	1.05052	1.03825
171	1.08267	1.07634	1.07049	1.06348	1.04856	1.03640
172	1.08064	1.07432	1.06848	1.06148	1.04660	1.03454
173	1.07860	1.07231	1.06647	1.05948	1.04465	1.03269
174	1.07657	1.07029	1.06446	1.05748	1.04269	1.03084
175	1.07453	1.06827	1.06245	1.05548	1.04073	1.02898
176	1.07250	1.06626	1.06044	1.05348	1.03878	1.02713
177	1.07047	1.06424	1.05843	1.05148	1.03682	1.02528
178	1.06843	1.06222	1.05642	1.04948	1.03486	1.02342
179	1.06640	1.06020	1.05441	1.04748	1.03291	1.02157
180	1.06436	1.05819	1.05240	1.04548	1.03095	1.01971
181	1.06233	1.05617	1.05039	1.04348	1.02900	1.01786
182	1.06029	1.05415	1.04838	1.04148	1.02704	1.01601
183	1.05826	1.05214	1.04637	1.03948	1.02508	1.01415
184	1.08994	1.08388	1.07872	1.07242	1.05801	1.04182

continued to next page...

Table 4.2: Showing the p-th quantiles of the *RankCover* statistic

Sample Sizes	p=0.1	p=0.05	p=0.025	p=0.01	p=0.001	p=0.0001
185	1.08820	1.08216	1.07699	1.07071	1.05634	1.04025
186	1.08647	1.08044	1.07526	1.06899	1.05468	1.03868
187	1.08473	1.07872	1.07353	1.06728	1.05301	1.03711
188	1.08300	1.07700	1.07181	1.06556	1.05134	1.03554
189	1.08126	1.07527	1.07008	1.06384	1.04968	1.03397
190	1.07952	1.07355	1.06835	1.06213	1.04801	1.03240
191	1.07779	1.07183	1.06663	1.06041	1.04635	1.03083
192	1.07605	1.07011	1.06490	1.05869	1.04468	1.02926
193	1.07432	1.06839	1.06317	1.05698	1.04301	1.02769
194	1.07258	1.06666	1.06145	1.05526	1.04135	1.02612
195	1.07084	1.06494	1.05972	1.05354	1.03968	1.02455
196	1.06911	1.06322	1.05799	1.05183	1.03801	1.02298
197	1.06737	1.06150	1.05626	1.05011	1.03635	1.02141
198	1.06564	1.05977	1.05454	1.04840	1.03468	1.01984
199	1.06390	1.05805	1.05281	1.04668	1.03301	1.01827
200	1.06216	1.05633	1.05108	1.04496	1.03135	1.01670
201	1.06043	1.05461	1.04936	1.04325	1.02968	1.01513
202	1.05869	1.05289	1.04763	1.04153	1.02802	1.01356
203	1.05696	1.05116	1.04590	1.03981	1.02635	1.01199
204	1.05522	1.04944	1.04417	1.03810	1.02468	1.01042
205	1.05348	1.04772	1.04245	1.03638	1.02302	1.00885
206	1.05175	1.04600	1.04072	1.03466	1.02135	1.00728
207	1.08098	1.07536	1.07029	1.06418	1.05092	1.03650
208	1.07947	1.07385	1.06879	1.06268	1.04946	1.03508
209	1.07795	1.07234	1.06729	1.06119	1.04800	1.03365
210	1.07643	1.07083	1.06579	1.05969	1.04654	1.03223
211	1.07492	1.06932	1.06429	1.05820	1.04508	1.03081
212	1.07340	1.06781	1.06279	1.05670	1.04361	1.02938
213	1.07189	1.06630	1.06129	1.05521	1.04215	1.02796
214	1.07037	1.06479	1.05979	1.05371	1.04069	1.02654
215	1.06886	1.06328	1.05829	1.05222	1.03923	1.02511
216	1.06734	1.06177	1.05679	1.05073	1.03777	1.02369
217	1.06582	1.06026	1.05529	1.04923	1.03631	1.02227
218	1.06431	1.05875	1.05379	1.04774	1.03484	1.02084
219	1.06279	1.05724	1.05229	1.04624	1.03338	1.01942
220	1.06128	1.05573	1.05078	1.04475	1.03192	1.01800
221	1.05976	1.05422	1.04928	1.04325	1.03046	1.01657
222	1.05825	1.05271	1.04778	1.04176	1.02900	1.01515
223	1.05673	1.05120	1.04628	1.04026	1.02753	1.01373
224	1.05521	1.04969	1.04478	1.03877	1.02607	1.01230
225	1.05370	1.04818	1.04328	1.03727	1.02461	1.01088
226	1.05218	1.04667	1.04178	1.03578	1.02315	1.00946

continued to next page...

Table 4.2: Showing the p-th quantiles of the *RankCover* statistic

Sample Sizes	p=0.1	p=0.05	p=0.025	p=0.01	p=0.001	p=0.0001
227	1.05067	1.04516	1.04028	1.03428	1.02169	1.00803
228	1.04915	1.04365	1.03878	1.03279	1.02023	1.00661
229	1.04763	1.04214	1.03728	1.03129	1.01876	1.00519
230	1.07488	1.06968	1.06493	1.05941	1.04715	1.03673
231	1.07355	1.06836	1.06362	1.05811	1.04591	1.03557
232	1.07222	1.06704	1.06231	1.05680	1.04466	1.03441
233	1.07090	1.06571	1.06100	1.05550	1.04342	1.03325
234	1.06957	1.06439	1.05969	1.05419	1.04218	1.03210
235	1.06824	1.06307	1.05838	1.05289	1.04094	1.03094
236	1.06691	1.06175	1.05707	1.05158	1.03970	1.02978
237	1.06559	1.06043	1.05576	1.05028	1.03846	1.02862
238	1.06426	1.05911	1.05444	1.04897	1.03722	1.02746
239	1.06293	1.05778	1.05313	1.04767	1.03598	1.02630
240	1.06161	1.05646	1.05182	1.04636	1.03474	1.02515
241	1.06028	1.05514	1.05051	1.04506	1.03350	1.02399
242	1.05895	1.05382	1.04920	1.04375	1.03226	1.02283
243	1.05763	1.05250	1.04789	1.04244	1.03102	1.02167
244	1.05630	1.05118	1.04658	1.04114	1.02978	1.02051
245	1.05497	1.04985	1.04527	1.03983	1.02854	1.01935
246	1.05364	1.04853	1.04395	1.03853	1.02730	1.01820
247	1.05232	1.04721	1.04264	1.03722	1.02606	1.01704
248	1.05099	1.04589	1.04133	1.03592	1.02482	1.01588
249	1.04966	1.04457	1.04002	1.03461	1.02358	1.01472
250	1.04834	1.04325	1.03871	1.03331	1.02234	1.01356
251	1.04701	1.04192	1.03740	1.03200	1.02110	1.01240
252	1.04568	1.04060	1.03609	1.03070	1.01986	1.01124
253	1.04436	1.03928	1.03478	1.02939	1.01862	1.01009
254	1.04303	1.03796	1.03346	1.02809	1.01738	1.00893
255	1.06817	1.06341	1.05913	1.05390	1.04269	1.03194
256	1.06702	1.06225	1.05796	1.05274	1.04155	1.03085
257	1.06587	1.06109	1.05680	1.05159	1.04041	1.02975
258	1.06471	1.05993	1.05563	1.05043	1.03926	1.02865
259	1.06356	1.05877	1.05447	1.04927	1.03812	1.02756
260	1.06241	1.05761	1.05330	1.04811	1.03698	1.02646
261	1.06126	1.05645	1.05213	1.04695	1.03583	1.02537
262	1.06010	1.05529	1.05097	1.04580	1.03469	1.02427
263	1.05895	1.05413	1.04980	1.04464	1.03355	1.02318
264	1.05780	1.05297	1.04863	1.04348	1.03241	1.02208
265	1.05664	1.05181	1.04747	1.04232	1.03126	1.02099
266	1.05549	1.05066	1.04630	1.04116	1.03012	1.01989
267	1.05434	1.04950	1.04514	1.04000	1.02898	1.01880
268	1.05319	1.04834	1.04397	1.03885	1.02783	1.01770

continued to next page...

Table 4.2: Showing the p-th quantiles of the *RankCover* statistic

Sample Sizes	p=0.1	p=0.05	p=0.025	p=0.01	p=0.001	p=0.0001
269	1.05203	1.04718	1.04280	1.03769	1.02669	1.01660
270	1.05088	1.04602	1.04164	1.03653	1.02555	1.01551
271	1.04973	1.04486	1.04047	1.03537	1.02441	1.01441
272	1.04857	1.04370	1.03930	1.03421	1.02326	1.01332
273	1.04742	1.04254	1.03814	1.03306	1.02212	1.01222
274	1.04627	1.04138	1.03697	1.03190	1.02098	1.01113
275	1.04511	1.04022	1.03581	1.03074	1.01983	1.01003
276	1.04396	1.03906	1.03464	1.02958	1.01869	1.00894
277	1.04281	1.03790	1.03347	1.02842	1.01755	1.00784
278	1.04166	1.03674	1.03231	1.02727	1.01641	1.00674
279	1.04050	1.03559	1.03114	1.02611	1.01526	1.00565
280	1.03935	1.03443	1.02997	1.02495	1.01412	1.00455
281	1.06263	1.05794	1.05390	1.04900	1.03793	1.02781
282	1.06161	1.05693	1.05288	1.04798	1.03696	1.02694
283	1.06059	1.05591	1.05187	1.04696	1.03600	1.02607
284	1.05957	1.05490	1.05085	1.04594	1.03503	1.02520
285	1.05855	1.05389	1.04983	1.04492	1.03406	1.02433
286	1.05753	1.05288	1.04882	1.04390	1.03310	1.02347
287	1.05652	1.05186	1.04780	1.04288	1.03213	1.02260
288	1.05550	1.05085	1.04679	1.04186	1.03117	1.02173
289	1.05448	1.04984	1.04577	1.04084	1.03020	1.02086
290	1.05346	1.04883	1.04475	1.03982	1.02923	1.01999
291	1.05244	1.04781	1.04374	1.03880	1.02827	1.01912
292	1.05143	1.04680	1.04272	1.03778	1.02730	1.01825
293	1.05041	1.04579	1.04170	1.03676	1.02633	1.01738
294	1.04939	1.04478	1.04069	1.03574	1.02537	1.01651
295	1.04837	1.04376	1.03967	1.03472	1.02440	1.01564
296	1.04735	1.04275	1.03865	1.03369	1.02343	1.01477
297	1.04633	1.04174	1.03764	1.03267	1.02247	1.01390
298	1.04532	1.04073	1.03662	1.03165	1.02150	1.01303
299	1.04430	1.03971	1.03561	1.03063	1.02054	1.01217
300	1.04328	1.03870	1.03459	1.02961	1.01957	1.01130
301	1.04226	1.03769	1.03357	1.02859	1.01860	1.01043
302	1.04124	1.03668	1.03256	1.02757	1.01764	1.00956
303	1.04023	1.03566	1.03154	1.02655	1.01667	1.00869
304	1.03921	1.03465	1.03052	1.02553	1.01570	1.00782
305	1.03819	1.03364	1.02951	1.02451	1.01474	1.00695
306	1.03717	1.03263	1.02849	1.02349	1.01377	1.00608
307	1.03615	1.03161	1.02748	1.02247	1.01280	1.00521
308	1.03513	1.03060	1.02646	1.02145	1.01184	1.00434
309	1.05700	1.05279	1.04895	1.04417	1.03414	1.02581
310	1.05608	1.05187	1.04804	1.04325	1.03324	1.02492

continued to next page...

Table 4.2: Showing the p-th quantiles of the *RankCover* statistic

Sample Sizes	p=0.1	p=0.05	p=0.025	p=0.01	p=0.001	p=0.0001
311	1.05517	1.05096	1.04712	1.04234	1.03233	1.02402
312	1.05426	1.05004	1.04621	1.04143	1.03143	1.02313
313	1.05335	1.04913	1.04529	1.04052	1.03052	1.02224
314	1.05244	1.04821	1.04438	1.03960	1.02962	1.02135
315	1.05153	1.04730	1.04347	1.03869	1.02871	1.02046
316	1.05062	1.04638	1.04255	1.03778	1.02781	1.01957
317	1.04971	1.04547	1.04164	1.03687	1.02691	1.01868
318	1.04880	1.04456	1.04072	1.03595	1.02600	1.01779
319	1.04789	1.04364	1.03981	1.03504	1.02510	1.01690
320	1.04698	1.04273	1.03889	1.03413	1.02419	1.01601
321	1.04606	1.04181	1.03798	1.03322	1.02329	1.01512
322	1.04515	1.04090	1.03706	1.03230	1.02238	1.01423
323	1.04424	1.03998	1.03615	1.03139	1.02148	1.01333
324	1.04333	1.03907	1.03523	1.03048	1.02057	1.01244
325	1.04242	1.03816	1.03432	1.02957	1.01967	1.01155
326	1.04151	1.03724	1.03341	1.02866	1.01876	1.01066
327	1.04060	1.03633	1.03249	1.02774	1.01786	1.00977
328	1.03969	1.03541	1.03158	1.02683	1.01695	1.00888
329	1.03878	1.03450	1.03066	1.02592	1.01605	1.00799
330	1.03787	1.03358	1.02975	1.02501	1.01514	1.00710
331	1.03696	1.03267	1.02883	1.02409	1.01424	1.00621
332	1.03605	1.03175	1.02792	1.02318	1.01334	1.00532
333	1.03513	1.03084	1.02700	1.02227	1.01243	1.00443
334	1.03422	1.02993	1.02609	1.02136	1.01153	1.00354
335	1.03331	1.02901	1.02518	1.02044	1.01062	1.00264
336	1.03240	1.02810	1.02426	1.01953	1.00972	1.00175
337	1.05310	1.04897	1.04528	1.04088	1.03216	1.02235
338	1.05227	1.04815	1.04446	1.04007	1.03131	1.02156
339	1.05145	1.04733	1.04364	1.03926	1.03046	1.02077
340	1.05063	1.04651	1.04282	1.03844	1.02961	1.01999
341	1.04980	1.04569	1.04200	1.03763	1.02876	1.01920
342	1.04898	1.04487	1.04118	1.03682	1.02791	1.01841
343	1.04816	1.04405	1.04037	1.03601	1.02706	1.01763
344	1.04734	1.04323	1.03955	1.03519	1.02622	1.01684
345	1.04651	1.04241	1.03873	1.03438	1.02537	1.01605
346	1.04569	1.04159	1.03791	1.03357	1.02452	1.01526
347	1.04487	1.04077	1.03709	1.03275	1.02367	1.01448
348	1.04404	1.03995	1.03627	1.03194	1.02282	1.01369
349	1.04322	1.03913	1.03546	1.03113	1.02197	1.01290
350	1.04240	1.03830	1.03464	1.03032	1.02112	1.01212
351	1.04158	1.03748	1.03382	1.02950	1.02028	1.01133
352	1.04075	1.03666	1.03300	1.02869	1.01943	1.01054

continued to next page...

Table 4.2: Showing the p-th quantiles of the *RankCover* statistic

Sample Sizes	p=0.1	p=0.05	p=0.025	p=0.01	p=0.001	p=0.0001
353	1.03993	1.03584	1.03218	1.02788	1.01858	1.00975
354	1.03911	1.03502	1.03136	1.02707	1.01773	1.00897
355	1.03828	1.03420	1.03055	1.02625	1.01688	1.00818
356	1.03746	1.03338	1.02973	1.02544	1.01603	1.00739
357	1.03664	1.03256	1.02891	1.02463	1.01518	1.00661
358	1.03582	1.03174	1.02809	1.02381	1.01434	1.00582
359	1.03499	1.03092	1.02727	1.02300	1.01349	1.00503
360	1.03417	1.03010	1.02645	1.02219	1.01264	1.00424
361	1.03335	1.02928	1.02564	1.02138	1.01179	1.00346
362	1.03252	1.02846	1.02482	1.02056	1.01094	1.00267
363	1.03170	1.02763	1.02400	1.01975	1.01009	1.00188
364	1.03088	1.02681	1.02318	1.01894	1.00924	1.00110
365	1.03006	1.02599	1.02236	1.01812	1.00840	1.00031
366	1.02923	1.02517	1.02154	1.01731	1.00755	0.99952
367	1.04869	1.04481	1.04132	1.03715	1.02815	1.01890
368	1.04795	1.04407	1.04059	1.03642	1.02743	1.01826
369	1.04722	1.04333	1.03986	1.03569	1.02670	1.01762
370	1.04648	1.04260	1.03912	1.03496	1.02598	1.01699
371	1.04575	1.04186	1.03839	1.03423	1.02525	1.01635
372	1.04501	1.04113	1.03765	1.03350	1.02453	1.01571
373	1.04428	1.04039	1.03692	1.03277	1.02380	1.01507
374	1.04354	1.03965	1.03619	1.03204	1.02308	1.01443
375	1.04281	1.03892	1.03545	1.03131	1.02235	1.01379
376	1.04207	1.03818	1.03472	1.03058	1.02162	1.01315
377	1.04134	1.03744	1.03398	1.02985	1.02090	1.01251
378	1.04060	1.03671	1.03325	1.02912	1.02017	1.01187
379	1.03987	1.03597	1.03252	1.02839	1.01945	1.01123
380	1.03913	1.03524	1.03178	1.02766	1.01872	1.01059
381	1.03840	1.03450	1.03105	1.02693	1.01800	1.00995
382	1.03766	1.03376	1.03031	1.02620	1.01727	1.00931
383	1.03693	1.03303	1.02958	1.02547	1.01655	1.00868
384	1.03619	1.03229	1.02884	1.02474	1.01582	1.00804
385	1.03546	1.03156	1.02811	1.02401	1.01510	1.00740
386	1.03472	1.03082	1.02738	1.02328	1.01437	1.00676
387	1.03399	1.03008	1.02664	1.02255	1.01364	1.00612
388	1.03325	1.02935	1.02591	1.02182	1.01292	1.00548
389	1.03252	1.02861	1.02517	1.02109	1.01219	1.00484
390	1.03178	1.02787	1.02444	1.02036	1.01147	1.00420
391	1.03105	1.02714	1.02371	1.01963	1.01074	1.00356
392	1.03031	1.02640	1.02297	1.01890	1.01002	1.00292
393	1.02958	1.02567	1.02224	1.01817	1.00929	1.00228
394	1.02884	1.02493	1.02150	1.01744	1.00857	1.00164

continued to next page...

Table 4.2: Showing the p-th quantiles of the *RankCover* statistic

Sample Sizes	p=0.1	p=0.05	p=0.025	p=0.01	p=0.001	p=0.0001
395	1.02811	1.02419	1.02077	1.01671	1.00784	1.00101
396	1.02737	1.02346	1.02004	1.01598	1.00711	1.00037
397	1.02664	1.02272	1.01930	1.01525	1.00639	0.99973
398	1.04487	1.04108	1.03781	1.03391	1.02503	1.01735
399	1.04421	1.04042	1.03715	1.03325	1.02437	1.01671
400	1.04354	1.03976	1.03648	1.03258	1.02371	1.01607
401	1.04288	1.03910	1.03582	1.03192	1.02305	1.01543
402	1.04221	1.03844	1.03516	1.03126	1.02239	1.01479
403	1.04155	1.03778	1.03449	1.03059	1.02173	1.01415
404	1.04088	1.03711	1.03383	1.02993	1.02107	1.01351
405	1.04022	1.03645	1.03317	1.02926	1.02041	1.01287
406	1.03955	1.03579	1.03250	1.02860	1.01975	1.01223
407	1.03889	1.03513	1.03184	1.02793	1.01909	1.01159
408	1.03822	1.03447	1.03118	1.02727	1.01843	1.01095
409	1.03756	1.03380	1.03051	1.02660	1.01777	1.01031
410	1.03689	1.03314	1.02985	1.02594	1.01710	1.00967
411	1.03623	1.03248	1.02919	1.02528	1.01644	1.00903
412	1.03556	1.03182	1.02852	1.02461	1.01578	1.00839
413	1.03490	1.03116	1.02786	1.02395	1.01512	1.00775
414	1.03423	1.03049	1.02720	1.02328	1.01446	1.00711
415	1.03357	1.02983	1.02653	1.02262	1.01380	1.00647
416	1.03290	1.02917	1.02587	1.02195	1.01314	1.00583
417	1.03224	1.02851	1.02521	1.02129	1.01248	1.00519
418	1.03158	1.02785	1.02454	1.02062	1.01182	1.00455
419	1.03091	1.02718	1.02388	1.01996	1.01116	1.00391
420	1.03025	1.02652	1.02322	1.01929	1.01050	1.00327
421	1.02958	1.02586	1.02255	1.01863	1.00984	1.00263
422	1.02892	1.02520	1.02189	1.01797	1.00918	1.00199
423	1.02825	1.02454	1.02123	1.01730	1.00852	1.00135
424	1.02759	1.02387	1.02056	1.01664	1.00786	1.00071
425	1.02692	1.02321	1.01990	1.01597	1.00720	1.00007
426	1.02626	1.02255	1.01924	1.01531	1.00654	0.99943
427	1.02559	1.02189	1.01857	1.01464	1.00588	0.99879
428	1.02493	1.02123	1.01791	1.01398	1.00522	0.99815
429	1.02426	1.02056	1.01725	1.01331	1.00455	0.99751
430	1.02360	1.01990	1.01658	1.01265	1.00389	0.99687
431	1.04110	1.03745	1.03418	1.03052	1.02236	1.01457
432	1.04050	1.03685	1.03359	1.02991	1.02176	1.01399
433	1.03990	1.03625	1.03299	1.02931	1.02116	1.01342
434	1.03929	1.03565	1.03239	1.02871	1.02056	1.01285
435	1.03869	1.03505	1.03179	1.02810	1.01995	1.01227
436	1.03809	1.03445	1.03119	1.02750	1.01935	1.01170

continued to next page...

Table 4.2: Showing the p-th quantiles of the *RankCover* statistic

Sample Sizes	p=0.1	p=0.05	p=0.025	p=0.01	p=0.001	p=0.0001
437	1.03749	1.03385	1.03059	1.02690	1.01875	1.01112
438	1.03688	1.03325	1.02999	1.02630	1.01815	1.01055
439	1.03628	1.03265	1.02940	1.02569	1.01755	1.00997
440	1.03568	1.03205	1.02880	1.02509	1.01694	1.00940
441	1.03507	1.03145	1.02820	1.02449	1.01634	1.00883
442	1.03447	1.03085	1.02760	1.02388	1.01574	1.00825
443	1.03387	1.03025	1.02700	1.02328	1.01514	1.00768
444	1.03327	1.02965	1.02640	1.02268	1.01454	1.00710
445	1.03266	1.02905	1.02580	1.02207	1.01393	1.00653
446	1.03206	1.02845	1.02521	1.02147	1.01333	1.00596
447	1.03146	1.02785	1.02461	1.02087	1.01273	1.00538
448	1.03086	1.02725	1.02401	1.02026	1.01213	1.00481
449	1.03025	1.02665	1.02341	1.01966	1.01152	1.00423
450	1.02965	1.02605	1.02281	1.01906	1.01092	1.00366
451	1.02905	1.02545	1.02221	1.01845	1.01032	1.00309
452	1.02844	1.02485	1.02161	1.01785	1.00972	1.00251
453	1.02784	1.02425	1.02102	1.01725	1.00912	1.00194
454	1.02724	1.02365	1.02042	1.01664	1.00851	1.00136
455	1.02664	1.02305	1.01982	1.01604	1.00791	1.00079
456	1.02603	1.02245	1.01922	1.01544	1.00731	1.00022
457	1.02543	1.02185	1.01862	1.01484	1.00671	0.99964
458	1.02483	1.02125	1.01802	1.01423	1.00611	0.99907
459	1.02423	1.02065	1.01742	1.01363	1.00550	0.99849
460	1.02362	1.02005	1.01683	1.01303	1.00490	0.99792
461	1.02302	1.01945	1.01623	1.01242	1.00430	0.99735
462	1.02242	1.01885	1.01563	1.01182	1.00370	0.99677
463	1.02182	1.01825	1.01503	1.01122	1.00310	0.99620
464	1.02121	1.01765	1.01443	1.01061	1.00249	0.99562
465	1.03769	1.03430	1.03122	1.02757	1.01995	1.01307
466	1.03714	1.03375	1.03067	1.02702	1.01939	1.01253
467	1.03659	1.03319	1.03012	1.02647	1.01884	1.01199
468	1.03604	1.03264	1.02957	1.02592	1.01828	1.01146
469	1.03549	1.03209	1.02902	1.02537	1.01773	1.01092
470	1.03494	1.03154	1.02847	1.02482	1.01717	1.01038
471	1.03439	1.03099	1.02791	1.02427	1.01662	1.00984
472	1.03384	1.03043	1.02736	1.02372	1.01606	1.00931
473	1.03329	1.02988	1.02681	1.02317	1.01551	1.00877
474	1.03274	1.02933	1.02626	1.02262	1.01495	1.00823
475	1.03219	1.02878	1.02571	1.02207	1.01440	1.00769
476	1.03164	1.02823	1.02516	1.02152	1.01384	1.00716
477	1.03109	1.02767	1.02461	1.02097	1.01329	1.00662
478	1.03054	1.02712	1.02406	1.02042	1.01273	1.00608

continued to next page...

Table 4.2: Showing the p-th quantiles of the *RankCover* statistic

Sample Sizes	p=0.1	p=0.05	p=0.025	p=0.01	p=0.001	p=0.0001
479	1.02999	1.02657	1.02351	1.01987	1.01218	1.00554
480	1.02944	1.02602	1.02296	1.01932	1.01163	1.00501
481	1.02889	1.02547	1.02241	1.01876	1.01107	1.00447
482	1.02834	1.02491	1.02186	1.01821	1.01052	1.00393
483	1.02779	1.02436	1.02131	1.01766	1.00996	1.00339
484	1.02724	1.02381	1.02076	1.01711	1.00941	1.00286
485	1.02669	1.02326	1.02021	1.01656	1.00885	1.00232
486	1.02614	1.02271	1.01966	1.01601	1.00830	1.00178
487	1.02559	1.02215	1.01911	1.01546	1.00774	1.00124
488	1.02504	1.02160	1.01856	1.01491	1.00719	1.00071
489	1.02449	1.02105	1.01801	1.01436	1.00663	1.00017
490	1.02394	1.02050	1.01746	1.01381	1.00608	0.99963
491	1.02340	1.01995	1.01691	1.01326	1.00552	0.99910
492	1.02285	1.01939	1.01636	1.01271	1.00497	0.99856
493	1.02230	1.01884	1.01580	1.01216	1.00441	0.99802
494	1.02175	1.01829	1.01525	1.01161	1.00386	0.99748
495	1.02120	1.01774	1.01470	1.01106	1.00330	0.99695
496	1.02065	1.01719	1.01415	1.01051	1.00275	0.99641
497	1.02010	1.01663	1.01360	1.00996	1.00219	0.99587
498	1.01955	1.01608	1.01305	1.00941	1.00164	0.99533
499	1.01900	1.01553	1.01250	1.00886	1.00108	0.99480
500	1.03464	1.03135	1.02835	1.02489	1.01768	1.01220

Table 4.3: Showing the p-th quantiles of the hybrid p-values

Sample Sizes	p=0.1	p=0.05	p=0.025	p=0.01	p=0.001	p=0.0001
20	0.06682	0.03219	0.01591	0.00659	0.00065	0.00007
21	0.06464	0.03226	0.01573	0.00632	0.00062	0.00005
22	0.06512	0.03175	0.01531	0.00620	0.00062	0.00006
23	0.06590	0.03182	0.01594	0.00618	0.00070	0.00007
24	0.06512	0.03226	0.01601	0.00647	0.00062	0.00006
25	0.06479	0.03165	0.01585	0.00622	0.00064	0.00006
26	0.06357	0.03102	0.01566	0.00631	0.00063	0.00006
27	0.06366	0.03112	0.01531	0.00607	0.00062	0.00007
28	0.06309	0.03098	0.01512	0.00599	0.00061	0.00007
29	0.06346	0.03038	0.01503	0.00600	0.00060	0.00006
30	0.06293	0.03024	0.01495	0.00590	0.00057	0.00005
31	0.06257	0.03042	0.01481	0.00586	0.00058	0.00007
32	0.06206	0.03057	0.01498	0.00579	0.00058	0.00006
33	0.06207	0.03016	0.01502	0.00606	0.00056	0.00005
34	0.06169	0.03010	0.01497	0.00592	0.00058	0.00005

continued to next page...

Table 4.3: Showing the p-th quantiles of the hybrid p-values

Sample Sizes	p=0.1	p=0.05	p=0.025	p=0.01	p=0.001	p=0.0001
35	0.06171	0.03003	0.01490	0.00584	0.00059	0.00005
36	0.06153	0.03016	0.01465	0.00572	0.00058	0.00007
37	0.06146	0.02965	0.01465	0.00574	0.00058	0.00005
38	0.06027	0.02944	0.01454	0.00565	0.00057	0.00006
39	0.06069	0.02942	0.01447	0.00566	0.00055	0.00005
40	0.06032	0.02926	0.01420	0.00564	0.00056	0.00005
41	0.06083	0.02960	0.01451	0.00572	0.00052	0.00006
42	0.05989	0.02923	0.01446	0.00574	0.00055	0.00005
43	0.06029	0.02928	0.01470	0.00577	0.00054	0.00005
44	0.06045	0.02904	0.01431	0.00558	0.00055	0.00006
45	0.05955	0.02907	0.01417	0.00570	0.00057	0.00005
46	0.05950	0.02885	0.01419	0.00553	0.00054	0.00006
47	0.05953	0.02891	0.01410	0.00558	0.00055	0.00006
48	0.05962	0.02908	0.01415	0.00553	0.00055	0.00005
49	0.05952	0.02874	0.01410	0.00549	0.00054	0.00005
50	0.05903	0.02858	0.01417	0.00561	0.00053	0.00005
51	0.05913	0.02857	0.01412	0.00558	0.00055	0.00006
52	0.05933	0.02896	0.01406	0.00540	0.00053	0.00005
53	0.05925	0.02884	0.01417	0.00557	0.00053	0.00004
54	0.05918	0.02879	0.01408	0.00553	0.00052	0.00005
55	0.05867	0.02861	0.01393	0.00550	0.00055	0.00005
56	0.05871	0.02862	0.01423	0.00554	0.00052	0.00005
57	0.05874	0.02838	0.01382	0.00552	0.00056	0.00005
58	0.05874	0.02865	0.01394	0.00543	0.00053	0.00005
59	0.05868	0.02845	0.01408	0.00550	0.00052	0.00006
60	0.05843	0.02840	0.01378	0.00545	0.00053	0.00005
61	0.05849	0.02840	0.01398	0.00547	0.00055	0.00006
62	0.05806	0.02831	0.01390	0.00541	0.00052	0.00005
63	0.05810	0.02812	0.01372	0.00546	0.00053	0.00005
64	0.05832	0.02852	0.01391	0.00544	0.00053	0.00005
65	0.05770	0.02831	0.01386	0.00543	0.00054	0.00006
66	0.05831	0.02817	0.01378	0.00543	0.00052	0.00005
67	0.05779	0.02805	0.01379	0.00539	0.00051	0.00005
68	0.05776	0.02800	0.01379	0.00551	0.00054	0.00005
69	0.05768	0.02780	0.01356	0.00534	0.00052	0.00005
70	0.05776	0.02806	0.01382	0.00536	0.00054	0.00005
71	0.05744	0.02778	0.01368	0.00536	0.00053	0.00005
72	0.05746	0.02776	0.01363	0.00540	0.00053	0.00005
73	0.05736	0.02798	0.01370	0.00544	0.00052	0.00005
74	0.05709	0.02770	0.01358	0.00537	0.00053	0.00005
75	0.05712	0.02766	0.01352	0.00526	0.00051	0.00005
76	0.05675	0.02759	0.01362	0.00538	0.00053	0.00006

continued to next page...

Table 4.3: Showing the p-th quantiles of the hybrid p-values

Sample Sizes	p=0.1	p=0.05	p=0.025	p=0.01	p=0.001	p=0.0001
77	0.05669	0.02769	0.01355	0.00531	0.00053	0.00005
78	0.05698	0.02765	0.01361	0.00534	0.00052	0.00005
79	0.05656	0.02766	0.01354	0.00538	0.00053	0.00005
80	0.05704	0.02791	0.01359	0.00534	0.00053	0.00005
81	0.05709	0.02756	0.01352	0.00530	0.00051	0.00005
82	0.05706	0.02765	0.01360	0.00530	0.00051	0.00005
83	0.05659	0.02730	0.01336	0.00527	0.00052	0.00005
84	0.05707	0.02768	0.01363	0.00539	0.00051	0.00005
85	0.05680	0.02755	0.01347	0.00526	0.00052	0.00005
86	0.05686	0.02750	0.01362	0.00538	0.00054	0.00005
87	0.05707	0.02766	0.01352	0.00535	0.00053	0.00006
88	0.05655	0.02756	0.01347	0.00530	0.00051	0.00005
89	0.05685	0.02762	0.01352	0.00532	0.00051	0.00005
90	0.05624	0.02709	0.01332	0.00524	0.00053	0.00005
91	0.05628	0.02728	0.01343	0.00527	0.00053	0.00005
92	0.05639	0.02727	0.01346	0.00529	0.00051	0.00005
93	0.05645	0.02736	0.01344	0.00530	0.00051	0.00005
94	0.05645	0.02751	0.01347	0.00536	0.00051	0.00005
95	0.05639	0.02755	0.01353	0.00530	0.00051	0.00005
96	0.05637	0.02727	0.01341	0.00529	0.00051	0.00005
97	0.05646	0.02716	0.01337	0.00528	0.00051	0.00005
98	0.05668	0.02731	0.01350	0.00524	0.00050	0.00005
99	0.05615	0.02736	0.01346	0.00527	0.00053	0.00006
100	0.05616	0.02737	0.01343	0.00527	0.00052	0.00005

APPENDIX C: TECHNICAL DETAILS FOR CHAPTER 3

C.1 Pre-processing of the GTEx data

We have used the GTEx pilot data (Lonsdale et al. 2013) for our analysis of real data. The data consists of genotype and expression data across nine tissues - adipose, artery, heart, lung, muscle, nerve, skin, thyroid and blood. There were 175 genotyped individuals who had expression data in at least one of the tissues. The tissues had shared samples in the sense that many individuals had expression data for more than one tissues. The sample sizes corresponding to these tissues were respectively $94, 112, 83, 119, 138, 88, 96, 105, 156$.

The elements of the genotype matrix are the minor allele frequencies (MAF) of donors in different SNP locations. Any missing value in this matrix was imputed by the average MAF of that locus across all donors. Loci that had less than 5% MAF for all donors were discarded and the final genotype matrix had about 7 million SNPs.

The expression levels were measured by the number of mapped reads per kilobase per million reads (RPKM). Genes with less than 10 donors with RPKM greater than 0.1 in some tissue were discarded resulting in about 22000 common genes. Finally, the expression values were inverse quantile normalized.

The SNPs located within 100 kilobases of the transcription start site of a gene were considered *cis* to that gene. This resulted in about 10 million gene-SNP pairs that were grouped by about 22000 genes.

There were a total of 19 covariates including 15 PEER factors, 3 principal components and 1 gender covariate. For each tissue, both the expression and the genotypes were residualized using linear regression on these 19 covariates. The residualized data were treated as the inputs Y and X of our model. While computing the z-statistics, the scaling factor was adjusted by the loss of degrees of freedom due to such residualization. Therefore the scaling factor $\sqrt{n-22}$ was used instead of the usual $\sqrt{n-3}$.

C.2 Details of the simulation procedures for *Z-REG-FDR*

The simulation procedure mentioned in Section 3.6 was used for simulating the data analysed in any simulation study with one causal SNP. The SNP matrix was prepared in different ways. For the data analysed in Table 3.3, the SNPs were simulated from from AR(1) structured normal distributions. For the data analysed in Table 3.2, the SNP data was picked up from the data on the tissue heart from GTEx. We used only 10000 genes among all the genes having at least 10 and at most 1000

For Table 3.1, however, the z-values were directly simulated from an AR(1) structure instead of using our usual simulation scheme. The only difference for causal locations and non-causal locations was in the expectation of z.

C.3 Details of the simulation procedures for two causal SNPs

For the simulation in case of two causal SNPs, it is important to note that it is not possible to have both the effect sizes unconstrained and maintain the variance of Y_i at the same time. As the correlation between the two causal SNPs, say $X_k^{(i)}$ and $X_l^{(i)}$, is already given by data, the two effect sizes, that are the correlations between Y_i and the respective SNPs, have to be such that the correlation matrix is positive definite. It can be shown that the required condition is

$$1 + 2\rho\rho_1\rho_2 - \rho^2 - \rho_1^2 - \rho_2^2 \geq 0$$

where $\rho = Cor(X_k^{(i)}, X_l^{(i)})$, $\rho_1 = Cor(X_k^{(i)}, Y_i)$, $\rho_2 = Cor(X_l^{(i)}, Y_i)$. We assume that the second causal SNP is 'secondary' in the sense that $|\rho_2| < |\rho_1|$. This along with $Sign(\rho_2) = Sign(\rho)Sign(\rho_1)$ ensures that the above condition is true, and hence the correlation matrix of $(Y, X_k^{(i)}, X_l^{(i)})$ is positive definite.

For Table 3.4, we used a simulated data where for half of the genes under alternative are assumed to have two causal SNPs and half are assumed to have one causal SNP.

For the genes with two causal SNPs, we simulate $\sqrt{n-3}\ tanh^{-1}(\rho_1)$ from $N(0, \sigma^2)$, and then simulate ρ_2 as

$$\rho_2 = Sign(\rho)Sign(\rho_1)|\rho_1|r$$

r is simulated from a $Beta(1, 2)$ distribution. For a gene with two causal SNPs, given ρ, ρ_1 and ρ_2, Y_i is simulated from a normal distribution with mean $\frac{1}{1-\rho^2}\{X_k^{(i)}(\rho_1 - \rho\rho_2) + X_l^{(i)}(\rho_2 - \rho\rho_1)\}$ and variance $\frac{1+2\rho\rho_1\rho_2-\rho^2-\rho_1^2-\rho_2^2}{1-\rho^2}$. Usual computation for conditional distribution of a multivariate normal distribution shows that such simulation procedure generates the expression so that the desired variance and covariances are maintained.

C.4 Details of the simulation procedures for inverse average

We simulated 20000 genes and 100 SNPs per gene for the inverse average examples. For the block example, it was assumed that the correlation within each block is 1 and outside block is 0. The block size was 10. The gene-SNP level lfdr were calculated based on such assumption. For the window type model, it was assumed that all the SNPs within a window of size 10 around the causal SNP will have significant gene-SNP level lfdr.

REFERENCES

Ailam, G. (1966), "Moments of coverage and coverage spaces," *Journal of Applied Probability*, 3, 550–555.

Albert, P. S., Ratnasinghe, D., Tangrea, J., and Wacholder, S. (2001), "Limitations of the case-only design for identifying gene-environment interactions," *American Journal of Epidemiology*, 154, 687–693.

Ardlie, K. G., Deluca, D. S., Segrè, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., Trowbridge, C. A., Maller, J. B., Tukiainen, T., Lek, M., et al. (2015), "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans," *Science*, 348, 648–660.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000), "Gene Ontology: tool for the unification of biology," *Nature genetics*, 25, 25–29.

Benjamini, Y. (2010), "Discovering the false discovery rate," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 405–416.

Benjamini, Y. and Bogomolov, M. (2014), "Selective inference on multiple families of hypotheses," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 297–318.

Benjamini, Y., Gavrilov, Y., et al. (2009), "A simple forward selection procedure based on false discovery rate control," *The Annals of Applied Statistics*, 3, 179–198.

Benjamini, Y. and Heller, R. (2007), "False discovery rates for spatial signals," *Journal of the American Statistical Association*, 102, 1272–1281.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

— (1997), "Multiple hypotheses testing with weights," *Scandinavian Journal of Statistics*, 24, 407–418.

— (2000), "On the adaptive control of the false discovery rate in multiple testing with independent statistics," *Journal of Educational and Behavioral Statistics*, 25, 60–83.

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006), "Adaptive linear step-up procedures that control the false discovery rate," *Biometrika*, 93, 491–507.

Benjamini, Y. and Yekutieli, D. (2001), "The control of the false discovery rate in multiple testing under dependency," *Annals of statistics*, 1165–1188.

— (2005), "Quantitative trait loci analysis using the false discovery rate," *Genetics*, 171, 783–790.

Bickel, P. and Xu, Y. (2009), "Discussion of: Brownian distance covariance," *The annals of applied statistics*, 1266–1269.

Birch, M. (1965), "The detection of partial association, II: the general case," *Journal of the Royal Statistical Society. Series B (Methodological)*, 111–124.

Black, M. (2004), "A note on the adaptive control of false discovery rates," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66, 297–304.

Blanchard, G. and Roquain, É. (2009), "Adaptive false discovery rate control under independence and dependence," *The Journal of Machine Learning Research*, 10, 2837–2871.

Bowman, A. and Azzalini, A. (2013), "R package sm: nonparametric smoothing methods (version 2.2-5)," .

Bowman, A. W. and Azzalini, A. (1997), *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations: The Kernel Approach with S-Plus Illustrations*, Oxford University Press.

Breiman, L. and Friedman, J. H. (1985), "Estimating optimal transformations for multiple regression and correlation," *Journal of the American statistical Association*, 80, 580–598.

Cai, T. T. and Sun, W. (2009), "Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks," *Journal of the American Statistical Association*, 104.

Cao, H., Sun, W., and Kosorok, M. R. (2013), "The optimal power puzzle: scrutiny of the monotone likelihood ratio assumption in multiple testing," *Biometrika*, ast001.

Casella, G. (1985), "An introduction to empirical Bayes data analysis," *The American Statistician*, 39, 83–87.

Chen, M. and Kendziorski, C. (2007), "A statistical framework for expression quantitative trait loci mapping," *Genetics*, 177, 761–771.

Chun, H. and Keleş, S. (2009), "Expression quantitative trait loci mapping with multivariate sparse partial least squares regression," *Genetics*, 182, 79–90.

Clark, P. J. and Evans, F. C. (1954), "Distance to nearest neighbor as a measure of spatial relationships in populations," *Ecology*, 445–453.

Cox, D. R. and Hinkley, D. V. (1979), *Theoretical statistics*, CRC Press.

Cuzick, J. (1985), "A Wilcoxon-type test for trend," *Statistics in medicine*, 4, 543–547.

Daub, C. O., Steuer, R., Selbig, J., and Kloska, S. (2004), "Estimating mutual information using B-spline functions–an improved similarity measure for analysing gene expression data," *BMC bioinformatics*, 5, 118.

de Siqueira Santos, S., Takahashi, D. Y., Nakata, A., and Fujita, A. (2013), "A comparative study of statistical methods used to identify dependencies between gene expression signals," *Briefings in bioinformatics*, bbt051.

Diggle, P. J. (1983), *Statistical analysis of spatial point patterns*, Academic Press London.

Diggle, P. J., Besag, J., and Gleaves, J. T. (1976), "Statistical analysis of spatial point patterns by means of distance methods," *Biometrics*, 659–667.

Donnelly, K. (1978), "Simulations to determine the variance and edge-effect of total nearest neighbor distance," *Simulation methods in archaeology*, 91–95.

Dunn, O. J. (1961), "Multiple comparisons among means," *Journal of the American Statistical Association*, 56, 52–64.

Eckerle, K. (1979), "Circular Interference Transmittance Study," *National Institute of Standards and Technology (NIST), US Department of Commerce, USA*.

Efron, B. (2004), "Large-scale simultaneous hypothesis testing," *Journal of the American Statistical Association*, 99.

— (2008), "Simultaneous inference: When should hypothesis testing problems be combined?" *The Annals of Applied Statistics*, 197–223.

Efron, B., Storey, J. D., and Tibshirani, R. (2001a), "Microarrays, Empirical Bayes Methods, and False Discovery Rates," in *Genet. Epidemiol*, Citeseer.

Efron, B. and Tibshirani, R. (2002), "Empirical Bayes methods and false discovery rates for microarrays," *Genetic epidemiology*, 23, 70–86.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001b), "Empirical Bayes analysis of a microarray experiment," *Journal of the American statistical association*, 96, 1151–1160.

Farcomeni, A. (2007a), "A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion," *Statistical Methods in Medical Research*.

— (2007b), "Some results on the control of the false discovery rate under dependence," *Scandinavian Journal of Statistics*, 34, 275–297.

Ferkingstad, E., Frigessi, A., Rue, H., Thorleifsson, G., and Kong, A. (2008), "Unsupervised empirical Bayesian multiple testing with external covariates," *The Annals of Applied Statistics*, 714–735.

Ferreira, J., Zwinderman, A., et al. (2006), "On the Benjamini–Hochberg method," *The Annals of Statistics*, 34, 1827–1849.

Flutre, T., Wen, X., Pritchard, J., and Stephens, M. (2013), "A statistical framework for joint eQTL analysis in multiple tissues," *PLoS genetics*, 9, e1003486.

Fujita, A., Sato, J. R., Demasi, M. A. A., Sogayar, M. C., Ferreira, C. E., and Miyano, S. (2009), "Comparing Pearson, Spearman and Hoeffding's D measure for gene expression association analysis," *Journal of bioinformatics and computational biology*, 7, 663–684.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2007), "Kernel Measures of Conditional Dependence." in *NIPS*, vol. 20, pp. 489–496.

Gatti, D. M., Shabalin, A. A., Lam, T.-C., Wright, F. A., Rusyn, I., and Nobel, A. B. (2009), "FastMap: fast eQTL mapping in homozygous populations," *Bioinformatics*, 25, 482–489.

Gauss, C. (1986), "English translation by A. Clarke," .

Gavrilov, Y., Benjamini, Y., and Sarkar, S. K. (2009), "An adaptive step-down procedure with proven FDR control under independence," *The Annals of Statistics*, 619–629.

Gebelein, H. (1941), "Das statistische Problem der Korrelation als Variations-und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung," *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 21, 364–379.

Gelfond, J. A., Ibrahim, J. G., and Zou, F. (2007), "Proximity model for expression quantitative trait loci (eQTL) detection," *Biometrics*, 63, 1108–1116.

Genovese, C. and Wasserman, L. (2002), "Operating characteristics and extensions of the false discovery rate procedure," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 499–517.

— (2004), "A stochastic process approach to false discovery control," *Annals of Statistics*, 1035–1061.

Genovese, C. R., Roeder, K., and Wasserman, L. (2006), "False discovery control with p-value weighting," *Biometrika*, 93, 509–524.

Goeman, J. J. and Finos, L. (2012), "The inheritance procedure: multiple testing of tree-structured hypotheses," *Statistical applications in genetics and molecular biology*, 11, 1–18.

Grabarnik, P. and Chiu, S. (2002), "Goodness-of-fit test for complete spatial randomness against mixtures of regular and clustered spatial point processes," *Biometrika*, 89, 411–421.

Green, G. H. and Diggle, P. J. (2007), "On the operational characteristics of the Benjamini and Hochberg false discovery rate procedure," *Statistical applications in genetics and molecular biology*, 6.

Gretton, A., Fukumizu, K., and Sriperumbudur, B. K. (2009), "Discussion of: Brownian distance covariance," *The annals of applied statistics*, 1285–1294.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2008), "A kernel statistical test of independence," .

Gretton, A. and Györfi, L. (2008), "Nonparametric independence tests: Space partitioning and kernel approaches," in *Algorithmic Learning Theory*, Springer, pp. 183–198.

Hall, P. (1984), "Mean and variance of vacancy for distribution of k-dimensional spheres within k-dimensional space," *Journal of applied probability*, 738–752.

— (1985), "Three limit theorems for vacancy in multivariate coverage problems," *Journal of Multivariate Analysis*, 16, 211–236.

— (1988), *Introduction to the theory of coverage processes*, John Wiley & Sons Incorporated.

Hall, P. et al. (1985), "On the coverage of k-dimensional space by k-dimensional spheres," *The Annals of Probability*, 13, 991–1002.

Hamed, K. H. and Ramachandra Rao, A. (1998), "A modified Mann-Kendall trend test for autocorrelated data," *Journal of Hydrology*, 204, 182–196.

Heller, R., Heller, Y., and Gorfine, M. (2013), "A consistent multivariate test of association based on ranks of distances," *Biometrika*, 100, 503–510.

Heller, R., Manduchi, E., Grant, G. R., and Ewens, W. J. (2009), "A flexible two-stage procedure for identifying gene sets that are differentially expressed," *Bioinformatics*, 25, 1019–1025.

Hines, W. and Hines, R. O. (1979), "The Eberhardt statistic and the detection of nonrandomness of spatial point distributions," *Biometrika*, 66, 73–79.

Hirschfeld, H. O. (1935), "A connection between correlation and contingency," in *Mathematical Proceedings of the Cambridge Philosophical Society*, Cambridge Univ Press, vol. 31, pp. 520–524.

Hochberg, Y. (1988), "A sharper Bonferroni procedure for multiple tests of significance," *Biometrika*, 75, 800–802.

Hoeffding, W. (1948), "A non-parametric test of independence," *The Annals of Mathematical Statistics*, 546–557.

Holgate, P. (1965a), "Some new tests of randomness," *The Journal of Ecology*, 261–266.

— (1965b), "Tests of randomness based on distance methods," *Biometrika*, 52, 345–353.

Holm, S. (1979), "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, 65–70.

Hu, J. X., Zhao, H., and Zhou, H. H. (2010), "False discovery rate control with groups," *Journal of the American Statistical Association*, 105.

Hubert, L. J. (1985), "Combinatorial data analysis: association and partial association," *Psychometrika*, 50, 449–467.

Jia, Z. and Xu, S. (2007), "Mapping quantitative trait loci for expression abundance," *Genetics*, 176, 611–623.

Jin, J. and Cai, T. T. (2007), "Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons," *Journal of the American Statistical Association*, 102, 495–506.

Jung, K., Friede, T., and Beißbarth, T. (2011), "Reporting FDR analogous confidence intervals for the log fold change of differentially expressed genes," *BMC bioinformatics*, 12, 288.

Kahaner, D., Moler, C. B., Nash, S., and Forsythe, G. E. (1989), *Numerical methods and software*, Prentice-Hall Englewood Cliffs, NJ.

Karr, A. (1991), *Point processes and their statistical inference*, vol. 7, CRC press.

Kendall, M. (1975), *Rank correlation methods*, Griffin, London.

Kendall, M. G. (1938), "A new measure of rank correlation," *Biometrika*, 81–93.

— (1942), "Partial rank correlation," *Biometrika*, 277–283.

Kendziorski, C., Chen, M., Yuan, M., Lan, H., and Attie, A. (2006), "Statistical methods for expression quantitative trait loci (eQTL) mapping," *Biometrics*, 62, 19–27.

Kendziorski, C., Newton, M., Lan, H., and Gould, M. (2003), "On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles," *Statistics in medicine*, 22, 3899–3914.

Kinney, J. B. and Atwal, G. S. (2014a), "Equitability, mutual information, and the maximal information coefficient," *Proceedings of the National Academy of Sciences*, 111, 3354–3359.

— (2014b), "Reply to Reshef et al.: Falsifiability or bust," *Proceedings of the National Academy of Sciences*, 111, E3364–E3364.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004), "Estimating mutual information," *Physical review E*, 69, 066138.

Lehmann, E., Romano, J. P., et al. (2005), "Generalizations of the familywise error rate," *The Annals of Statistics*, 33, 1138–1154.

Lehmann, R. (1977), "General derivation of partial and multiple rank correlation coefficients," *Biometrical Journal*, 19, 229–236.

Li, G., Shabalin, A. A., Rusyn, I., Wright, F. A., and Nobel, A. B. (2013), "An Empirical Bayes Approach for Multiple Tissue eQTL Analysis," *arXiv preprint arXiv:1311.2948*.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013), "The genotype-tissue expression (GTEx) project," *Nature genetics*, 45, 580–585.

Maghsoodloo, S. (1975), "Estimates of the quantiles of Kendall's partial rank correlation coefficient," *Journal of Statistical Computation and Simulation*, 4, 155–164.

Mann, H. B. (1945), "Non-parametric test against trend," *Econometrika*, 13, 245–259.

Meinshausen, N. (2008), "Hierarchical testing of variable importance," *Biometrika*, 95, 265–278.

Miles, R. (1969), "The asymptotic values of certain coverage probabilities," *Biometrika*, 56, 661–680.

Moon, Y.-I., Rajagopalan, B., and Lall, U. (1995), "Estimation of mutual information using kernel density estimators," *Physical Review E*, 52, 2318.

Moran, P. (1951), "Partial and multiple rank correlation," *Biometrika*, 26–32.

— (1973), "A central limit theorem for exchangeable variates with geometric applications," *Journal of Applied Probability*, 837–846.

Moran, P. A. (1974), "The volume occupied by normally distributed spheres," *Acta Mathematica*, 133, 273–286.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004), "Detecting differential gene expression with a semiparametric hierarchical mixture method," *Biostatistics*, 5, 155–176.

Owen, A. B. (2005), "Variance of the number of false discoveries," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 411–426.

Paninski, L. (2003), "Estimation of entropy and mutual information," *Neural Computation*, 15, 1191–1253.

Petretto, E., Bottolo, L., Langley, S. R., Heinig, M., McDermott-Roe, C., Sarwar, R., Pravenec, M., Hübner, N., Aitman, T. J., Cook, S. A., et al. (2010), "New insights into the genetic control of gene expression using a Bayesian multi-tissue approach," *PLoS computational biology*, 6, e1000737.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007), "PLINK: a tool set for whole-genome association and population-based linkage analyses," *The American Journal of Human Genetics*, 81, 559–575.

Qiu, X. and Yakovlev, A. (2006), "Some comments on instability of false discovery rate estimation," *Journal of Bioinformatics and Computational Biology*, 4, 1057–1068.

Qiuheng, T., Jiang, H., and Yiming, D. (2014), "Model selection method based on maximal information coefficient of residuals," *Acta Mathematica Scientia*, 34, 579–592.

Rényi, A. (1959), "On measures of dependence," *Acta mathematica hungarica*, 10, 441–451.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. (2011), "Detecting novel associations in large data sets," *Science*, 334, 1518–1524.

Ripley, B. (1979), "Tests of 'randomness' for spatial point patterns," *Journal of the Royal Statistical Society. Series B (Methodological)*, 368–374.

Ripley, B. and Silverman, B. (1978), "Quick tests for spatial interaction," *Biometrika*, 65, 641–642.

Roeder, K. and Wasserman, L. (2009), "Genome-wide significance levels and weighted hypothesis testing," *Statistical science: a review journal of the Institute of Mathematical Statistics*, 24, 398.

Sarkar, S. K. (2002), "Some results on false discovery rate in stepwise multiple testing procedures," *Annals of statistics*, 239–257.

— (2008), "On methods controlling the false discovery rate," *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, 135–168.

Saviotti, P. (1996), *Technological evolution, variety, and the economy*, E. Elgar.

Schwartzman, A. (2008), "Empirical null and false discovery rate inference for exponential families," *The Annals of Applied Statistics*, 1332–1359.

Schwartzman, A. and Lin, X. (2011), "The effect of correlation in false discovery rate estimation," *Biometrika*, 98, 199–214.

Shabalin, A. A. (2012), "Matrix eQTL: ultra fast eQTL analysis via large matrix operations," *Bioinformatics*, 28, 1353–1358.

Šidák, Z. (1967), "Rectangular confidence regions for the means of multivariate normal distributions," *Journal of the American Statistical Association*, 62, 626–633.

Simon, N. and Tibshirani, R. (2014), "Comment on "Detecting Novel Associations In Large Data Sets" by Reshef Et Al, Science Dec 16, 2011," *arXiv preprint arXiv:1401.7645*.

Smith, T. E. (2004), "A Scale-Sensitive Test of Attraction and Repulsion Between Spatial Point Patterns," *Geographical analysis*, 36, 315–331.

Spearman, C. (1904), "The proof and measurement of association between two rings," *Amer. J. Psychol.*, 15, 72–101.

Speed, T. (2011), "A correlation for the 21st century," *Science*, 334, 1502–1503.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998), "Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," *Molecular biology of the cell*, 9, 3273–3297.

Storey, J. D. (2002), "A direct approach to false discovery rates," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 479–498.

— (2003), "The positive false discovery rate: A Bayesian interpretation and the q-value," *Annals of statistics*, 2013–2035.

Storey, J. D. and Tibshirani, R. (2003), "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences*, 100, 9440–9445.

Strimmer, K. (2008), "A unified approach to false discovery rate estimation," *BMC bioinformatics*, 9, 303.

Sun, L., Craiu, R. V., Paterson, A. D., and Bull, S. B. (2006), "Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies," *Genetic epidemiology*, 30, 519–530.

Sun, W. and Cai, T. T. (2007), "Oracle and adaptive compound decision rules for false discovery rate control," *Journal of the American Statistical Association*, 102, 901–912.

— (2009), "Large-scale multiple testing under dependence," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 393–424.

Székely, G. J. and Rizzo, M. L. (2009), "Brownian distance covariance," *The annals of applied statistics*, 3, 1236–1265.

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

Szekely, G. J., Rizzo, M. L., et al. (2014), "Partial distance correlation with methods for dissimilarities," *The Annals of Statistics*, 42, 2382–2412.

Torabi, H. and Vahidi-Asl, M. G. (2009), "Testing for "Randomness in Spatial Point Patterns, Using the Number of Empty-Quadrants in the Region," *Applied Mathematical Sciences*, 3, 1595–1608.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001), "Missing value estimation methods for DNA microarrays," *Bioinformatics*, 17, 520–525.

van der Laan, M. J., Dudoit, S., and Pollard, K. S. (2004), "Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives," *Statistical Applications in Genetics and Molecular Biology*, 3.

Wang, W., Wei, Z., and Sun, W. (2010), "Simultaneous set-wise testing under dependence, with applications to genome-wide association studies," *Stat. Interface*, 3, 501–511.

Westfall, P. and Young, S. (1993), *Resampling-based multiple testing: Examples and methods for p-value adjustment*, vol. 279, John Wiley & Sons.

Wright, F. A., Shabalin, A. A., and Rusyn, I. (2012), "Computational tools for discovery and interpretation of expression quantitative trait loci," *Pharmacogenomics*, 13, 343–352.

Yang, T. Y. and Jeong, S. (2013), "Grouped False-Discovery Rate for Removing the Gene-Set-Level Bias of RNA-seq," *Evolutionary bioinformatics online*, 9, 467.

Yekutieli, D. (2008), "Hierarchical false discovery rate–controlling methodology," *Journal of the American Statistical Association*, 103, 309–316.

Yekutieli, D. and Benjamini, Y. (1999), "Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics," *Journal of Statistical Planning and Inference*, 82, 171–196.

Zhao, H. and Zhang, J. (2014), "Weighted p-value procedures for controlling FDR of grouped hypotheses," *Journal of Statistical Planning and Inference*.

Zhao, Z. and Gene Hwang, J. (2012), "Empirical Bayes false coverage rate controlling confidence intervals," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74, 871–891.