Statistical Methods for Large-scale Multiple Testing Problems

by

Yu Gao

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2019

© Yu Gao 2019

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Dr. Shelley Bull Professor, University of Toronto
Supervisor(s):	Dr. Kun Liang Assistant Professor, University of Waterloo
Internal Member:	Dr. Paul Marriott Professor, University of Waterloo
	Dr. Richard J. Cook Professor, University of Waterloo
Internal-External Member:	Dr. Brendan J. McConkey Associate Professor, University of Waterloo

Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Chapters 2–4 are based on the research papers that are co-authored with my supervisor Dr. Kun Liang. My contributions include proposing the main methods, implementing the algorithms, performing the simulations, conducting the real data analyses, and writing the manuscripts. My supervisor gave me valuable comments and advice on these works.

Abstract

Large-scale multiple testing is an important research field in statistics, and it is widely applied in different areas, for example, DNA microarray analysis, brain imaging, and astronomical surveys. In a large-scale multiple testing problem, we often simultaneously test thousands or even millions of hypotheses, and many statistical challenges would emerge from a large number of hypotheses. In this thesis, we face up to some of the challenges, and focus on developing statistical methods for large-scale multiple testing problems with specific applications in pharmacovigilance databases and genome-wide association studies (GWAS).

In Chapter 2, we study the multiple testing problem in pharmacovigilance databases. Pharmacovigilance databases are established to monitor the adverse drug reactions of marketed drugs, and it is of interest to detect the combinations of drugs and adverse events that exhibit stronger associations than some threshold. The null hypotheses that are tested are composite, and the distribution of test statistics under null hypotheses may be hard to derive. Moreover, the count of reports for combinations is discrete, and the test statistics are also discrete. We first derive the optimal test statistics to maximize the power of detection while controlling the false discovery rate (FDR). We then propose a nonparametric empirical Bayes method to estimate the test statistics and demonstrate its performance advantage through simulations. We apply the proposed method to the pharmacovigilance database in the United Kingdom, and detect additional signals.

In Chapters 3–4, we study the applications of multiple testing in GWAS. GWAS are widely used to identify the genetic variants that are associated with human diseases or traits. We are often trying to identify the associated genetic variants among millions of genotyped ones, while there are only a few thousand subjects included in the sample due to economic reasons. The common practice of GWAS is to compare the marginal p-values to an overly-stringent threshold to control the family-wise error rate, and the procedure is known to be lacking in power given fixed sample size. Moreover, the neighboring genetic variants are often highly correlated with each other, and the local dependence further implicates the multiple testing problem. We are interested in improving the detection power by proposing new statistical methods that employ the FDR that is a more powerful error rate.

Specifically, in Chapter 3, we focus on genetic studies on continuous traits. We propose a novel approach to take into consideration the effects of local dependence among genetic variants. We propose a search algorithm to find tentative causal variants, and compute adjusted p-values by accounting for the effects of tentative causal variants. Then, the

adjusted p-values of non-causal variants would be uniformly distributed, and conventional FDR control procedures are applicable. Through simulations and application to the North Finland Birth Cohort (NFBC) study, we show that our procedure is advantageous over other candidate methods.

In Chapter 4, we focus on the case-control genetic studies with shared control. A common phenomenon "pleiotropy" exists as some genetically related diseases often share associated variants, and we can leverage the pleiotropy to improve the probability of identifying weakly associated variants by integratively analyzing related diseases. However, the related GWAS often share part of the control sample, and the shared control sample would induce a positive correlation between test statistics. We propose a four-component bivariate normal mixture model for the z-values from two GWAS, use an expectation-maximization (EM) algorithm to estimate the parameters, and further estimate the FDR. An adaptive pruning procedure is proposed to tackle the problem of local dependence. Through simulations and application to the data of schizophrenia and bipolar disorder, we show the proposed procedure outperforms other methods.

Acknowledgements

Foremost, I would love to express my sincere gratitude to my supervisor Dr. Kun Liang for his continuous support and patient guidance throughout my Ph.D study, for his enthusiasm, and for his immense knowledge. I really appreciate the remarkable ideas, the valuable comments and the efforts he took to guide me through the whole study.

My gratitude also extends to other examining committee members, Dr. Shelley Bull, Dr. Paul Marriot, Dr. Richard J. Cook, and Dr. Brendan J. McConkey. I appreciate their precious time and efforts to make insightful comments on the thesis.

I also want to thank all the friends I met at the University of Waterloo.

Last but not least, I want to express my thanks to my wife Jiawen Li, and our parents Wenfang Zhang, Wenzuo Gao, Suijun Li, and Jian Li for their constant support and care. Moreover, I would also take the opportunity to thank my daughter Everly Gao for bringing joy to the whole family.

Dedication

This is dedicated to my family.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this chapter, we provide general background information about the topics studied in the dissertation. We first introduce the hypothesis testing problem, from simple hypothesis testing to multiple hypothesis testing. We also describe the procedures to control two widely used error rates: the family-wise error rate and the false discovery rate. We are interested in the application of multiple testing in a large-scale setting where thousands or even millions of hypotheses are simultaneously tested. We focus on two specific multiple testing applications that motivate our methodology development, and describe the context of pharmacovigilance and genetic studies. In the first application in pharmacology, we aim to detect adverse events of marketed drugs from spontaneous reports and alert pharmacovigilance experts to unexpected strong associations of drug and adverse event. In the second application in genetics, we aim to identify the genetic variants that are associated with complex human diseases or traits via genome-wide association studies (GWAS).

1.1 Hypothesis Testing

1.1.1 Simple Hypothesis Testing

Hypothesis testing is an important method for statistical inference. The whole population is often not observable because it is either too large or not accessible, and the true value of a population parameter is usually not known. We can use hypothesis testing and estimation to make inference on the parameter in the population with a set of observed data sampled from the population. Hypothesis testing is different from estimation, as estimation is trying

to derive the most likely value of the population parameter, while hypothesis testing is determining whether a statement/hypothesis about the parameter value is likely to be true or not.

Hypothesis testing often tries to determine between two competing hypotheses about the population parameters using statistical evidence. For example, one hypothesis may state that male and female students at the University of Waterloo study for the same length of time in a semester, while the other hypothesis claim that female students study longer than male students. The first hypothesis is actually tested, and is referred to as the null hypothesis, denoted by H_0. The null hypothesis is assumed to be true unless there is strong evidence against it. The other hypothesis is known as the alternative hypothesis, and is true when the null hypothesis is false, often denoted by H_1 or H_A. The null and alternative hypotheses need to be disjoint, and sometimes complementary of each other.

If a null hypothesis specifies a single value for the population parameter of interest, then it is called a simple null hypothesis. On the contrary, a composite null hypothesis specifies the parameter as a range of values instead. For example, in Chapter 2 we are testing whether the odds ratio of a combination of drug and adverse event is no greater than a tested value, and we are testing a composite null hypothesis; while in the genetic studies in Chapters 3 and 4, we are testing whether a genetic marker is associated with some trait/disease or not, i.e., the coefficient in a model is 0 or not, and it is a simple null hypothesis.

Hypothesis testing aims to determine whether the likelihood that the value of the population parameter is likely to be true or not. There are four steps in the hypothesis testing (Privitera, 2011), and they are summarized as follows,

1. State the null and alternative hypotheses. The null hypothesis states a value/range for the population parameter to be tested, while the alternative hypothesis states a range that directly contradicts the null hypothesis.

2. Set the level of significance. It is a criterion we set for making a decision for hypothesis testing. We can compute the likelihood of observing the sample if the null hypothesis were assumed to be true, and compare the likelihood to the level of significance. The level of significance is often set at 5% or 10%.

3. Calculate the test statistic. When computing the likelihood, we need a random variable to evaluate how likely it is to observe the sample if the null hypothesis were assumed to be true, and the random variable is called the test statistic.

4. Draw a conclusion. The value of test statistic can be used to draw a conclusion about whether to reject the null hypothesis. We can compute the probability of having the observed test statistic given that the value of the population parameter in the null hypothesis were assumed to be true, and this probability is called p-value. If the p-value is no greater than the level of significance, we reject the null hypothesis and the statement of the parameter value in the null hypothesis is not true; otherwise, there is no evidence to reject the null hypothesis.

For a null hypothesis H_0, depending on whether the status of the null hypothesis is true or false and the decision is to reject the null hypothesis or not, there are four possible categories. Among the four categories, there are two erroneous cases. The statistical test may reject the null hypothesis when H_0 is true, and this type of error is called *type I error*. The associated risk is the *type I error rate*, denoted by α. The statistical test may not reject the null hypothesis when H_0 is false, and this is a *type II error*. Its associated risk is the *type II error rate*, denoted by β. The outcomes are summarized in Table 1.1.

Table 1.1: Outcomes of testing a null hypothesis

	Not to reject H_0	Reject H_0
H_0 is true	true negative; $(1-\alpha)$	false positive; type I error rate (α)
H_0 is false	false negative; type II error rate (β)	true positive; power $(1-\beta)$

In a hypothesis testing problem, we often try to control the type I error rate at a nominal level and maximize the power at the same time.

1.1.2 Multiple Hypothesis Testing

Now we move on to the multiple testing problem where multiple null hypotheses are tested simultaneously. The progress in science often accelerates after breakthroughs in technology. As statisticians, we need to develop new statistical methods for every new wave of scientific data. Simultaneous hypothesis testing first drew attention from mathematicians and statisticians in the 1950s, and methods for multiple comparisons in the setting of analysis of variance (ANOVA) have been developed. Such methods for multiple comparisons adjustment focus on correcting for a modest number of comparisons, motivated in the ANOVA setting. Over the past several decades, new technologies such as microarray and high-throughput sequencing bring the multiple testing challenge into a new era, in which a large number of hypotheses, on the scale of thousands or even millions, are simultaneously tested. Multiple hypothesis testing is now commonplace in many scientific fields, such as

pharmacology, genetics, and astronomy. For example, in genetics it is of interest to discover among thousands of genes the ones that are differentially expressed between the cases and controls.

Let us consider a multiple testing problem, where m null hypotheses are simultaneously tested, among which m_0 are true null hypotheses,

$$H_{01},\ H_{02},\ \ldots,\ H_{0m}, \tag{1.1}$$

and suppose the corresponding p-values are $p_1,\ p_2,\ \ldots,\ p_m$. A test procedure will declare each hypothesis "significant" or "non-significant", or equivalently, will reject or accept each hypothesis. For a hypothesis in the m hypotheses, depending on its true status of the null hypothesis and whether the test procedure rejects it or not, the test result would fall in one of the four categories in Table 1.1. Table 1.2 summarizes the outcomes of simultaneously testing m hypotheses by the number of hypotheses that fall in each of the four categories. Here, we consider the number of type I errors (false positives; V) instead of the type-I error rate (α), and the number of type II errors (false negatives; T) instead of the type-II error rate (β).

Table 1.2: Outcomes of testing m null hypotheses

	Declared non-significant	Declared significant	Total
True null	true negatives (U)	false positives (V)	m_0
False null	false negatives (T)	true positives (S)	$m - m_0$
Total	$m - R$	R	m

The probability of making a type I error would increase exponentially with the number of null hypotheses that are tested, and the naive use of the type I error rate for each hypothesis may not be suitable in multiple testing. Consider an illustrating example, we simultaneously test $m = 1,000,000$ hypotheses, on the same scale as the studies in Chapters 2–4, and we apply the type I error rate $\alpha = 0.05$ to each test. We would expect $500,000$ type I errors simply by chance, and this may not be acceptable by most researchers. Therefore, the type I error rate for each hypothesis is not suitable for multiple testing.

Hochberg and Tamhane (1987) suggest that it is meaningful to consider a combined measure of error rate for a family of hypotheses. It is necessary to consider alternative error rates instead of the type I error rate in the multiple testing problem. Here we describe the family-wise error rate and the false discovery rate in details, which are the two most widely used error rates for multiple testing applications.

Family-wise Error Rate

The first alternative error rate is the family-wise error rate (FWER), also known as the overall Type I error rate. It is defined as the probability of committing one or more false positives in a family of hypotheses, i.e.,

$$\text{FWER} = \Pr\{V > 0\}.$$

Controlling the family-wise error rate at a nominal level α implies that the probability of having at least one false positive is at most α. A few procedures (for instance, see Bonferroni 1936; Šidák 1967 and Hochberg 1988) have been proposed to control the family-wise error rate. The classical Bonferroni's procedure is most widely used, and it rejects the null hypothesis H_{0i} if

$$p_i \leq \alpha/m.$$

For example, there are often millions of genetic markers tested simultaneously in genome-wide association studies (GWAS), and the common practice is to compare the p-values of genetic markers to the Bonferroni-corrected threshold $0.05/1,000,000 = 5 \times 10^{-8}$ to control the family-wise error rate at 0.05.

The Bonferroni's procedure applies a more stringent threshold that takes the number of hypotheses into account to the p-values. It is straightforward to apply and can control the family wise error rate for any dependence structure among the hypotheses. However, the Bonferroni's procedure applies an overly-stringent threshold to the p-values, and may miss some true signals while controlling the probability of at least one false positive. The Bonferroni's procedure has limited detection power, especially when the number of hypotheses tested is large.

False Discovery Rate

In their seminal paper, Benjamini and Hochberg (1995) introduce the false discovery rate (FDR) as a powerful alternative to the family-wise error rate. From then on, the FDR has gained an increasing amount of popularity from researchers in various fields, and has even become the standard methodology in some areas, such as the expression quantitative trait loci (eQTL) community. The concept of family-wise error rate was introduced when the multiple testing problem often contains a modest number of individual cases m, while the FDR is more suitable to the new era of multiple testing applications with m in thousands (Efron, 2010).

The proportion of false positives V among all the rejections R can be viewed as a random variable, and the FDR is defined as

$$\text{FDR} = E\left(\frac{V}{R \vee 1}\right),$$

where $R \vee 1 = max(R, 1)$. By definition, the FDR equals to 0 when no rejection is made.

There are two main approaches in the control of the FDR, frequentist and Bayes. The most well-known frequentist method is the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995). Let us order all the p-values from the smallest to the largest, denoted by $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(i)} \leq \ldots \leq p_{(m)}$. For a nominal level α in $(0, 1)$, the BH procedure find the index k such that $k = max\{i : p_{(i)} \leq i\alpha/m\}$ and reject all the hypotheses whose p-value are no larger than $p_{(k)}$. The BH procedure applied at level α controls the FDR at level less than or equal to $m_0/m\alpha$. The equality holds when testing continuous data, as the p-values are uniformly distributed under the null hypotheses and $\Pr(P_i \leq k/m\alpha) = k/m\alpha$ for $i = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, m$. The BH procedure controls the FDR under independence and a special positive dependence condition (Benjamini and Hochberg 1995; Benjamini et al. 2001).

A related concept, the marginal FDR (mFDR) is defined as

$$\text{mFDR} = \frac{E(V)}{E(R)},$$

and Genovese and Wasserman (2002) show that mFDR = FDR + $O(m^{-1/2})$ when test statistics are independent and the proportion of null hypotheses remains constant. Therefore, the mFDR is asymptotically equivalent to the FDR.

Alternatively, we can also consider a two-group mixture model in a Bayesian framework of the FDR. Corresponding to $H_{01}, H_{02}, \ldots, H_{0m}$, assume we have the test statistics $X_1, X_2, \ldots, X_m$. $\theta_1, \theta_2, \ldots, \theta_m$ are the null indicators, where $\theta_i = 1$ indicates H_{0i} is true null, and θ_i's independently follow Bernoulli(π_0) distribution, where π_0 is the true null probability. Let F denote the cumulative distribution function (CDF) of X, and we have the following two-group model,

$$F(x) = \pi_0 F_0(x) + (1 - \pi_0) F_1(x),$$

where π_0 is the proportion of null hypotheses, and F_0 and F_1 are the CDFs of X under the null and non-null hypotheses.

For a rejection region S, Efron and Tibshirani (2002) define the Bayesian FDR (Fdr) as

$$\begin{aligned}\text{Fdr}(S) &= \Pr(\theta = 1 | x \in S) \\ &= \frac{\pi_0 F_0(S)}{\pi_0 F_0(S) + (1-\pi_0) F_1(S)} \\ &= \frac{\pi_0 F_0(S)}{F(S)}.\end{aligned}$$

Note that the Bayesian FDR is the posterior probability that a hypothesis is true null given it is in the rejection region. Under the two-group model, it is straightforward to show mFDR(S) = Fdr(S). Hence, the Bayesian FDR is equivalent to the marginal FDR under the two-group model, and we can approximate the FDR by the marginal FDR or the Bayesian FDR asymptotically.

The local FDR (fdr) was proposed by Efron et al. (2001) and is defined as the posterior probability of a hypothesis with a test statistic x being a true null,

$$\text{fdr}(x) = \frac{\pi_0 f_0(x)}{f(x)},$$

where $f_0(x)$ and $f(x)$ are the probability density functions of X under the null hypothesis and the overall density of X, respectively. Efron and Tibshirani (2002) show that the Bayesian FDR is the conditional expectation of the local FDR, i.e., $\text{Fdr}(S) = E\{\text{fdr}(x) | x \in S\}$. This result suggests that for any rejection region S, its Bayesian FDR can be empirically estimated as $\sum_{x_i \in S} \text{fdr}(x_i)/|S|$, where $|S|$ is the number of test statistics within S.

The estimate of the local FDR is then an important step for the FDR control. For a rejection region S, the proportion of null hypotheses π_0 can be estimated as

$$\hat{\pi}_0 = \frac{|S|}{m F_0(S)},$$

where $f_1(x) = 0$, for $x \in S$, i.e., we here have the "zero assumption" that all the test statistics of the non-null hypotheses are outside of the region S (Efron, 2010). For example, this set can be $[0.5, 1]$ if we are working on p-values. The null density function $f_0(x)$ is typically assumed to be known as we know the exact distribution of the test statistic X under null, for example, it is $Uniform[0, 1]$ for p-values and the standard normal distribution for z-values. Efron (2004a) also introduce an empirical estimation of the distribution under the null hypothesis. The mixture density distribution $f(x)$ can be estimated by simple kernel density estimates or Poisson regression estimates for z-values (Efron, 2010).

With the estimates of the local FDR, the following FDR control procedure at level α naturally emerges,

1. Order all the hypotheses by the fdr estimates from the smallest to the largest, denoted by $H_{0(i)}$ corresponding to $\widehat{\text{fdr}}_{(i)}$;
2. Reject the hypotheses $H_{0(i)}$ for $i = 1, \ldots, k$, where $k = max\{j = 1, 2, \ldots, m : \frac{\sum_{i=1}^{j} \widehat{\text{fdr}}_{(i)}}{j} \leq \alpha\}$.

1.2 Pharmacovigilance

The word "pharmacovigilance" originates from two roots: pharmakon that means "drug" in Greek and vigilare that means "to keep watch" in Latin. Pharmacovigilance is a branch in pharmacological science focusing on drug safety via collecting, monitoring, detecting and further preventing adverse effects of drugs on the market, also known as adverse drug reactions (ADRs). Pharmacovigilance is crucial to assuring the safety of drugs by reliable and timely exchange of information on drug safety issues (WHO et al., 2002).

Many pharmacovigilance databases have been established to collect reports of adverse effects of marketed drugs, for example, the FDA's Adverse Event Reporting System (FAERS; Gupta 2008) in the United States, the European Medicines Agency EudraVigilance Database (Vermeer et al., 2013) and the VigiBase (Lindquist, 2008) from the World Health Organization.

1.2.1 Medical Dictionary for Regulatory Activities

The Medical Dictionary for Regulatory Activities (MedDRA) was developed by the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) in the 1990s. The MedDRA is a highly specific standardized medical terminology dictionary for the convenient sharing of information for medical products, and it is used by regulatory authorities in the pharmaceutical industry, from pre-marketing clinical trials to post-marketing pharmacovigilance monitoring. It is the standard adverse event classification dictionary that is endorsed by the ICH.

The MedDRA dictionary has a hierarchical structure. This structure is organized by System Organ Class (SOC), further divided in three levels, the "high level group terms (HLGT), the "high level terms" (HLT), the "preferred terms" (PT), and the "lowest level

terms" (LLT), from the most general to the most specific. The data entry of individual cases are coded at the LLT level, and the output of counts are often at the PT level.

1.2.2 Collection of reports

There are mainly two types of reports of drugs and adverse events. The first one is a solicited report. A solicited report is gathering information about suspected ADRs for specific drugs and adverse events, and it is derived from organized data collection systems, for example, clinical trials, post-approval named patient use programs, other patient support and disease management programs, surveys of patients or healthcare providers, and information gathering on efficacy or patient compliance (Srba, 2014). Adverse event reports from a solicited report cannot be considered spontaneous. The second one is an unsolicited report. Unsolicited reports are obtained from other sources, for example, spontaneous reports, scientific literature reports, and media reports. See more details about different kinds of ADR reports in European Medicines Agency (2012).

Spontaneous reports

We give a detailed description of spontaneous reports as they are used in pharmacovigilance databases. A spontaneous report is a kind of unsolicited communication from healthcare professionals or patients to a regulatory authority or other organizations (for example, the WHO, Poison Control Center), and it describes the ADRs from a patient who was given one or more medicinal products and that does not derive from a study or any organized data collection scheme (European Medicines Agency, 2012). Stimulated reporting should be considered spontaneous, such as a publication in the press, or questioning of healthcare professionals by company representatives.

The reports that are directly from patients should also be handled as spontaneous reports irrespective of any subsequent "medical confirmation", a process required by some regulatory authorities. As European Medicines Agency (2012) increases the importance of patients in the existing context of spontaneous reporting ADRs, even if the reports received from patients do not qualify for regulatory reporting, the cases should still be retained, and emphasis should be placed on the report instead of on its source.

1.2.3 Pharmacovigilance databases

Pharmacovigilance databases often contain millions of spontaneous reports of adverse events and associated drugs. Typically, a health care provider can submit a spontaneous report through a spontaneous reporting scheme when he/she believes one of his/her patients is suffering adverse events related to a drug being taken, or a patient can submit a spontaneous report on his/her own. In many countries, spontaneous reports can be filed electronically under some specific standard.

Drug Analysis Print
Drug name: ABATACEPT

Drug name:	ABATACEPT	Report type:	Spontaneous
Report run date:	24-Jul-2014	Report origin:	UNITED KINGDOM
Data lock date:	24-Jul-2014 19:00:05	Route of admin:	ALL
Period covered:	01-Jul-1963 to 24-Jul-2014	Reporter type:	ALL
Earliest reaction date:	26-Sep-2007	Reaction:	ALL
MedDRA version:	MedDRA 17.0	Age group:	ALL

	Single active constituent		Multiple active constituent		Total unique reports*	
Reaction Name	All	Fatal	All	Fatal	All	Fatal
SOC						
HLT						
PT						
Eye disorders						
Iris and uveal tract infections, irritations and inflammations						
Uveitis	1	0	0	0	1	0
Lid, lash and lacrimal infections, irritations and inflammations						
Eyelid oedema	1	0	0	0	1	0
Ocular disorders NEC						
Eye pain	1	0	0	0	1	0
Eye disorders SOC TOTAL	3	0	0	0	3	0

Figure 1.1: Example of aggregated spontaneous reports from Yellow Card Scheme

In the United Kingdom, the Medication and Healthcare Products Regulatory Agency (MHRA) operates post-marketing surveillance for reporting, investigating and monitoring of adverse drug reactions. The spontaneous reporting scheme in the United Kingdom was introduced by the launch of the Yellow Card Scheme, after the thalidomide tragedy 1964. The doctors and dentists (since 1964), pharmacists (since the 1990s), nurses (since 2002), and patients (since 2004) have been invited to submit suspected ADRs. Now, hundreds of thousands of spontaneous reports have been submitted to the MHRA via the Yellow Card Scheme (http://yellowcard.mhra.gov.uk/).

The databases then aggregate individual spontaneous reports based on the combination of drugs and events. Figure 1.1 shows an example of aggregated spontaneous reports of an autoimmune disease drug Abatacept and adverse events reported related to eye disorders.

We can see that related adverse drug reactions are grouped in a hierarchical structure by MedDRA for Medicines Regulation. The structure used in the United Kingdom pharmacovigilance database is composed of three levels, the "preferred term" (PT), the "high level term" (HLT) and the "system organ class" (SOC), from the most specific to the most general. For example, eye pain (PT) is grouped under the broader heading ocular disorders NEC (HLT), which is contained within eye disorders (SOC). We focus on the adverse drug reactions of the high level term in the application of our project.

We can extract the number of reported cases for each combination of adverse event and drug from a pharmacovigilance database, and the pharmacovigilance data can be summarized as a large $I \times J$ contingency table crossing I adverse events and J drugs. For a specific pair (adverse effect i, drug j), a 2 by 2 contingency table can be obtained by collapsing the large contingency table, as presented in Table 1.3. In the table, n_{ij}, $n_{i.}$, $n_{.j}$ and n denote the number of reports involving the pair (i, j), the marginal count for the adverse event i, the marginal count for the drug j and the total count in the contingency table, respectively. Our aim is, among the combinations of drugs and adverse events, to detect the pairs of (i, j) where drug j leads to significantly more adverse events i than expected by chance.

Table 1.3: Two by two contingency table for the adverse event-drug combination (i, j)

	Drug j	Other drugs	Total
Adverse event i	n_{ij}	$n_{i.} - n_{ij}$	$n_{i.}$
Other adverse events	$n_{.j} - n_{ij}$	$n + n_{ij} - n_{i.} - n_{.j}$	$n - n_{i.}$
Total	$n_{.j}$	$n - n_{.j}$	n

1.2.4 Automatic Detection Methods

Due to a large number of possible combinations of drugs and adverse events, automatic detection methods need to be developed. The automatic detection methods identify some combinations of drugs and adverse events that show unusually strong association and report these detections to pharmacovigilance experts so that the reported associations can be further analyzed. We here introduce some automatic detection methods that are commonly used by regulatory agencies and drug monitoring systems.

Proportional Reporting Ratio

The proportional reporting ratio (PRR) is proposed by Evans et al. (2001), and it is defined as the ratio of the proportion with which a specific adverse event is reported for the drug of interest and the proportion with which the same adverse event is reported for all the others drugs, i.e., for drug j and adverse event i,

$$\text{PRR}_{ij} = \frac{n_{ij}(n - n_{.j})}{n_{.j}(n_{i.} - n_{ij})}.$$

A proportional reporting ratio indicates the extent to which a specific adverse event is reported for individuals that are taking or not taking a specific drug. If the proportional reporting ratio is greater than 1, it suggests that the adverse event is more commonly reported for individuals taking the drug.

A signal is generated if there is a PRR no smaller than 2, chi-squared test (on one degree of freedom with Yates's correction) statistic no smaller than 4 and three or more cases.

Reporting Odds Ratio

The reporting odds ratio (ROR) is introduced by van Puijenbroek et al. (2002), and it is defined as the observed odds ratio, for drug j and adverse event i,

$$\text{ROR}_{ij} = \frac{n_{ij}(n - n_{.j} - n_{i.} + n_{ij})}{(n_{i.} - n_{ij})(n_{.j} - n_{ij})}.$$

The logarithm of reporting odds ratio is assumed to be normally distributed with variance estimated as $var\{ln(ROR_{ij}\} = 1/n_{ij} + 1/(n - n_{.j} - n_{i.} + n_{ij}) + 1/(n_{i.} - n_{ij}) + 1/(n_{.j} - n_{ij})$. Then, if the lower bound of the two-sided 95% confidence interval is greater than 0, a signal is generated.

Gamma Poisson Shrinker

The Gamma Poisson shrinker model is proposed by DuMouchel (1999). It is assumed that for a cell (i, j), $n_{ij} \sim Poisson(\phi_{ij}E_{ij}))$, where ϕ_{ij} is the relative risk, and E_{ij} is the expected number of reports for cell (i, j) if the independence between drug j and adverse event i were true,

$$E_{ij} = \frac{n_{i.}n_{.j}}{n}.$$

The prior distribution of all the ϕ's is a mixture of two gamma distributions,

$$\phi_{ij} \sim \omega Gamma(\alpha_1, \beta_1) + (1-\omega) Gamma(\alpha_2, \beta_2),$$

and the parameters are estimated by maximizing the products of the marginal likelihoods of $n_{ij}|E_{ij}$. Then, the posterior distribution is computed as

$$\begin{aligned}&\phi_{ij}|n_{ij}, E_{ij} \\ &\sim \hat{\omega} Gamma(\hat{\alpha}_1 + n_{ij}, \hat{\beta}_1 + E_{ij}) + (1-\hat{\omega}) Gamma(\hat{\alpha}_2 + n_{ij}, \hat{\beta}_2 + E_{ij}).\end{aligned}$$

DuMouchel and Pregibon (2001) propose to rank the cells by the 5% quantile of the distribution of $\phi_{ij}|n_{ij}, E_{ij}$ and a threshold of 2 is recommended by Szarfman et al. (2002).

1.2.5 Related Research in Multiple Testing

The automatic signal detection methods mentioned above use empirical thresholds to detect signals, and do not consider the multiplicity adjustment due to a large number of hypotheses tested simultaneously. We consider the problem of detecting adverse drug effects in the setting of multiple testing, and focus on estimating and controlling the FDR (Benjamini and Hochberg, 1995). There are two main challenges here. The first one is that the number of reports of adverse event and drug is discrete, and the test statistics testing their association is also discrete; the second one is that we are interested in detecting strong associations between adverse events and drugs, which lead to testing of composite null hypotheses.

Most existing methods can only address one of the two challenges, except Ahmed et al. (2009) and Ahmed et al. (2010), and we briefly describe the statistical methods that have been proposed to test composite null hypotheses with continuous data, to test simple null hypotheses with discrete data, and to test composite null hypotheses with discrete data.

Composite Null Hypotheses with Continuous Data

Sun and McLain (2012) consider the multiple testing problem of composite null hypotheses in heteroscedastic models. Specifically, Suppose X_i, $i = 1, 2, \ldots, m$ are independent observations from a random mixture model,

$$X_i = \mu_i + \epsilon_i, \ \epsilon_i \sim N(0, \sigma_i^2),$$

where μ_i denotes the unknown effects sizes for comparing two competing conditions, and ϵ_i denotes the heteroscedastic errors with variances σ_i^2.

They assume that μ_i has a mixture density function with a point mass at zero,

$$f_\mu(\cdot) = (1-\omega)\delta_0(\cdot) + \omega g(\cdot),$$

where $\omega = \Pr(\mu_i \neq 0)$ is the non-zero proportion, δ_0 is the dirac delta function, and g is a continuous density function of non-zero effects.

The composite null hypotheses $H_{i0} : \mu_i \in A$ are tested simultaneously, and A is a set of unimportant effects including 0 and small uninteresting effects, referred to as "indifference region" by Sun and McLain (2012).

Sun and McLain (2012) formulate the problem in a decision theory framework. They formulate the density function of X as a mixture model on A and A^c, i.e.,

$$f(x|\sigma^2) = (1-\tilde{\omega})\tilde{f}_0(x|\sigma^2) + \tilde{\omega}\tilde{f}_1(x|\sigma^2),$$

where

$$\tilde{\omega} = \omega \int_{A^c} g(\mu)d\mu,$$

$\tilde{f}_0$ and $\tilde{f}_1$ are the corresponding density functions on A and A^c, respectively.

Then, they prove the optimal test statistic to minimize the loss as an expression of the weighted sum of false positives and false negatives is,

$$T_{OR}(X_i, \sigma_i^2) = \frac{(1-\tilde{\omega})\tilde{f}_0(X_i, \sigma_i^2)}{f(X_i, \sigma_i^2)},$$

and it actually reduces to the local FDR when $A = \{0\}$ and the errors are homoscedastic (Sun and McLain, 2012). See details about the parameter estimates in Sun and McLain (2012).

Sun and McLain (2012) test the composite null hypotheses with continuous test statistics, and their method cannot be applied to the multiple testing problem with composite null hypotheses and discrete data. However, their formulation of the local FDR inspires us to derive the local FDR as our test statistics in Chapter 2.

Discrete Data with Simple Null Hypotheses

Many methods have been proposed to test multiple simple null hypotheses with discrete test statistics. For discrete data, the BH procedure applied at level α controls the FDR

at level less than $m_0/m\alpha$, and Gilbert (2005) argues that the original BH procedure can be made more powerful. Gilbert (2005) proposes a modification to the BH procedure that consists of two steps. The first step is to remove the null hypotheses whose test statistics do not reach some level of significance, and the second step is to apply the BH procedure to the remaining hypotheses. One drawback of Gilbert (2005) is that they ignore the discreteness of the remaining hypotheses.

Kulinskaya and Lewin (2009) suggests using randomized p-values from randomized tests. Specifically, for the null distribution of a discrete test statistic X, let c be the value of X such that $\Pr(X \geq c) > \alpha$ and $\Pr(X > c) < \alpha$. Then, a randomized p-value that achieves the exact level α test is computed as $P(c) = \Pr(X > c) + U \Pr(X = c)$, where $U \sim Uniform(0, 1)$. The randomized p-values are continuous and uniformly distributed under the null hypotheses. However, the randomized p-values are not interpretation-friendly due to the randomness.

Heyse (2011) exploits the discreteness of the test statistics and propose to use FDR adjusted p-values that are modified for discreteness for the FDR control. Define $Q_i(P)$ as the largest p-value no greater than P that is achievable for hypothesis $i = 1, 2, \ldots, m$. $Q_i(P)$ is zero if the p-values are not achievable under P or an extreme value of P. They propose to compute adjusted p-values as

$$P^*_{|m|} = P_{(m)},$$

$$P^*_{|j|} = min\{P_{|j+1|}, \sum_{i=1}^{m} Q_i(P_{(j)})/j\},$$

for values of $j \leq m - 1$. Then, the hypotheses with adjusted p-values $P^*_{|j|} \leq \alpha$, $j = 1, 2, \ldots, m$ are rejected. Heyse (2011) demonstrates its power advantage over Gilbert (2005) by simulations, however it may fail to control the FDR, see examples in Gur (2011) and Döhler et al. (2018).

Composite Null Hypotheses with Discrete Data

Let ψ denote the odds ratio, which is a measure of the association between adverse event and drug. We test the following null hypotheses for adverse event i and drug j,

$$H_{0_{ij}} : \psi_{ij} \leq \psi_0 \text{ versus } H_{1_{ij}} : \psi_{ij} > \psi_0,$$

where ψ_0 is a tested value for odds ratio, and $\psi_0 = 1$ is often of interest.

Ahmed et al. (2009) extend the two existing Bayesian methods used in automatic signal detection for pharmacovigilance, the Bayesian confidence propagation neural network (BCPNN; Bate et al. 1998) and the Gamma Poisson Shrinker (GPS; DuMouchel 1999), to the multiple testing setting. By simulations, Ahmed et al. (2009) show that the extended GPS method has better performance than the extended BCPPN method and the original GPS and BCPPN. However, the GPS assumes a two-component gamma distribution for relative risks, and the parametric assumption can be too strong in reality.

Ahmed et al. (2010) propose a frequentist method that models the distribution of mid-p-values (Lancaster, 1961) as a mixture model after filtering out the hypotheses with small cell counts. In Table 1.3, assume the marginal counts $(n_{i\cdot}, n_{\cdot j}, n)$ are fixed, and the random variable $N_{ij}|n_{i\cdot}, n_{\cdot j}, n; \psi_0$ follows an non-central hypergeometric distribution (Agresti and Kateri, 2011). The p-value for cell (i, j) can be computed as $p_{ij} = \Pr(N_{ij} \geq n_{ij}|n_{i\cdot}, n_{\cdot j}, n; \psi_0)$. The corresponding mid-$p$-value is computed as $\Pr(N_{ij} > n_{ij}|n_{i\cdot}, n_{\cdot j}, n; \psi_0) + \frac{1}{2}\Pr(N_{ij} = n_{ij}|n_{i\cdot}, n_{\cdot j}, n; \psi_0)$. The distribution of p-values under null hypothesis is assumed to be a non-decreasing function expressed as a mixture of a uniform distribution and a non-decreasing function (Ahmed et al., 2010). Through a simulation study, Ahmed et al. (2010) show their method is able to control the FDR. However, filtering the hypotheses with small cell counts is not always satisfactory as it naively discards some hypotheses. Also, it is non-trivial to select the threshold for the cell count filtering.

1.3 Genome-wide Association Studies

In the last two decades, genome-wide association studies (GWAS) have become an important tool to identify the genetic variants that are associated with human diseases or traits, as can be seen from Figure 1.2. As of September 2016, GWAS have found more than $24,000$ unique variant-trait associations from $2,518$ publications from the GWAS catalog `http://www.genome.gov/gwastudies/` (Welter et al., 2013).

1.3.1 Basic Concepts

All living creatures, including human and animals, are composed of functional and reproductive cells. Inside the cells reside the hereditary material, the genome. The genome genetically influences the process of development of cells and further the characteristics of the living creatures, and the characteristics are called phenotype. The phenotypes can be traits, for example human height and blood pressure, or diseases, for example type I diabetes and autism.

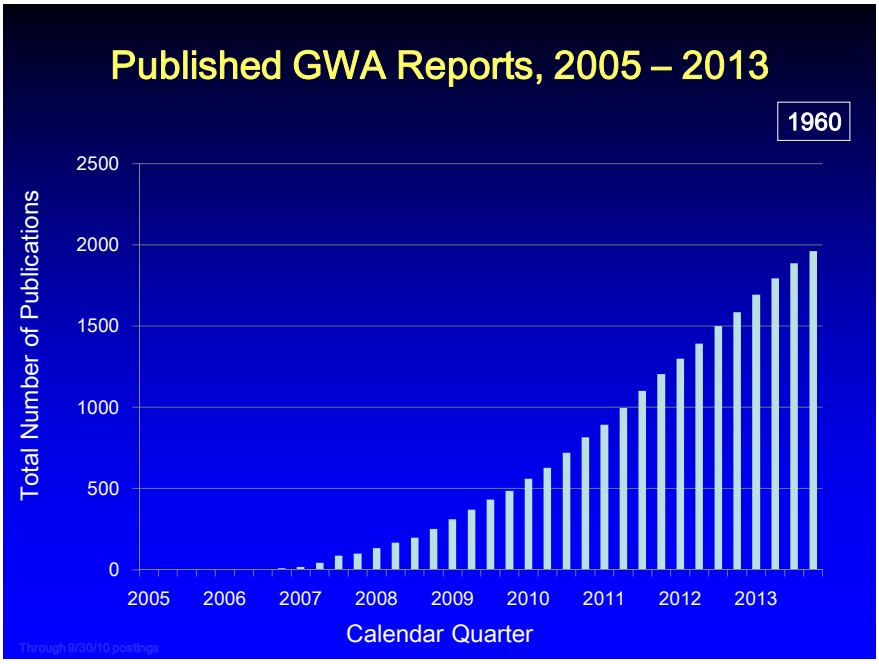

Figure 1.2: Number of published GWA reports, figure from the GWAS catalogue

The genome is made of chromosomes. Different living creatures have a varying number of chromosomes. For example, human have 23 pairs of chromosomes while monkeys and Ape have 24 pairs. Among the 23 pairs of chromosomes in human, 22 pairs are autosomes and they look the same in males and females, while the 23rd pair differs between male and female and are called the sex chromosomes. When a baby is born, for each pair of chromosomes he inherits one chromosome from his father, and the other from his mother.

Each chromosome contains a string of 4 nucleotides: Adenine (A), Thymine (T), Cytosine (C) and Guanine (G), and the nucleotides arrange in pairs to form Deoxyribonucleic Acid (DNA) that has a double helix structure. A segment of DNA codes for the amino acid sequence for a particular protein, and this segment is called a gene.

The most common genetic variants spanning the whole genome are single-nucleotide variants. A single-nucleotide polymorphism (SNP) is a variation of a single nucleotide at a specific location in the genome. At a SNP location, there are typically two alleles

(base-pairs) that commonly exist within a population: a major allele and a minor allele. The definition of major or minor is based on the frequency of that allele appears, and the frequency of a SNP is defined as the frequency of the minor allele. For example, at a specific SNP position, the two possible nucleotide variations are C and A and they are called the alleles for this position. If the C nucleotide only appears in a small fraction of the individuals in the population, say 20%, and the A nucleotide appears in most of the population, then C is the minor allele and A is the major allele at this position, and the frequency of this SNP is 20%.

1.3.2 Hardy-Weinberg Equilibrium

The concept of Hardy-Weinberg equilibrium was proposed by Hardy (1908) and Weinberg (1908), and it states that allele frequencies and genotype frequencies in a population will remain constant under the assumptions that the population has infinite size, is randomly mating, and is free from evolutionary influences, for example natural selection, migration, and mutation.

Let us consider a simple case with a locus that have two alleles, denoted by A and a. Assume the frequencies for A and a are $f_A = p$ and $f_a = q$, respectively, where $p + q = 1$. By the Hardy-Weinberg equilibrium, the expected frequency for genotypes are given,

$$f_{AA} = p^2,$$

$$f_{Aa} = 2pq,$$

$$f_{aa} = q^2.$$

The frequencies for alleles A and a, and the frequencies for genotypes AA, Aa, and aa will remain constant from generation to generation with no evolutionary influences.

In reality, not all the assumptions are necessarily satisfied, and a deviation may be observed. Statistical methods that compare the expected and observed frequencies, such as Guo and Thompson (1992), can be used to test the equilibrium. These tests are often used in the quality control of genetic studies, to screen out the genetic variants that severely violate the Hardy-Weinberg equilibrium.

1.3.3 Linkage Disequilibrium

The SNPs at different loci are not independent of each other, which greatly complicates the multiple testing problem in GWAS. In fact, the SNPs in a close neighborhood tend to be

correlated with each other. The frequency of one combination of different alleles is higher or lower than what would be when they were independent, and this phenomenon is known as linkage disequilibrium. There are several factors that may lead to the dependence between SNPs, including inheritance pattern, population structure, and genetic factors. One important factor is recombination, which refers to the fact that during the process of production of offspring, chromosomes break at locations and the loci on two sides separate from each other. The chromosomes do not break equally easily at different locations, and the loci in a segment that is difficult to break tend to be highly correlated with each other.

Linkage disequilibrium can be measured in different ways, and a comparison of different measures can be found in Devlin and Risch (1995). We describe several commonly used statistics that measure the pairwise linkage disequilibrium between two loci. Considering two neighboring loci in a population, suppose allele A occurs with frequency f_A at one locus, allele B occurs with frequency f_B at the other locus, and the haplotype AB occurs with frequency f_{AB}. If the two alleles A and B are independent, then $f_{AB} = f_A f_B$. If the independence is not satisfied, there would be a discrepancy between the two quantities f_{AB} and $f_A f_B$. The discrepancy is one of the earliest measures of linkage disequilibrium, and the coefficient of linkage disequilibrium is defined as

$$D = f_{AB} - f_A f_B.$$

A greater value of D implies that the alleles A and B are more in linkage disequilibrium. The definition of D is simple, however not always convenient in real applications, as the range of D depends on the allele frequencies and this makes the comparison of the strength of linkage disequilibrium difficult. Several other measures of linkage disequilibrium have been proposed to facilitate the comparison, and we mention two most common alternatives.

The maximum of D can be derived as

$$D_{max} = \begin{cases} min\{f_A f_B, (1-f_A)(1-f_B)\} & \text{if D} < 0; \\ min\{f_A(1-f_B), (1-f_A)f_B\} & \text{if D} > 0. \end{cases}$$

Lewontin (1964) propose to normalize D by dividing its maximum,

$$D' = \frac{D}{D_{max}}.$$

Another important measure r is proposed by Hill and Robertson (1968), and it is defined as

$$r = \frac{D}{\sqrt{f_A(1-f_A)f_B(1-f_B)}},$$

and it can be easily proved that r is equivalent to the Pearson correlation coefficient between the indicator variables of the major alleles or minor alleles. The values of r are between -1 and 1. A value of $r = 1$ or -1 indicates the two alleles are in perfect linkage disequilibrium and there are only two possible combinations for the haplotypes, while a value of $r = 0$ indicates the independence between the alleles.

In the GWAS projects of Chapters 3–4, we shall use the correlation coefficient r to measure the level of linkage disequilibrium.

1.3.4 Genetic Models

A genetic model describes the relationship between genotype and phenotype. The phenotype can be binary for characteristics or diseases, for example curly hair and type I diabetes, or continuous for traits, for example body mass index (BMI). A causal variant is the locus that contributes to the variation in the phenotype. Depending on the number of loci that have an influence on the phenotype, there are simple traits and complex traits. Simple traits only have one causal locus, for example whether people are able to curl up the sides of tongue and sickle cell anemia, while complex traits have more than one causal locus, for example whether people have curly hair and schizophrenia.

For simple traits, there are several models that can be used to describe the genetic relationship between genotype and phenotype. We first define a penetrance function,

$$\Pr(Y|G) = \begin{cases} a\ conditional\ probability & \text{if Y is binary;} \\ a\ conditional\ density & \text{if Y is continuous,} \end{cases}$$

where Y is the phenotype and G is the genotype for a causal variant. Assume the disease allele is D and its counterpart is d, and the genotype G can be DD, Dd and dd. To simplify, we assume the phenotype is binary and Y represents whether a disease (or characteristic) is observed.

Mendelian models, also named deterministic models describe the genetic models for simple traits. Depending on the different modes of inheritance, we can have several genetic models. In a dominant model, the disease allele D is dominant in determining the phenotype and one occurrence of D would lead to the disease, i.e.,

$$\Pr(Y = 1|DD) = \Pr(Y = 1|Dd) = 1,\ \ \Pr(Y = 1|dd) = 0.$$

The genotype scores are often coded as 1 for DD or Dd, and 0 for dd.

In a recessive model, the disease allele D is recessive, and only the genotype DD would lead to the disease, i.e.,

$$\Pr(Y = 1|DD) = 1, \;\; \Pr(Y = 1|Dd) = \Pr(Y = 1|dd) = 0.$$

The genotype scores are coded as 1 for DD, and 0 for Dd or dd.

Another important model is called the additive/dosage model, which indicates the disease risk is additive in terms of the occurrence of disease allele D,

$$\Pr(Y = 1|DD) > \Pr(Y = 1|Dd) > \Pr(Y = 1|dd).$$

The genotype scores are often coded as the number of disease alleles, 2 for DD, 1 for Dd, and 0 for dd.

In general, the mode of inheritance is unknown, and the additive model is most commonly used in genetic studies.

For complex traits, we often assume that each locus contributes a small amount to the phenotypic variability, and the additive model is also the most common. In Chapters 3–4, we focus on developing statistical methods for detecting associated variants of complex traits using the additive model.

1.3.5 Univariate Test

The number of SNPs genotyped in genetic studies can be close to $1,000,000$ or more, while the number of subjects is only a few thousand. This is a "big n, small p" problem, and univariate tests have been proposed to measure the association between SNPs and the phenotype. Assume Y is the vector of phenotypes for n individuals, and X is the n by M matrix that contains the genotype scores of M SNPs that are coded as the number of minor alleles $\{0 = no\ minor\ allele, 1 = 1\ minor\ allele, 2 = 2\ minor\ alleles\}$ in the additive model.

For continuous phenotype, a simple linear regression is usually used,

$$Y = \beta_0 + \beta_i X_i + \epsilon,$$

for $i = 1, 2, \ldots, M$, where β_0 is the intercept, β_i is the coefficient, X_i is the vector of genotype scores for SNP i, and ϵ is the error term. The maximum likelihood estimates of the parameters in the model can be computed, and the t test statistics and corresponding p-values can be obtained for all SNPs.

For discrete phenotypes, several methods have been used to test the association. We can compute a contingency table crossing the genotype and phenotype, and χ^2 test is then applied to test the independence of the table. If all the cells in the table have expected frequencies > 5, a Pearson χ^2 test can be conducted. Otherwise, alternatives such as Fisher's exact test can be used. The Cochran-Armitage test (Cochran 1954; Armitage 1955), also known as trend test, is also commonly used in case-control studies.

Another common test used in discrete phenotypes is the logistic regression model. Similar to the simple linear regression used in continuous phenotypes, a logistic regression model can be fitted to the genotype scores of each SNP,

$$\log \frac{p_j}{1-p_j} = \beta_0 + \beta_i X_i,$$

for $i = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, n$, where $p_j = \Pr(Y_j = 1|X_i)$ is the probability of individual j having the disease or characteristic given SNP i. The parameters can be estimated by maximized likelihood method, and the Wald statistics and corresponding p-values are computed. Note that it can be proved that the Cochran-Armitage test is equivalent to the score test under the logistic regression model.

After computing marginal p-values for all the SNPs, the common practice in GWAS is to compare the p-values to the Bonferroni-corrected threshold 5×10^{-8} to control the family-wise error rate at level 0.05. However, as we argued in Section 1.1.2 and will demonstrate in Chapter 3, the control of the family-wise error rate can be overly-conservative and have limited detection power, and we are interested in developing statistical methods for GWAS that are able to reasonably control the false discovery rate (FDR) instead.

1.3.6 Penalized Multiple Regression

To overcome the low detection power problem of marginal p-values, many methods have been proposed to apply penalized regression methods to all SNPs, for instance see Hoggart et al. (2008), Wu et al. (2009), and Hoffman et al. (2013).

A penalized multiple regression can be used in GWAS, and the estimates of regression coefficients are given as,

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}}\ l(y|\beta) - \sum_j p_\theta(\beta_j),$$

where y is the vector of phenotype values, β is the vector of coefficients, $l(\cdot)$ is the log-likelihood function for a linear regression (for continuous phenotypes) or a logistic regression

(for discrete phenotypes), and $p(\cdot)$ is the penalty function indexed by the vector of tuning parameters θ. The penalty function $p(\cdot)$ needs to satisfy the sparsity property as only a small subset of the SNPs are causal SNPs.

Different penalty functions have been used by researchers. Wu et al. (2009) use the Lasso penalty, Hoggart et al. (2008) use the normal-exponential-gamma (NEG) penalty due to Griffin and Brown (2011), and Hoffman et al. (2013) compare various penalties including Lasso (Tibshirani, 1996), Adaptive Lasso (Zou, 2006), NEG, minimax concave penalty (Zhang et al., 2010) and others.

The penalized regression methods are able to improve the detection power compared to the marginal p-values, however the selection of associated SNPs is often arbitrary due to the high correlations among neighboring SNPs. Moreover, another drawback of penalized regression in GWAS is that they only compute the coefficient estimates with no standard deviations or p-values, and provide no error rate control, which is not always convenient. In Chapters 3–4, we aim to approximately control the FDR in GWAS.

Chapter 2

Simultaneous testing of composite null hypotheses with discrete data: an application to pharmacovigilance

2.1 Introduction

As a motivating example, we look at the post-marketing surveillance of pharmaceutical drugs. Drugs are widely used to treat or prevent diseases, but they could also cause minor to severe adverse drug reactions (ADRs), such as nausea, skin rash, or even death. The World Health Organization, the European Union and many countries have established pharmacovigilance databases to monitor the marketed drugs for detection of ADRs.

In any pharmacovigilance database, we can extract the number of reported cases for each combination of adverse event and drug. Therefore, the pharmacovigilance data can be summarized as a large $I \times J$ contingency table crossing I adverse events and J drugs. For a specific pair (adverse effect i, drug j), a 2 by 2 contingency table can be obtained by collapsing the large contingency table, as presented in Table 1.3. In the table, n_{ij}, $n_{i.}$, $n_{.j}$ and n denote the number of reports involving the pair (i, j), the marginal count for the adverse event i, the marginal count for the drug j and the total count in the contingency table, respectively.

Let ψ denote the odds ratio, which is a measure of the association between adverse event and drug. In pharmacovigilance, we are interested in finding strong positive associations between drugs and ADRs. Thus, we test the following hypothesis for adverse event i and

drug j,

$$H_{0_{ij}} : \psi_{ij} \leq \psi_0 \text{ versus } H_{1_{ij}} : \psi_{ij} > \psi_0, \tag{2.1}$$

where ψ_0 is a target threshold for odds ratio. An adverse event-drug combination (i, j) will be investigated further if $H_{0_{ij}}$ is rejected, which means that drug j may lead to significantly more adverse event i than expected by chance. It is common to choose $\psi_0 = 1$, but larger ψ_0 threshold are also widely used to focus on strong associations.

In Equation (2.1), the alternative hypothesis is one-sided, and the null hypothesis is composite. Testing a composite null hypothesis is challenging because the distribution of test statistic under the null is difficult to characterize. Furthermore, the number of reports of any adverse event-drug combination is discrete, and the corresponding test statistic is also discrete. Finally, there are multiple associations being tested simultaneously. As a concrete example, the pharmacovigilance database in the United Kingdom contained $1,617$ drugs and $1,310$ adverse events by September 2014, and there are slightly over 2 million unique combinations of adverse events and drugs.

To take the multiplicity into account, we will focus on estimating and controlling the false discovery rate (FDR; Benjamini and Hochberg 1995), which is a more powerful alternative than the family-wise error rate. Few statistical methods have been proposed to test multiple composite null hypotheses with discrete data, and most existing methods can only address either the issue of discreteness or composite null. Methods to control the FDR for discrete data under simple null hypothesis include Gilbert (2005), Kulinskaya and Lewin (2009), Heyse (2011), Habiger (2015), Chen et al. (2018) and Döhler et al. (2018), among others. On the other hand, Sun and McLain (2012) study multiple testing of composite null hypotheses with continuous test statistics.

Methods that are specific to the multiple testing of pharmacovigilance signal detection have been developed. Existing methods including the reporting odds ratio (ROR; van Puijenbroek et al. 2002), the Bayesian confidence propagation neural network (BCPNN; Bate et al. 1998) and the Gamma Poisson Shrinker (GPS; DuMouchel 1999) are used to automatically detect signals from a large number of reports. One notable drawback of these methods is that they use ad hoc thresholds to detect signals and do not consider the multiplicity. Ahmed et al. (2010) propose a frequentist method that models the mid-p-value distribution through a finite mixture model after filtering out the hypotheses with small cell counts. Through a simulation study based on the French pharmacovigilance database, Ahmed et al. (2010) showed their method's FDR estimates are conservative with respect to the true FDR levels. However, their threshold for cell count filtering is arbitrary. More importantly, filtering the hypotheses with small cell counts is equivalent to filtering the hypotheses with large p-values and can lead to anti-conservative results.

Ahmed et al. (2009) extend the two existing Bayesian methods used in pharmacovigilance signal detection, the BCPNN by Bate et al. (1998) and the GPS by DuMouchel (1999), to the multiple testing setting. The extended GPS method showed better performance than the extended BCPPN method and the original GPS and BCPPN. The GPS is a parametric method which assumes a two-component gamma distribution for relative risks. However, the strong parametric assumption of GPS may not always be reasonable for real applications.

The chapter is organized as follows. In Section 2, we investigate an oracle procedure based on the local FDR, which is shown to have the best power given a constraint on the FDR level. Then, we propose a nonparametric empirical Bayes method to approximate the oracle statistic. In Section 3, we conduct simulation studies to illustrate the advantages of our proposed procedure over existing methods. In Section 4, we apply the proposed procedure to the UK pharmacovigilance database.

2.2 Method

In Benjamini and Hochberg (1995), a simple procedure based on p-values (the BH procedure) was proposed and shown to control the FDR under independence. They assume p-values under true null follow independent $Uniform(0,1)$ distribution, which is a reasonable assumption when the null hypotheses are simple and the observations are continuous. If the null hypotheses are composite and the observations are discrete, the p-values under true null are stochastically larger than $Uniform(0,1)$, and a direct application of the BH procedure can be severely conservative. In the pharmacovigilance example, let N_{ij} be the number of reports of the ith adverse event after taking the jth drug, a reasonable model is that N_{ij} follows a noncentral hypergeometric distribution given the marginal counts $n_{i.}$, $n_{.j}$, n and the true odds ratio ψ_{ij} (Agresti, 2002). Given the odds ratio threshold ψ_0 that is tested, the p-values corresponding to the null hypothesis in Equation (2.1) can be computed by the Fisher's exact test as $p_{ij} = \Pr(N_{ij} \geq n_{ij} | n_{i.}, n_{.j}, n; \psi_0)$. Ahmed et al. (2010) propose a frequentist method to estimate the FDR by modeling the mid-p-values (Lancaster, 1961) under null hypotheses as a mixture of a uniform distribution and a non-decreasing density function. The fundamental question we will address first is whether (mid-)p-values are the best statistics to use for testing multiple composite null hypotheses with discrete data.

2.2.1 Oracle Procedure to Control the FDR

Assume a set of unknown parameters $\theta_1, \theta_2, \ldots, \theta_m$ are independently distributed as

$$\theta_i \sim g(\theta),$$

for $i = 1, 2, \ldots, m$. Suppose $X_1,\ X_2,\ \ldots,\ X_m$ are the the discrete test statistics that are independently distributed with the probability mass function,

$$p(x) = \int_{\Theta} f(x|\theta) g(\theta) d\theta,$$

where $f(x|\theta)$ is a known distribution for the test statistics X given the true value of θ, and Θ is the support for the parameter θ. Consider the problem of simultaneously testing m independent composite null hypotheses,

$$H_{0_i} : \theta_i \in \Theta_0 \text{ versus } H_{1_i} : \theta_i \notin \Theta_0, \tag{2.2}$$

for $i = 1, 2, \ldots, m$, where Θ_0 is the support of θ for true null hypotheses.

The probability mass function of X corresponding to true null hypotheses is

$$p_0(x) = \int_{\Theta_0} f(x|\theta) g(\theta) d\theta.$$

Let V denote the number of falsely rejected null hypotheses and R the total number of rejections, and the FDR is defined as

$$\text{FDR} = E\left(\frac{V}{R \vee 1}\right),$$

where $R \vee 1 = max(R, 1)$. For any rejection region S, the marginal FDR (mFDR) is defined as

$$\begin{aligned} \text{mFDR} &= \frac{E(V)}{E(R)} \\ &= \frac{\sum_{x \in S} p_0(x)}{\sum_{x \in S} p(x)}. \end{aligned}$$

The Bayesian FDR (Fdr) was proposed by Efron and Tibshirani (2002), and it is defined as

$$\begin{aligned} \text{Fdr}(S) &= \Pr(\theta \in \Theta_0 | x \in S) \\ &= \frac{\sum_{x \in S} p_0(x)}{\sum_{x \in S} p(x)}. \end{aligned}$$

Note that the Bayesian FDR is the posterior probability that a hypothesis is true null given it is in the rejection region, and $\text{mFDR}(S) = \text{Fdr(S)}$.

The local FDR (fdr; Efron et al. 2001), as the posterior probability of being true null, can be defined as follows,

$$\text{fdr}(x) = \frac{p_0(x)}{p(x)}.$$

In the literature of multiple testing with simple null and continuous test statistics, the local FDR based procedure has been shown to be superior to p-value based procedures (Sun and Cai, 2007). It is straightforward to show that the Bayesian FDR is the conditional expectation of the local FDR.

Theorem 1. $\text{Fdr}(S) = E_p\{\text{fdr}(x)|x \in S\}$.

Proof.

$$\begin{aligned} E_p\{\text{fdr}(x)|x \in S\} &= \frac{E_p\{\text{fdr}(x); x \in S\}}{\Pr(x \in S)} \\ &= \frac{\sum_{x\in S}[p_0(x)/p(x)]p(x)}{\sum_{x\in S} p(x)} \\ &= \frac{\sum_{x\in S} p_0(x)}{\sum_{x\in S} p(x)} \\ &= \text{Fdr}(S). \end{aligned}$$

□

This result suggests that for a set of rejections $\mathcal{R}$, its Bayesian FDR can be empirically estimated as $\sum_{x_i\in S} \text{fdr}(x_i)/|\mathcal{R}|$, for $i = 1, 2, \ldots, m$, where $|\mathcal{R}|$ is the number of rejections.

Genovese and Wasserman (2002) showed that under independence, mFDR = FDR + $O(m^{-1/2})$. Therefore, the FDR and the Bayesian FDR are asymptotically equivalent, as the marginal FDR is equivalent to the Bayesian FDR. This suggests that we can asymptotically control the FDR by controlling the Bayesian FDR.

We now establish the optimality of the local FDR based procedure in the context of testing composite nulls with discrete data.

Theorem 2. *For the multiple testing problem in (2.2), define the oracle rejection region of the form* $S_{OR} = \{x : \text{fdr}(x) \le C\}$, *where* C *is a cut-off for the local* FDR. *Suppose the Bayesian* FDR *level of* S_{OR}, $\text{Fdr}(S_{OR}) = \alpha$. *For any rejection region* S *such that* $\text{Fdr}(S) \le \alpha$, S_{OR} *yields more expected true positives.*

The proof is given in Appendix, and it is similar in spirit to Proposition 1 of Du et al. (2014), which focus on the setting of simple null and continuous test statistics. Theorem 2 shows that the local FDR is the optimal test statistic to form the rejection region. Then, we propose an oracle procedure to control the Bayesian FDR as follows,

1. Order all the hypotheses by their fdr values and separate them into q groups with distinct ordered fdr values denoted by $\text{fdr}_{(1)}, \ldots, \text{fdr}_{(q)}$;
2. Reject the hypotheses in groups $i = 1, \ldots, k$ with the smallest fdr values, where $k = max\{j : \frac{\sum_{i=1}^{j} m_i \text{fdr}_{(i)}}{\sum_{i=1}^{j} m_i} \leq \alpha\}$, where m_i is the number of fdr values in group i.

Because of the discreteness of the data, we handle ties by dividing them into groups of identical fdr values in the oracle procedure. It is easy to see that the Bayesian FDR level with $\text{fdr}_{(k)}$ as the cut-off is $\frac{\sum_{i=1}^{k} m_i \text{fdr}_{(i)}}{\sum_{i=1}^{k} m_i}$.

2.2.2 Adaptive Procedure to Control the FDR

The oracle procedure uses the fdr, which require the oracle knowledge of the mixing distribution g. To approximate the oracle procedure, we first need to estimate g.

We use the predictive recursion (PR) method (Newton, 2002) to estimate the mixing distribution g. We choose the PR because of its computation efficiency and convergence properties under weak conditions (Tokdar et al., 2009). This computation efficiency is crucial for applications with a large number of observations, which is the case for pharmacovigilance data.

More specifically, the sampling distribution of cell count N_{ij} given $n_{i.}, n_{.j}$ and n follows a noncentral hypergeometric distribution with parameter ψ_{ij}. The initial density estimate of $g(\psi)$ is chosen as a uniform distribution, and it is shown to have no significant effects on the final estimate $g_m(\psi)$ due to its convergence property (Tokdar et al., 2009). Moreover, as the predictive recursion algorithm is a sequential procedure, the final estimate depends on the order of data. As suggested by the literature, We randomly permute the data and obtain an average density estimate $\bar{g}_m(\psi)$. Further details about our implementation of the predictive recursion are included in Appendix.

After we obtain the PR estimate of the odds ratio ψ density $\bar{g}_m(\psi)$, we estimate the local FDR as

$$\widehat{\text{fdr}}(x) = \frac{\int_{\Theta_0} f(x|\theta)\bar{g}_m(\theta)d\theta}{\int_{\Theta} f(x|\theta)\bar{g}_m(\theta)d\theta},$$

where Θ_0 is the parameter support corresponding to null hypothesis, i.e. $\psi \leq \psi_0$ in our multiple testing problem for pharmacovigilance data.

We replace the local FDR in the oracle procedure with its estimates, and naturally have the adaptive procedure to approximate the oracle procedure.

2.3 Simulations

In this section, candidate methods and data generation process are first described. Then simulation results to test $\psi_0 = 1$ and $\psi_0 = 5$ are presented to illustrate the performance of candidate methods.

2.3.1 Candidate Methods

We consider the following methods:

`midP`: the mid-p-value frequentist method proposed in Ahmed et al. (2010). The adverse event-drug combinations with small cell counts n_{ij}'s are filtered. Ahmed et al. (2010) considered filtering criteria of $n_{ij} < 1$ and $n_{ij} < 3$, and here we filter all combinations with empty cells, i.e. $n_{ij} < 1$.

`Pval`: because the use of mid-p-values and the filtering of small cell counts can lead to anti-conservative results, we consider the use of all p-values. The distribution of p-values has a U shape due to the testing of composite nulls. Similar to Ahmed et al. (2010), we applied the LBE method (Pounds and Cheng, 2006) to estimate the flattened proportion in the center of the distribution of p-values and adjust the FDR estimate accordingly. Technical details of this p-value based procedure are presented in Appendix.

`GPS`: the extended GPS method suggested in Ahmed et al. (2009). `GPS` is implemented to test the relative risk (denoted by ϕ) of the combinations of drugs and adverse events instead of the odds ratio ψ, but there is no one-to-one mapping between them. However, it is equivalent to test $\phi \leq 1$ and $\psi \leq 1$, and we will only show the simulation result of `GPS` when $\psi_0 = 1$. The estimate of the local FDR can be obtained as the posterior probability $\Pr(\phi_{ij}|n_{ij} \leq \phi_0)$, and the same rejection procedure as our proposed method can be applied to hypothesis testing.

`PRfdr`: our proposed method with the density of odds ratio ψ estimated by predictive recursion (PR). As in Martin and Tokdar (2009), 10 random permutations of the data

were performed, and the average of the density estimates was computed as $\bar{g}_m$ to avoid the dependence on the order of data.

`KDfdr`: same as `PRfdr` except that we use the kernel density estimate of odds ratios based on the true ψ_{ij} values. `KDfdr` is considered as the "oracle" because the true ψ_{ij} values are not known in real applications.

2.3.2 Data Generation

We generate data similarly as simulation studies of Ahmed et al. (2010). We first obtain the marginal counts for adverse events and drugs of the UK pharmacovigilance data as $n_{1.}, \ldots, n_{I.}$ and $n_{.1}, \ldots, n_{.J}$, respectively. For one iteration, the probability of observing adverse event i for drug j is

$$p_{ij} = \frac{\lambda_{ij} r_{i.} r_{.j}}{\sum_{ij} \lambda_{ij} r_{i.} r_{.j}},$$

where $r_{1.}, \ldots, r_{I.} \sim Dirichlet(n_{1.}, \ldots, n_{I.})$ and $r_{.1}, \ldots, r_{.J} \sim Dirichlet(n_{.1}, \ldots, n_{.J})$.

For the purpose of computational efficiency and comparison with the original data, a subset with 200 drugs and 200 adverse events randomly sampled from the original UK data was also used in the simulation. Three different settings of λ distributions are investigated:

1. the sampled data with $ln(\lambda_{ij}) \sim 0.3N(-1, 0.01) + 0.5N(0, 0.1) + 0.2N(1, 1)$;
2. the sampled data with $ln(\lambda_{ij}) \sim 0.5N(0, 1) + 0.5N(1, 2)$;
3. the original data with $ln(\lambda_{ij}) \sim logistic(0, 0.5)$.

The first two settings correspond to a three-component and a two-component normal distribution for $ln(\lambda)$, respectively, and the third setting is the same as that in Ahmed et al. (2010).

With the probabilities p_{ij} known, the cell counts follow a multinomial distribution and can be generated as

$$n_{11}, n_{12}, \ldots, n_{IJ} \sim Multinomial(n, (p_{11}, p_{12}, \ldots, p_{IJ})),$$

where n is the sum of cell counts in the UK data.

The true odds ratio for cell (i, j) can be computed as

$$\psi_{ij} = \frac{p_{ij}(1 + p_{ij} - p_{i\cdot} - p_{\cdot j})}{(p_{i\cdot} - p_{ij})(p_{\cdot j} - p_{ij})},$$

and the true relative risk is

$$\phi_{ij} = \frac{p_{ij}}{p_{i\cdot}p_{\cdot j}},$$

where $p_{i\cdot}$ and $p_{\cdot j}$ are the ith row sum and the jth column sum of p.

2.3.3 Simulation Results

We repeat the simulation 200 times and report the average realized FDR levels and power. The target FDR levels are between 0.01 and 0.2.

Figure 2.1 shows the realized FDR levels versus the target FDR levels of five candidate methods in the three simulation settings for testing $\psi_0 = 1$ and $\psi_0 = 5$. We can clearly see that, the realized FDR levels for `KDfdr` and `PRfdr` are very close to the target FDR levels, which means that they can control the true FDR very well, and the `PRfdr` yields almost the same results to the "oracle" `KDfdr`. Both `Pval` and `midP` are able to control the FDR in sets 1 and 2, as the realized FDR levels are smaller than the target FDR; `midP` performs better than `Pval` as the realized FDR levels of `midP` are closer to the target FDR level compared to those for `Pval`. It can also be seen that the improvement in controlling the FDR by the adoption of mid-p-values decreases for $\psi_0 = 5$ than $\psi_0 = 1$. In set 3, the FDR control of `Pval` for $\psi_0 = 1$ is first slightly liberal, then turns slightly conservative; `Pval` yields a conservative FDR control for $\psi_0 = 5$. The use of mid-p-values in set 3 makes `midP` very liberal in FDR control. `GPS` is liberal in terms of the FDR control in set 1 as the realized FDR levels are greater than the target FDR levels, while it is conservative in set 2. In set 3, `GPS` is slightly liberal for the target FDR levels larger than 0.1.

Figure 2.2 shows the number of true positives versus the target FDR levels. Among the methods that can control the FDR in each setting, `KDfdr` and `PRfdr` are noticeably more powerful than `Pval`, `midP` and `GPS`. Compared to `Pval`, `midP` has more true positives, and the improvement of power for $\psi_0 = 1$ is greater than that for $\psi_0 = 5$. In sets 1 and 3, the number of true positives for `GPS` is slightly larger than those of `KDfdr` and `PRfdr`, and the reason is `GPS` liberal in the FDR control and yields more signals, including true positives and false positives, than expected. This result also applies to `midP` for $\psi_0 = 1$ in set 3.

Figure 2.3 plots the true positive rate versus the false positive rate, which is known as the receiver operating characteristic (ROC) curve. We can compare the goodness of ranking of test statistics in a multiple testing procedure by the ROC curve. A multiple testing procedure is composed of ranking and thresholding test statistics, and a good ranking of test statistics will often lead to a good testing procedure. The lines for `KDfdr` and `PRfdr` are mostly overlapping, and their rankings of test statistics are significantly better than those for `Pval` and `midP`, which can explain why `KDfdr` and `PRfdr` can improve the control of FDR and yield more power. We can also see that the ranking of `Pval` is slightly better than that of `midP` as the mid-p-values have distorted the original ranking of p-values. For $\psi_0 = 1$, the lines for `GPS` are overlapping with those for `KDfdr` and `PRfdr` in sets 1–3, and they have similar rankings of test statistics.

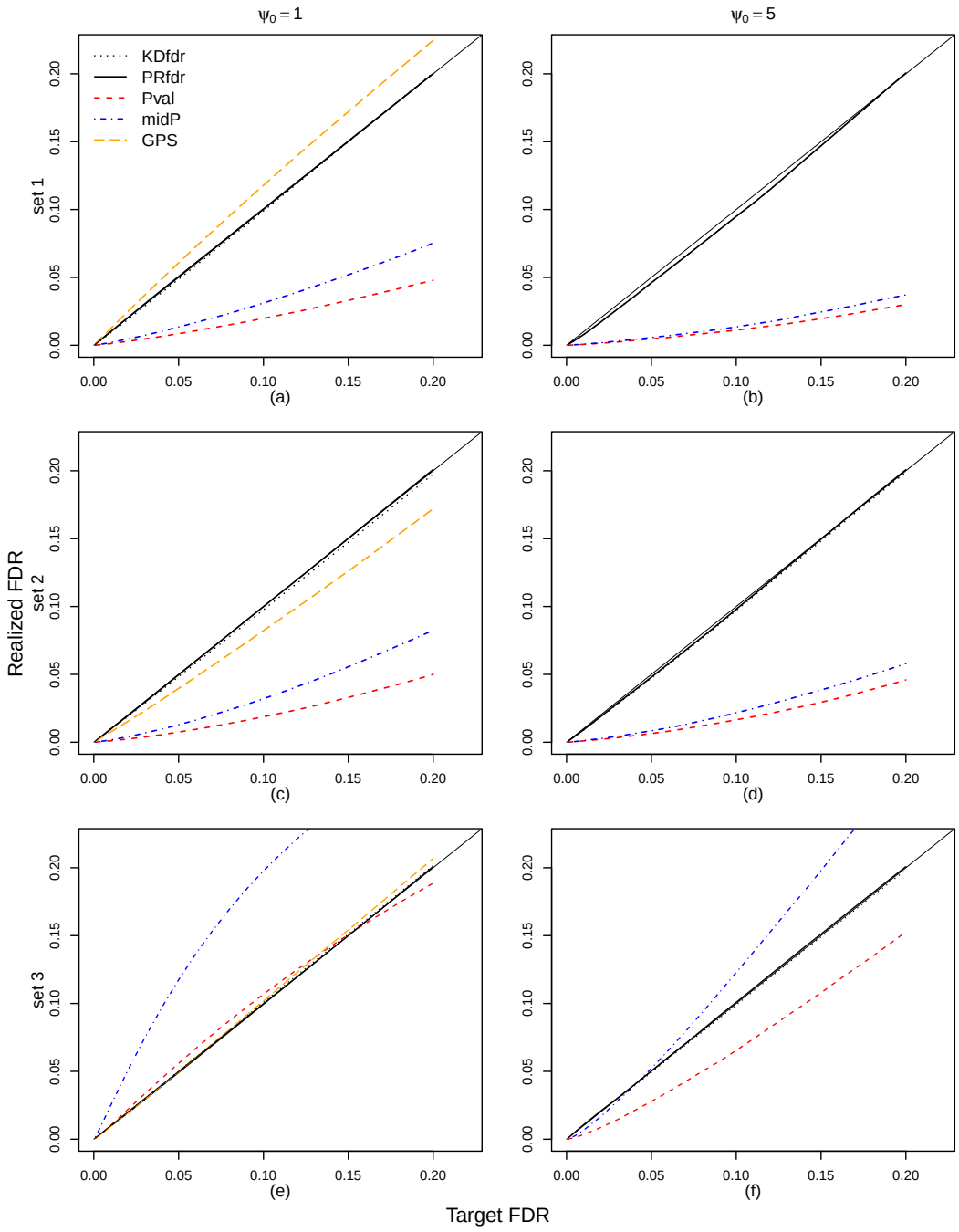

Figure 2.1: Realized FDR levels (y axis) versus target FDR levels (x axis). Rows 1, 2 and 3 show the plots from sets 1, 2 and 3, respectively. Reference diagonal line as gray line

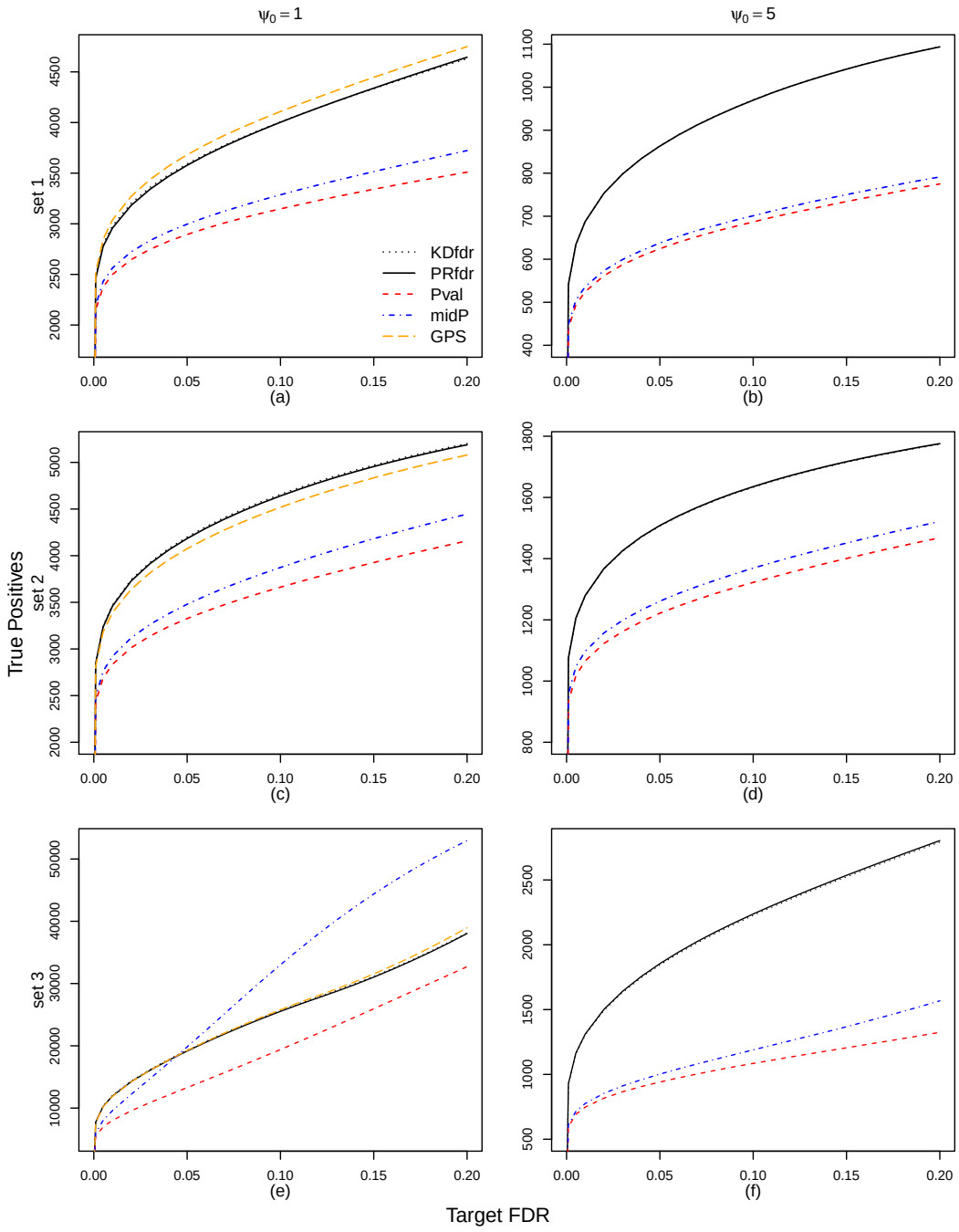

Figure 2.2: Number of true positives (y axis) versus target FDR levels (x axis). Rows 1, 2 and 3 show the plots from sets 1, 2 and 3, respectively

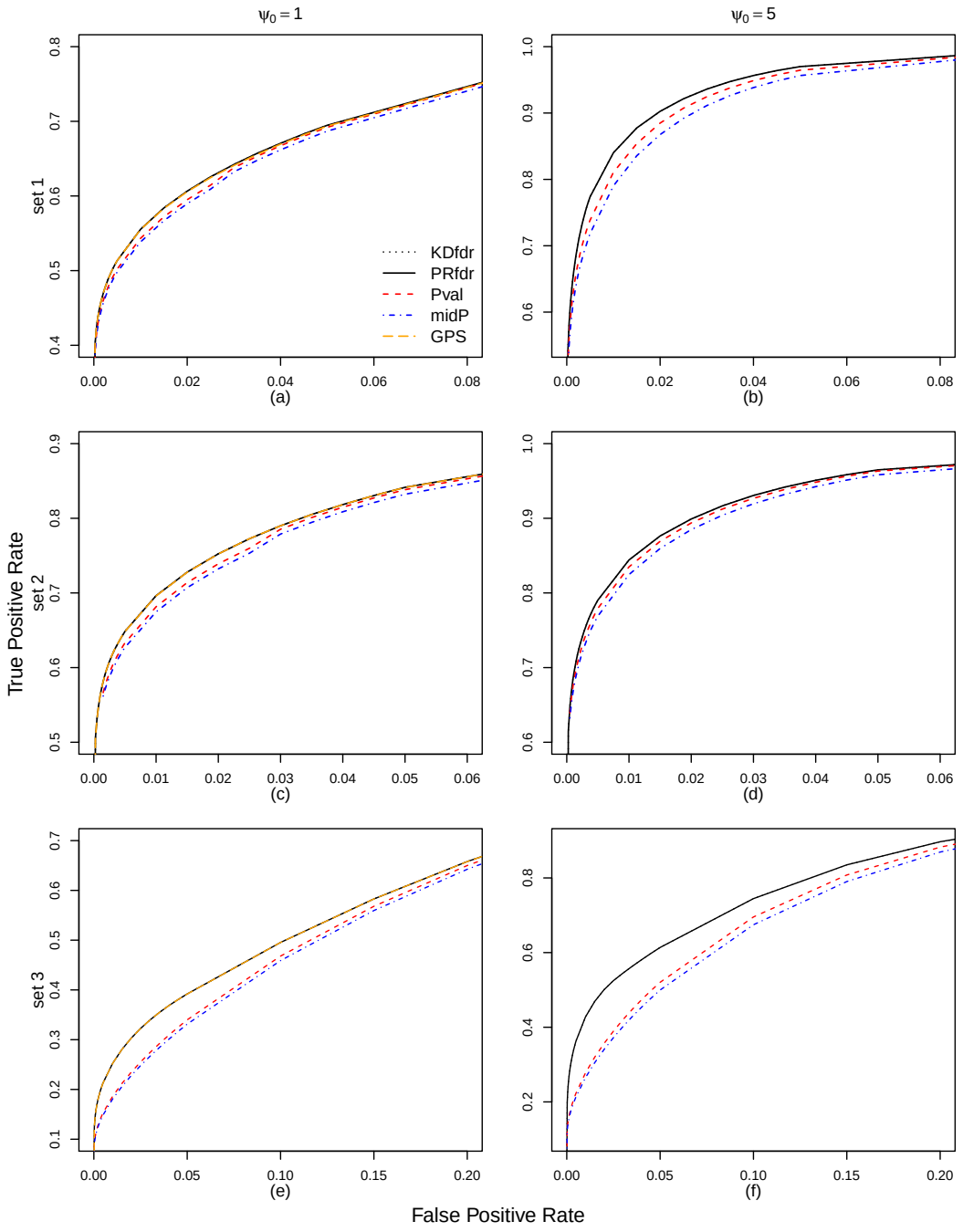

Figure 2.3: True positive rate (y axis) versus false positive rate (x axis). Rows 1, 2 and 3 show the plots from sets 1, 2 and 3, respectively

As a summary, we have shown that `KDfdr` and `PRfdr` outperform `Pval`, `midP` and `GPS` in the sense that `KDfdr` and `PRfdr` can better control the FDR, yield more true positives and provide a more efficient ranking of test statistics. `Pval` can control the FDR in sets 1–3, but the realized FDR levels are not as close to the target FDR levels and it yields significantly less power; `midP` can improve the control of FDR and power in sets 1 and 2, but it is liberal in the FDR control in set 3. `GPS` can be liberal or conservative in the control of FDR and it has less power than `KDfdr` and `PRfdr` even when the realized FDR levels are below the target FDR levels.

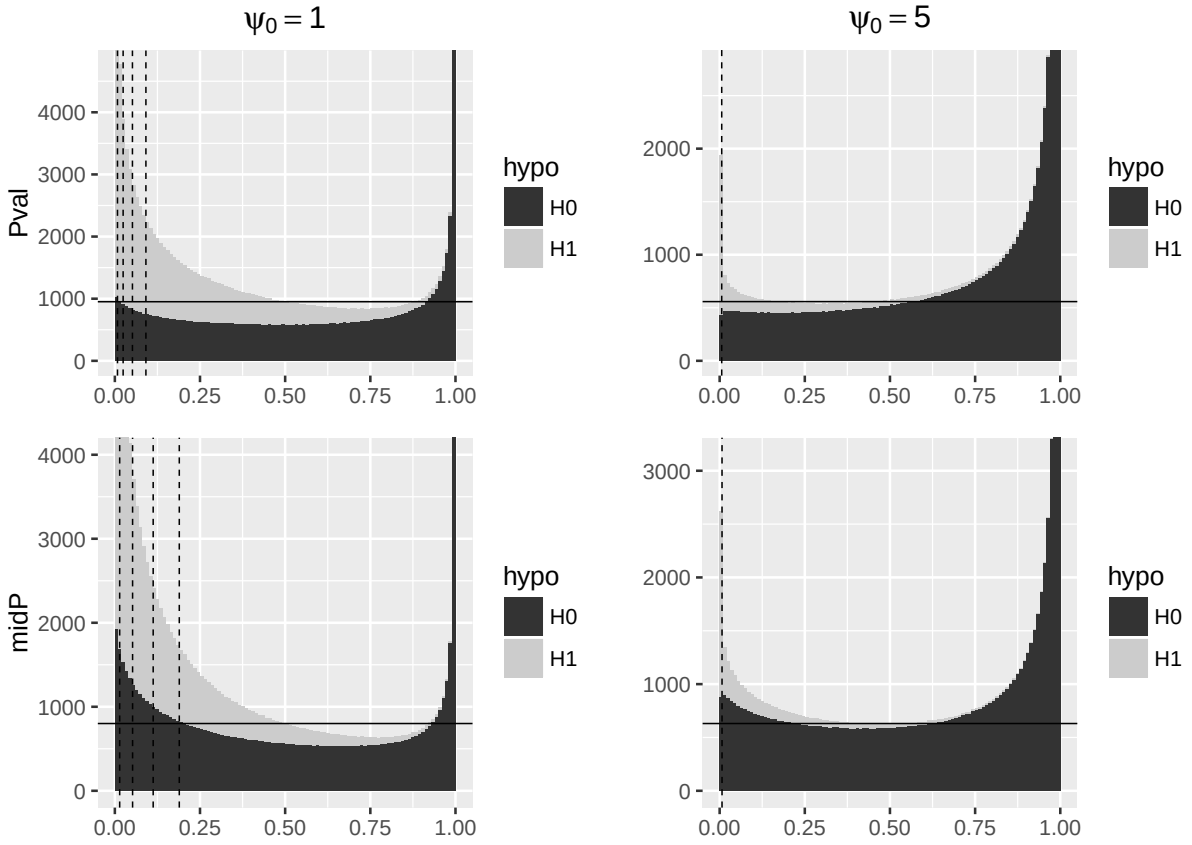

Figure 2.4: Average histograms of p-values and mid-p-values

The result that `midP` fails to control the FDR in set 3 is contradictory to the observation and recommendation in Ahmed et al. (2010), where they argued that the use of mid-p-values can improve the performance of the multiple testing procedures in pharmacovigilance databases. One key assumption in Ahmed et al. (2010) is that the distribution

of mid-p-values under null hypotheses is a non-decreasing function. In Figure 2.4, the average histograms, over 200 iterations, of p-values and mid-p-values from null (black area) and alternative (grey area) hypotheses for $\psi_0 = 1$ and $\psi_0 = 5$ are plotted. The horizontal solid line indicates the LBE estimate of the flattened proportion $\pi_0\pi_{0*}$, while the vertical dash lines show the threshold values for p-values or mid-p-values corresponding to the target FDR levels 0.05, 0.1, 0.15 and 0.2 for $\psi_0 = 1$ and 0.2 for $\psi_0 = 5$. We can see that the distribution of p-values or mid-p-values from null hypotheses are not necessarily non-decreasing. The histogram of mid-p-values for $\psi_0 = 1$ shows the most decreasing distribution on the left side, mid-p-values for $\psi_0 = 5$ the second, p-values for $\psi_0 = 1$ the third, and p-values for $\psi_0 = 5$ the least. This observation contradicts the assumption of non-decreasing function. Their simulation results are based on the French pharmacovigilance data, which is not publicly available. The non-deceasing true null density assumption may be more appropriate for the French data but certainly is not universally satisfied.

2.4 Application

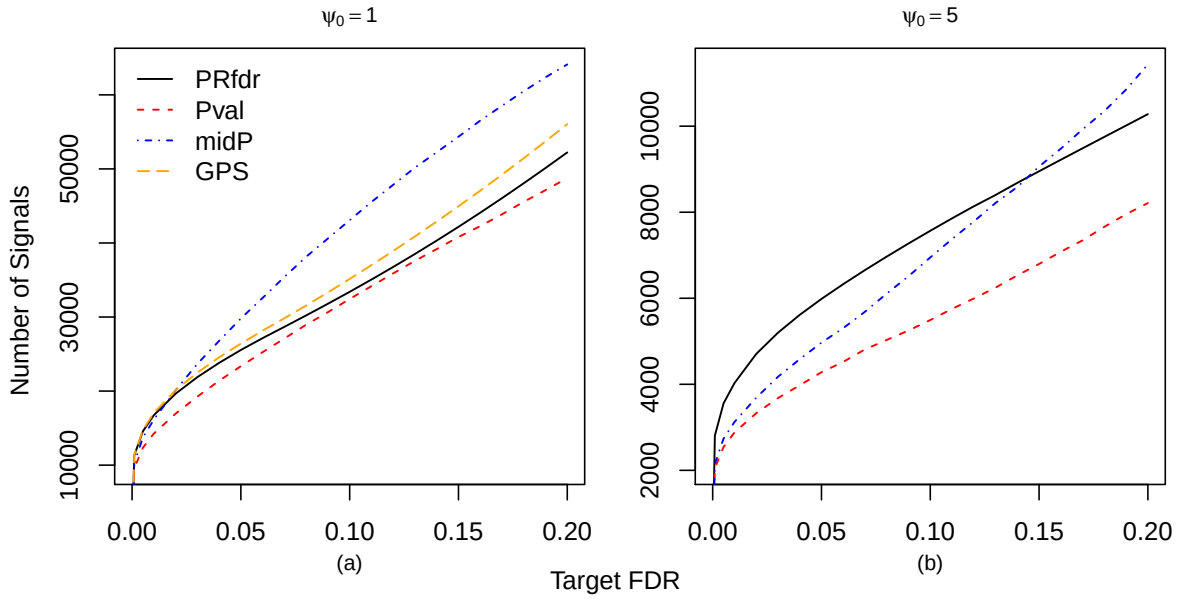

Figure 2.5: Number of signals (y axis) versus target FDR levels (x axis)

The proposed procedure `PRfdr` along with `Pval`, `midP` and `GPS` were applied to the UK pharmocovigilance data. As of September 2014, there were $1,617$ drugs and $1,310$ adverse events in the UK database. The numbers of detected signals are reported for each target FDR level between 0.001 and 0.2. As can be seen in Figure 2.5, for $\psi_0 = 1$, `midP` detected the largest number of signals, `GPS` the second, `PRfdr` the third, and `Pval` the smallest. For $\psi_0 = 5$, the number of signals for `PRfdr` was larger than that for `midP` for small target FDR levels, and `PRfdr` detected fewer signals than `midP` when the target FDR level is greater than about 0.14. `Pval` also detected the fewest signals when $\psi_0 = 5$ is tested.

Both `midP` and `GPS` fail to control the FDR in some simulation settings and are not further compared here. For $\psi_0 = 1$ and $\psi_0 = 5$, our proposed procedure `PRfdr` detected more signals than `Pval`. Among the additional detected signals by `PRfdr` for $\psi_0 = 1$, we find the combination of the pain-killing drug oxycodone and the adverse event nausea or vomiting. The FDR-adjusted p-value by `PRfdr` for the combination is 0.005, while the FDR-adjusted p-value by `Pval` is 0.051. Nausea or vomiting is known as a common side effect of oxycodone, for instance see Portenoy et al. (2007), and this association would be discovered by `PRfdr`, but not by `Pval` at target FDR level 0.05.

2.5 Conclusion

We study the multiple testing problem with composite null hypotheses and discrete data. Pharmacovigilance data is a specific application, and signals can be automatically detected if the association between a combination of drug and adverse event is higher than the tested value.

Ahmed et al. (2010) proposed to adopt mid-p-values on filtered cells, and it is assumed that the distribution of p-values under H_0 is non-decreasing. This assumption can adapt the mixture model of p-values from simple null hypotheses to composite null hypotheses, however it can be violated in some cases and the control of FDR will not be guaranteed. In Ahmed et al. (2009), they extended the GPS model to the multiple testing setting and showed by simulations the extended GPS model outperforms other Bayesian methods. The extended GPS model uses a two-component gamma distribution to approximate the true distribution of relative risk, however in reality the relative risk can follow any distribution. The approximation may not work well in some cases, for example in simulation set 1 where the true relative risk follows a three-component distribution and in set 2 a two-component distribution, and the control of FDR can be liberal or conservative, as shown in the simulation.

In this chapter, we assume no specific distribution on the odds ratio. We propose a non-parametric empirical Bayes method based on the local FDR to approximately control the FDR and achieve the maximum power among all the procedures that can control the FDR. The proposed procedure is free of assumptions on the distribution of true odds ratio.

Further research can be conducted to analyze the goodness of the predictive recursion density estimate. This non-parametric density estimation method exhibits convergence property and efficient computation. However, there are certain cases where the accuracy of density estimation may not be satisfactory, as Padilla et al. (2015) pointed out that the mixing density estimate by predictive recursion can be over-smooth. Moreover, it remains unknown whether the local FDR estimate is consistent, and thus the optimality of the adaptive procedure is not guaranteed.

Chapter 3

Control the False Discovery Rate of GWAS with Adjusted Block P-values

3.1 Introduction

In the last two decades, genome-wide association studies (GWAS) have been widely used to identify the genetic variants that are associated with human diseases or traits. As of September 2018, there are $71,673$ unique variant-trait associations reported from $3,567$ publications in the GWAS catalog (Welter et al., 2013). In spite of the encouraging success of GWAS in studying the genetic basis of human traits, there still exist many challenges. For most complex human traits, the reported variants only explain a small proportion of the trait variation, which is referred to as "missing heritability" (Manolio et al., 2009). For example, there are about 180 single-nucleotide polymorphisms (SNPs) found to be associated with human height (Allen et al., 2010). However, these SNPs can only explain about 10% of the phenotypic variation while the total genetic contribution to human height is estimated to be about 80% (Visscher et al., 2008).

In GWAS, the number of SNPs genotyped can be close to $1,000,000$ or more, while the number of subjects is on the scale of a few thousand. This is known as the "large p, small n" problem, and fitting a multiple regression model with all SNPs is not feasible. The common practice is to test the marginal effect of each SNP separately. Then a stringent statistical threshold of 5×10^{-8} on the marginal p-values is used to control the family-wise error rate (FWER) at level 0.05. The FWER controlling procedure on marginal p-values is simple to apply and can accommodate arbitrary dependence among SNPs. However, such an approach is ill-suited for the GWAS studies of complex diseases or traits for two

reasons. First, in large-scale multiple testing applications such as GWAS, controlling the FWER is a less powerful approach than controlling the false discovery rate (Benjamini and Hochberg, 1995). Second, complex traits are often associated with many SNPs with small or modest effect sizes, and the power to detect causal SNPs (signals) by simple linear regression deteriorates with the increase of the number of signals. As an illustrating example, we randomly simulate 10 to 100 signals with uniformly distributed effect sizes and compute the p-values from marginal linear regression. Details about the simulation of genotype and phenotype data can be found later in Section 3 of our simulation studies. We plot the proportion of signals with p-values no larger than 5×10^{-8} as a function of the number of signals in Figure 3.1. The proportion decreases steadily as the number of signals increases, and only about 30% of the signals have p-values under the threshold when we have 100 signals. To overcome the inefficiency of marginal p-values in GWAS, many researchers propose to use penalized regression, for instance, see Hoggart et al. (2008), Wu et al. (2009), and Hoffman et al. (2013). However, the error rates of the selected variables from penalized regression are unclear.

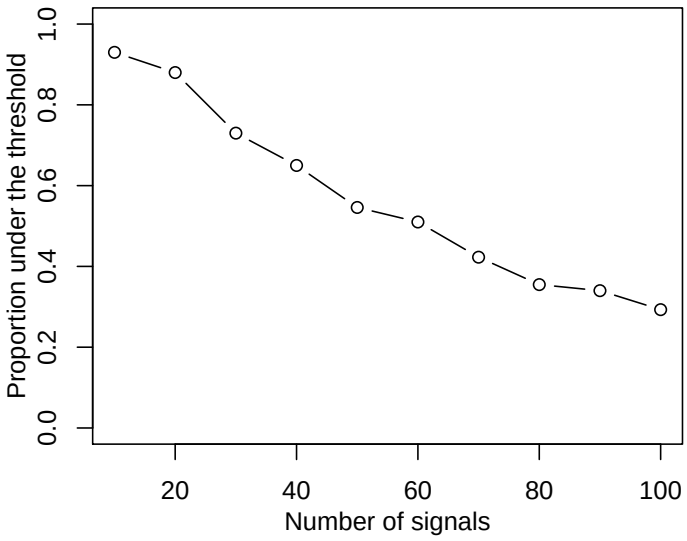

Figure 3.1: Proportion of signals with marginal p-values under the threshold of 5×10^{-8}

We consider the control of the false discovery rate (FDR) in GWAS. Most FDR controlling procedures require independence, but the GWAS test statistics are dependent due to linkage disequilibrium (LD) between SNPs.

As an important feature, GWAS typically only genotype a small subset of all SNPs due to practical and economic reasons, and there is no guarantee that all causal SNPs are genotyped. The dependence between SNPs is utilized to select the genotyped SNPs such that most SNPs are reasonably correlated with a genotyped one. But this feature also makes the definition of true associations on the SNP level difficult. If we use the strict definition of true associations as those between causal SNPs and the phenotype, we will miss many causal SNPs that are not genotyped. On the other hand, if we treat the SNPs that have correlations with the causal SNPs higher than some threshold as having true associations, it is non-trivial to set an appropriate threshold and there could be several correlated SNPs for any causal SNP. Furthermore, no matter whether a causal SNP is genotyped, the neighboring SNPs of the causal SNP tend to have small marginal p-values due to their correlations with the causal SNP, and we call the phenomenon as *signal leakage*. Naive applications of existing FDR procedures are likely to declare these SNPs discoveries, but classifying such discoveries as true associations can distort the true FDR because there can be multiple correlated SNPs for each causal SNP.

In a thought-provoking paper, Brzyski et al. (2017) carefully illustrate many fundamental difficulties of defining true associations and FDR on the SNP level. As many SNPs can be significantly correlated with a causal SNP, they iteratively group the correlated SNPs around the tentative causal SNPs into blocks and define true associations on the block level. More specifically, at each step of their blocking algorithm and among the un-grouped SNPs, they first find the SNP with the smallest marginal p-value as a new block representative, then by a fixed threshold they group all SNPs that are correlated with the representative into the new block. Further tests are carried out on these block representatives. To control the FDR, they use results from selective inference (Benjamini and Bogomolov, 2014) by applying existing testing procedures at a more stringent FDR level. There are two procedures proposed. The first testing procedure applies the linear step-up Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) to the marginal p-values of the block representatives; the second one regresses trait values to block representatives by a special penalized regression method called geneSLOPE. Specifically, they extend SLOPE (Bogdan et al., 2015) to the GWAS setting by modifying the penalty sequence. SLOPE (Bogdan et al., 2015) is an extension of the Lasso (Tibshirani, 1996), and it can achieve the FDR control under orthogonal designs. Brzyski et al. (2017) show approximate FDR control of their testing procedures in simulation studies, and the geneSLOPE procedure performs better than the univariate testing procedure.

However, the procedures in Brzyski et al. (2017) rely on several tuning parameters, and their performance can be sensitive to these parameters, which include the initial screening p-value threshold, the block correlation threshold, and the true positive correlation thresh-

old. To be more specific, we will focus on the last parameter first. In their simulation studies, a block representative is classified as a true positive if the correlation between the representative and any causal SNP is no smaller than a threshold of 0.3. It is clear that the realized FDR and power depend on this threshold. Furthermore, this definition of true positive still allows multiple true associations for each causal SNP.

Similar to Brzyski et al. (2017), we aim to control the FDR on the block level. We first introduce a new definition of the true associations that are one-to-one to the causal SNPs. We obtain a list of tentative causal SNPs through forward selection. We then compute the adjusted block p-value for each block by modeling the effect of signal leakage. Finally, the BH procedure naturally applies to the block p-values and achieves the FDR control.

The chapter is organized as follows. In Section 2, we first introduce the multiple testing problem for GWAS and define the FDR on the block level. Then, we describe our methods to control the FDR. In Section 3, we illustrate the performance of our procedure with a simulation study, and compare our procedure to other candidate methods. In Section 4, we apply our procedure to the North Finland Birth Cohort (NFBC) study (Sabatti et al., 2009) and show additional true discoveries and potential true discoveries. We conclude with some remarks and conclusions in Section 5.

3.2 Methods

In this section, we describe the multiple testing problem for GWAS and define the false discovery rate (FDR; Benjamini and Hochberg 1995) on the block level. Some neighboring single nucleotide polymorphisms (SNPs) exhibit relatively high dependence, and the distribution of test statistics of null hypotheses correlated with non-null hypotheses is distorted due to signal leakage. We derive the effects of causal SNPs on the z-values of correlated non-causal SNPs. Given all the causal SNPs, we propose an oracle procedure to take the effects of signal leakage into consideration. We group the SNPs based on the correlation structure, compute block p-values by adjusting for the effects of causal SNPs outside the blocks, and apply the BH procedure to adjusted block p-values for the FDR control. The locations of causal SNPs are unknown, and we propose an adaptive procedure to find tentative causal SNPs by forward selection and estimate adjusted block p-values.

3.2.1 Multiple Testing Problem for GWAS

We conduct a GWAS study to identify the loci in the genome that are associated with the phenotype of interest. Suppose we have collected the trait values Y and genotype scores X for a collection of M SNPs across the genome from a sample of n individuals. The phenotype of interest can be a continuous trait (e.g., height, blood pressure) or a discrete trait (i.e., disease and no disease). For simplicity, we focus on GWAS studies with continuous traits. Under the commonly used additive model of inheritance, X is the number of minor alleles and takes values in $\{0, 1, 2\}$.

For an unknown set of causal SNPs $\mathcal{C}$, we assume the commonly used linear additive model

$$Y_i = \beta_0 + \sum_{j \in \mathcal{C}} \beta_j W_{ij} + \epsilon_i, \ i = 1, 2, \ \ldots, n, \tag{3.1}$$

where W_{ij} denotes the genotype for the jth causal SNP of the ith subject and $\epsilon_i \sim N(0, \sigma^2)$ is the random error. Because the causal SNPs are not necessarily genotyped, we use W instead of X to emphasize their difference.

Similar to Brzyski et al. (2017), we proceed to test on the block level. The exact blocking procedure used is not crucial, and we use the hierarchical clustering method as in Candes et al. (2018). More specifically, we use the single-linkage clustering with the correlation between genotype scores of SNPs as the similarity measure. The advantage of this blocking procedure is that it only depends on the Pearson correlation between SNPs and is independent of test statistics. Other blocking procedures can also be used, for example, the sequential blocking method of Brzyski et al. (2017) and the interval graph-based method of Kim et al. (2017).

Assume all M SNPs are grouped into m relatively independent blocks, then we define the truly associated blocks as follows. If a causal SNP is genotyped, the block that contains it is truly associated with the phenotype. If a causal SNP is not genotyped, we find the genotyped SNP that correlates the most with the causal one as its representative, then we define the block that contains the representative as truly associated. Finally, we define the rest of the blocks as not associated with the phenotype. Unlike the true association definition in Brzyski et al. (2017), our definition establishes a one-to-one mapping between the truly associated blocks and causal SNPs and does not depend on tuning parameters.

Then the multiple testing problem simplifies to testing whether a block is truly associated with the phenotype or not,

$$H_j : \textit{ block } j \textit{ is not associated with the phenotype}$$

for $j = 1, 2, \ldots, m$. We define the FDR on the block level as $E[V/max\{R, 1\}]$ where V and R are the numbers of falsely and total rejected blocks, respectively.

3.2.2 Adjusted Block p-values

Fitting a multiple regression to all the SNPs is not feasible, as the number of SNPs is often much larger than the number of subjects in GWAS. A marginal linear model

$$Y_i = b_0 + b_j X_{ij} + c_i,$$

can be fitted to test $H_j : b_j = 0$ for SNP $j = 1, 2, \ldots, M$, and we can compute its z-value as

$$z_j = \frac{X_j^T Y}{\sqrt{n-1}\sigma S_j},$$

where S_j is the standard deviation of elements in X_j. Note that in reality, σ is unknown and needs to be estimated.

For a causal SNP c and a non-causal SNP k, it can be shown that the asymptotic joint distribution of z-values is a multivariate normal distribution, i.e.

$$\begin{pmatrix} Z_k \\ Z_c \end{pmatrix} \sim MVN\left(\begin{pmatrix} \frac{\sqrt{n-1}}{\sigma} S_c r_{kc} \beta_c \\ \frac{\sqrt{n-1}}{\sigma} S_c \beta_c \end{pmatrix}, \begin{pmatrix} 1 & r_{kc} \\ r_{kc} & 1 \end{pmatrix}\right), \tag{3.2}$$

where r_{kc} is the Pearson correlation coefficient of X_k and X_c, and S_c is the standard deviation of X_c. The details of the derivation can be found in Appendix. A similar result is also shown in Hormozdiari et al. (2014), and a parallel result for the joint distribution of association statistics in case-control GWAS studies can be found in Han et al. (2009).

Then, conditioning on Z_c, Z_k follows a normal distribution given as follows,

$$Z_k | Z_c \sim N(r_{kc} Z_c, 1 - r_{kc}^2).$$

That is, conditioning on Z_c, the expectation of Z_k is simply the product of Z_c and the Pearson correlation coefficient between them, and the variance is a function of the correlation. The conditional expectation of Z_k implies that simply decreasing the dependence between the SNPs that are tested, for example in Brzyski et al. (2017), may not work well, and the effects of causal SNPs need to be adjusted properly. We can use the conditional distribution to remove the effects of causal SNPs on non-causal ones and calculate adjusted p-values.

Then by applying Equation (3.2) and its conditional form in a multivariate case, for each block we can specify the joint distribution of z-values conditioning on the z-values of causal SNPs outside this block. For any block K, denote the index set of n_K SNPs in the block as $\mathcal{K}$, and we write the conditional distribution as

$$W_{\mathcal{K}|\mathcal{C}} \equiv Z_{\mathcal{K}}|Z_{\mathcal{C}\setminus\mathcal{K}} \sim MVN(\mu_{\mathcal{K}}, \Sigma_{\mathcal{K}}),$$

where $W_{\mathcal{K}|\mathcal{C}} = (W_{\mathcal{K}|\mathcal{C},1}, \ldots, W_{\mathcal{K}|\mathcal{C},n_K})^T$ is defined as the random vector of $Z_{\mathcal{K}}$ conditional on $Z_{\mathcal{C}\setminus\mathcal{K}}$, $\mu_{\mathcal{K}} = (\mu_{\mathcal{K},1}, \ldots, \mu_{\mathcal{K},n_K})^T$ is the conditional mean vector, and $\Sigma_{\mathcal{K}}$ is the conditional variance-covariance matrix. Assume the z-values in block K are $z_{\mathcal{K}} = (z_{\mathcal{K},1}, \ldots, z_{\mathcal{K},n_K})^T$. Denote the element-wise maximum departure of z-values from the conditional mean as $d = \max_{i=1,\ldots,n_K} \mid z_{\mathcal{K},i} - \mu_{\mathcal{K},i} \mid$. Then the adjusted block p-value for block k is

$$\Pr(\max_{i=1,\ldots,n_K} \mid W_{\mathcal{K}|\mathcal{C},i} - \mu_{\mathcal{K},i} \mid \geq d),$$

which can be computed using the multivariate normal distribution function.

We compute adjusted block p-values by removing the effects of causal SNPs on non-causal SNPs. Then, we apply the BH procedure to adjusted block p-values at a nominal level α to control the FDR.

3.2.3 Algorithms to Find Tentative Causal SNPs

To adjust for the signal leakage, we need the knowledge of the causal SNPs and to compute the z-values. However, the locations of causal SNPs are unknown in reality and is the aim of a GWAS study. Moreover, the computation of z-values depends on the unknown σ in Equation (3.1). In Figure 3.1, the detection power of marginal test decreases as the number of signals increases because the unmodeled signals inflate the variance estimate. We propose to estimate the causal SNPs and σ simultaneously through a forward selection algorithm, which sequentially adds tentative causal SNPs until the consecutive reduction

in σ estimates is reasonably small. The details are as follows.

Algorithm 1: Simple forward selection algorithm

1. Calculate p-values for all SNPs across the genome with marginal linear regression, and select the most significant SNP into the set of tentative causal SNPs $\mathcal{P}$;
2. Fit a linear regression to the SNP in $\mathcal{P}$, and denote the estimate of σ as $\hat{\sigma}_{old}$;
3. Compute the p-values for the SNPs outside $\mathcal{P}$ conditioning on the SNPs in $\mathcal{P}$ with multiple linear regression, and select the most significant SNP into $\mathcal{P}$;
4. Re-fit a multiple linear regression to the SNPs in $\mathcal{P}$, and the new estimate of σ is $\hat{\sigma}_{new}$;
5. If $|\hat{\sigma}_{new} - \hat{\sigma}_{old}|/\hat{\sigma}_{old} \leq c$, where c is a threshold, the search stops. Otherwise, set $\hat{\sigma}_{old} = \hat{\sigma}_{new}$ and return to step 3.

We only use the p-values from marginal linear regression in step 1 to select the most significant SNP and then use conditional p-values in the remaining steps. Our goal is to correct for most of the signal leakage and obtain a good estimate of σ. To this end, a relatively stringent threshold of c is preferred to include most of the strong signals or their highly correlated neighbors. By default, we set $c = 0.5\%$.

The simple search algorithm is intuitive but requires intense computation. A GWAS study often contains millions of SNPs, and we need to compute conditional p-values for all the SNPs outside $\mathcal{P}$ at each iteration. For the purpose of computation efficiency, we proceed with a batch mode that only recomputes conditional p-values for the SNPs whose p-values are under a threshold p, for example $p = 0.05$, in each batch, and then repeat the batches until convergence. Moreover, we may first scan all the SNPs and use available variable selection methods, such as Lasso, to select a subset of SNPs as a starting point, then find tentative causal SNPs in the selected SNPs. The batch algorithm is described in

Algorithm 2.

Algorithm 2: Batch search algorithm

1 Use Lasso to select a subset of SNPs $\mathcal{S}$;

2 Initialize the candidate set as $\mathcal{C} = \emptyset$, the set of tentative causal SNPs $\mathcal{P} = \emptyset$, and the old estimate of σ as a reasonably large number, for example $\hat{\sigma}_{old} = 1 \times 10^8$;

3 Start a batch. Set the new batch indicator $\delta = 1$. Compute a p-value for each SNP in $\mathcal{S} \setminus \mathcal{P}$ conditioning the SNPs in $\mathcal{P}$, set the candidate set $\mathcal{C}$ as the set of SNPs that have p-values smaller than p;

4 Compute the p-values for the SNPs in $\mathcal{C}$ conditioning on the SNPs in $\mathcal{P}$. Update the candidate set $\mathcal{C}$, that is, keep only the SNPs with p-values smaller than p in $\mathcal{C}$. Set $\delta = 0$;

5 Select the most significant SNP in $\mathcal{C}$ into $\mathcal{P}$. Fit a multiple linear regression to the SNPs in $\mathcal{P}$, and obtain a new estimate of σ as $\hat{\sigma}_{new}$;

6 Check the convergence criterion.

(a) If $|\hat{\sigma}_{new} - \hat{\sigma}_{old}|/\hat{\sigma}_{old} > c$, keep searching within the current batch by going to step 4;

(b) If $\delta = 1$ and $|\hat{\sigma}_{new} - \hat{\sigma}_{old}|/\hat{\sigma}_{old} \leq c$, the search stops;

(c) If $\delta = 0$ and $|\hat{\sigma}_{new} - \hat{\sigma}_{old}|/\hat{\sigma}_{old} \leq c$, start a new batch by going to step 3;

3.3 Simulations

We conduct simulation studies to investigate the control of the FDR and power of our method.

3.3.1 Candidate Methods

We consider the following candidate methods in the simulation:

`Adaptive`: Our proposed method. We first find tentative causal SNPs with the batch search algorithm 2, estimate σ in Equation (3.1) by fitting a multiple linear model to the tentative causal SNPs, and estimate marginal z-values for all the SNPs. Then we compute the adjusted block p-values conditional on the tentative causal SNPs outside the blocks, and apply the BH procedure at a target FDR level α.

`Oracle`: Same as `Adaptive`, except that we use the true locations of causal SNPs and the true σ.

`GeneSLOPE`: The proposed GeneSLOPE procedure in Brzyski et al. (2017) using our SNP block partitions. As in their paper, we screen out the SNPs that have marginal p-values smaller than 0.05 and apply the GeneSLOPE procedure to the block representatives.

3.3.2 Simulation Setting

We applied HapGen2 (Su et al., 2011) to simulate genotype data across 22 chromosomes and used the European population in the 1000 Genomes project (Consortium et al., 2010) as the reference data. Then, we randomly sampled $M = 10,000$ SNPs for $n = 2,000$ individuals. We used the commonly used additive model (3.1), and computed the genotype scores $X \in \{0, 1, 2\}$ for each SNP as the number of minor alleles.

We randomly selected the set of causal SNPs $\mathcal{C}$ among all the SNPs and varied the number of causal SNPs, which took values $\{10, 20, \ldots, 100\}$. The effect size β was randomly generated and simulated by two distributions: set 1, $Unif(0.6\sqrt{2logM}, 1.4\sqrt{2logM})$ as in Brzyski et al. (2017); set 2, $N(0, 2)$. Random errors ϵ_i's were simulated from $N(0, 1)$. Then, we calculated the trait values Y base on the additive model (3.1).

To facilitate the comparison of candidate methods, we applied the same blocking method as in Candes et al. (2018), i.e., hierarchical clustering with the absolute value of Pearson correlation between genotype scores X as the similarity measure and 0.3 and 0.5 as the single-linkage cutoff ρ. The target FDR level $\alpha = 0.1$.

3.3.3 Simulation Results

We now illustrate the performance of candidate methods with realized FDR levels and the number of true positives. The following results are based on 100 iterations. For $\rho = 0.3$ in Figure 3.2, we can see that the realized FDR levels of `Oracle` and `Adaptive` are close to the target FDR level $\alpha = 0.1$. In set 1 where the effect sizes are uniformly distributed, `GeneSLOPE` has inflated FDR levels, and the largest realized FDR level is about 50% over

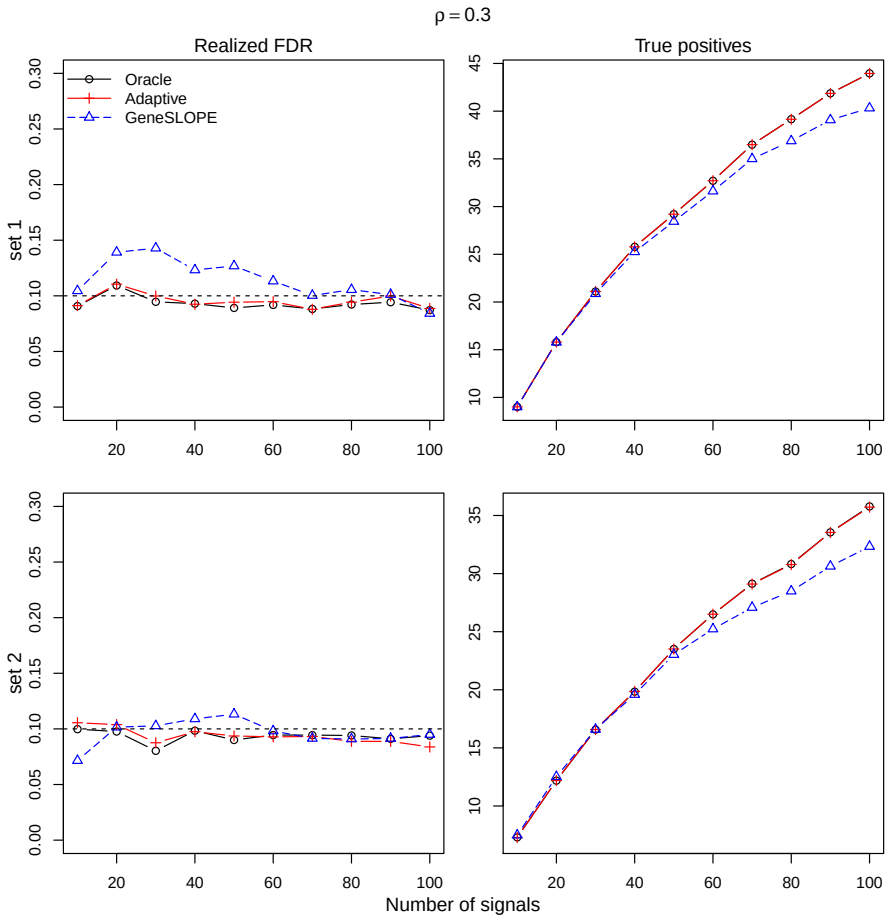

Figure 3.2: Realized FDR levels and true positives for various numbers of signals ($\rho = 0.3$)

the target level. We also see that `Oracle` and `Adaptive` are more powerful than `GeneSLOPE`, especially when the number of signals is large. The power advantage of our procedure over `GeneSLOPE` is more pronounced in set 2 where many effect sizes are small or modest.

The simulation results for $\rho = 0.5$ are shown in Figure 3.3. The realized FDR levels of `Oracle` and `Adaptive` are very close to the target FDR level. On the other hand, the realized FDR levels of `GeneSLOPE` can be highly inflated with a maximum of about 0.31. The inflation of the realized FDR levels is also shown in the simulation results of Brzyski et al. (2017), though to a less extent. The increasing trend of realized FDR may be due to the fact that `GeneSLOPE` fails to account for the dependence between blocks and the effects of signal leakage are more pronounced when the number of signals is large. Our procedure is

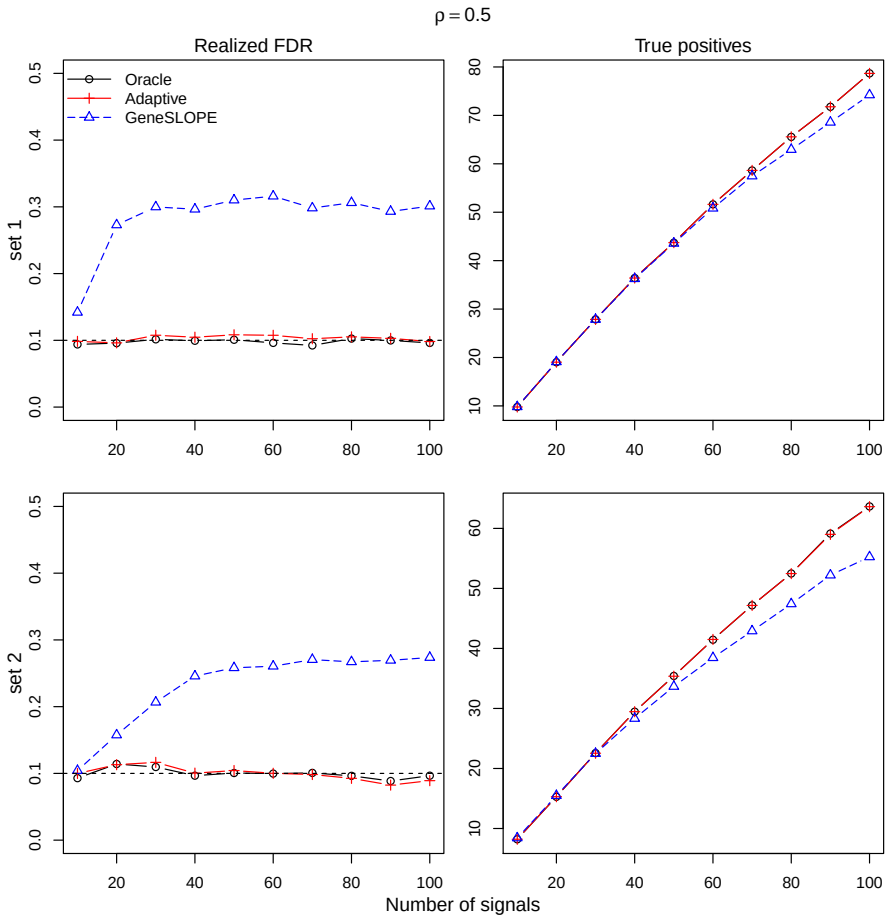

Figure 3.3: Realized FDR levels and true positives for various numbers of signals ($\rho = 0.5$)

more powerful than `GeneSLOPE` even when the realized FDR levels of `GeneSLOPE` is almost 3 times of ours.

In Figure 3.4, we compare the performance of candidate methods with blocking thresholds $\rho \in \{0.3, 0.4, \ldots, 0.7\}$. The number of causal SNPs simulated is set to 100. We see that `Oracle` and `Adaptive` are able to achieve the FDR control at all the levels of blocking thresholds. `GeneSLOPE` has a reasonable FDR control when $\rho = 0.3$ but fails to control the FDR when ρ is larger. The realized FDR of `GeneSLOPE` has a noticeable increasing trend, and the largest realized FDR is above 0.4 when $\rho = 0.7$. When a larger blocking threshold is used, the dependence between blocks gets stronger, thus exacerbating the effects of signal leakage. As for the detection power, we can see that `Adaptive` manages to detect

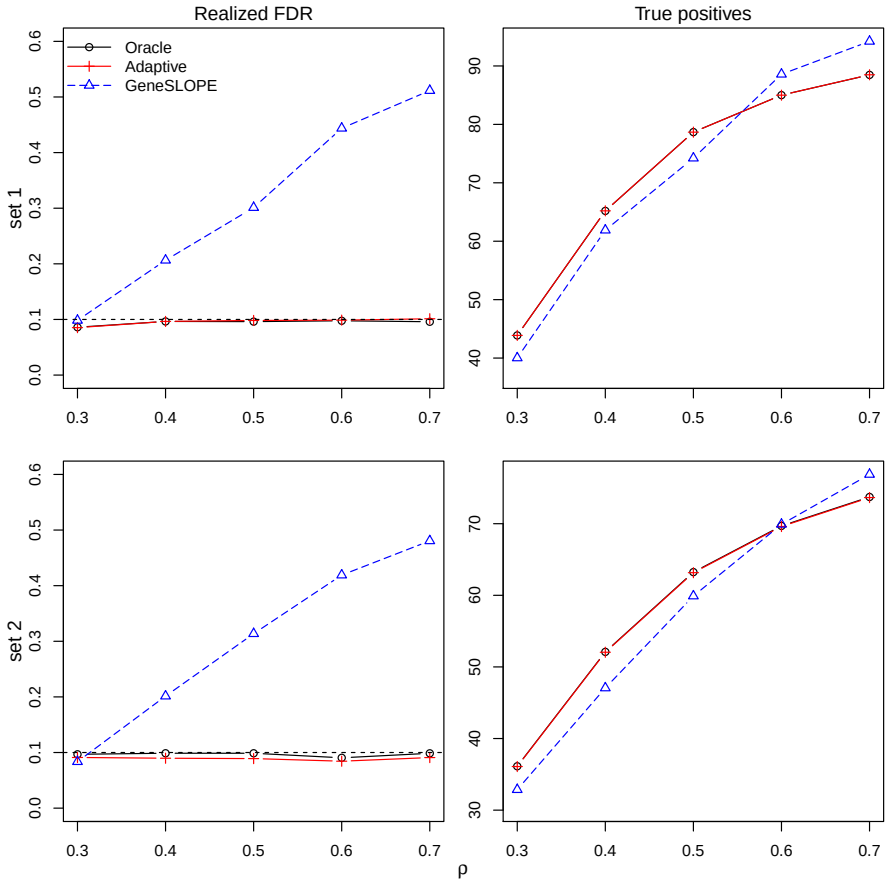

Figure 3.4: Realized FDR levels and true positives for various blocking thresholds ρ

more true positives, even when `GeneSLOPE` has much inflated realized FDR levels. We can also see that the number of true positives increases with the blocking thresholds, and this is because there could be multiple signals residing in the same block when the blocking threshold is small.

We also run simulations with the sequential blocking procedure in Brzyski et al. (2017) and both their and our definitions of truly associated blocks. The results are similar to the results above and are shown in the Appendix.

3.4 Application

We apply our proposed procedure to the North Finland Birth Cohort (NFBC) study (Sabatti et al., 2009). We obtained access to raw phenotype and genotype data from dbGaP with accession number phs000276.v2.p1. The raw data contains the genotype and phenotype data for $364,590$ genetic markers and $5,402$ subjects.

We aim to identify the causal SNPs that are associated with the four lipid phenotypes: high-density lipoproteins (HDL), low-density lipoproteins (LDL), triglycerides (TG), and total cholesterol (CHOL). We applied the proposed procedure `Adaptive` along with `GeneSLOPE` (Brzyski et al., 2017) to the NFBC data and compared the results to the discoveries that are reported by Global Lipids Genetics Consortium (GLGC; Willer et al. 2013), which is a much larger study that contains $188,577$ subjects.

We used the same pre-processing protocol in Sabatti et al. (2009) to filter the subjects. Among the $5,402$ subjects in the study, 487 were excluded from the sample as they were either taking medication for diabetes or not fasted, and the remaining subjects were used for further analysis. Then, we fitted multiple linear regression to the four lipids, with independent variables as gender, pregnant status, oral conception and the first five principle components of genotype scores as population structure variables. We computed the residuals from linear regression and used them as adjusted trait values in the following analysis.

We also conducted quality control procedures in terms of genetic markers. We used the R package "snpStats" (Clayton, 2012) and excluded the markers with call rate $\leq 95\%$, minor allele frequency < 0.01 or Hardy-Weinberg equilibrium p-value < 0.0001. We also imputed missing genotype data for a SNP as the mean of genotype scores for the SNP, as in Brzyski et al. (2017).

We divided the SNPs into blocks with hierarchical clustering and used $\rho = 0.5$ as the single-linkage cutoff. For our proposed procedure `Adaptive`, we treat the SNPs on different chromosomes as independent. We first searched for tentative causal SNPs inside each chromosome and aggregated the results across the 22 chromosomes. Then, we fitted a multiple linear regression to the set of tentative causal SNPs and obtained an estimate of σ. We computed the z-values for all the SNPs and adjusted p-values for blocks, then applied the BH procedure. For `GeneSLOPE`, we selected block representatives as the SNPs with the smallest marginal p-values, and used block representatives.

We compared the rejections by `Adaptive` and `GeneSLOPE` to the discoveries in the GLGC study. Specifically, for each rejected block, we found the smallest marginal p-value reported in GLGC for the SNPs within 1Mb distance from any SNP in the block. We

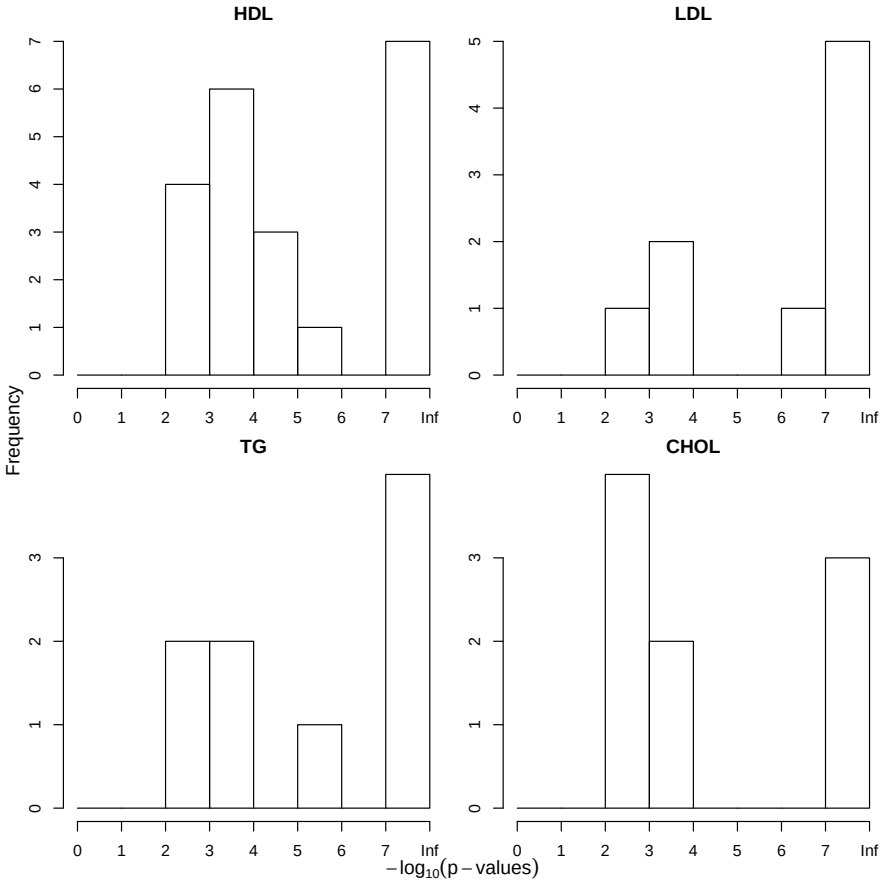

Figure 3.5: Histogram of minimum marginal p-values for the SNPs in GLGC within 1Mb distance from each rejection by `Adaptive`

treated the rejections that have the smallest marginal p-values smaller than 5×10^{-8} as true discoveries, similar to Brzyski et al. (2017). It turned out that our procedure `Adaptive` generated (about) the same number of true discoveries as `GeneSLOPE`. However, `Adaptive` successfully identified more associated blocks than `GeneSLOPE`. In Figure 3.5, we plotted the histogram of minimum p-values for rejections by `Adaptive`. We can clearly see that `Adaptive` identified some additional blocks with modest effect sizes and their minimum marginal p-values are on the magnitude of 10^{-7} to 10^{-2}.

The common practice of GWAS is to compare marginal p-values to the Bonferroni-corrected cutoff 5×10^{-8} and report the SNPs with p-values under the cutoff as discoveries.

There could be two drawbacks in this procedure. The first one is the stringent Bonferroni-corrected cutoff aims to control the family-wise error rate (FWER) at level 0.05, but it is known to be very conservative, especially for complex traits. The second one is that the marginal p-values may not be powerful enough and fail to identify weak or modest associations, and this is due to the fact that marginal regression leads to an inflated estimate of residual variance σ^2 (Hoffman et al., 2013).

The adjusted p-values for additional rejected blocks from `Adaptive` are quite small, for instance, for HDL the block p-values are between 5.19×10^{-6} and 3.66×10^{-5}. We also fitted multiple linear regression to the representative SNPs for rejected blocks from `Adaptive`, and the additional rejections turned out to be very significant. For example, the p-values for HDL are between 1.51×10^{-5} and 2.58×10^{-4}. The additional rejections for the other three phenotypes were also significant. Therefore, `Adaptive` yields about the same number of true discoveries, and the additional rejections can be potential true discoveries.

The adjusted block p-values from blocks with no association would follow closer to uniform distribution after we remove the effects of signal leakage. This allows us to apply the recent development of multiple testing methods to further improve the detection power. For instance, Lei and Fithian (2018) proposed a framework `AdaPT` to use side information in multiple testing problems. We computed the average of minor allele frequency of SNPs in a block as the block minor allele frequency. Then, we applied `AdaPT` to the NFBC data with the adjusted block p-value as test statistics and the block minor allele frequency as side information, and yielded more rejections. For example, `AdaPT` identified 4 more blocks that have minimal marginal p-values within 1Mb distance smaller than 5×10^{-8} in GLGC.

3.5 Conclusion

The use of the FDR in GWAS is still limited compared with its wide applications in other fields, for example expression quantitative trait loci (eQTL) studies. This is mainly due to the fact that neighboring SNPs are often correlated with each other, and the SNPs that are correlated with causal SNPs tend to yield small p-values due to signal leakage. Therefore, one causal SNP may correspond to several rejections of null hypotheses of no association if we naively use test statistics from marginal regression and do not account for the local dependence. The definition of true positive is also troublesome as there may be a few SNPs that are correlated with one causal SNP, and the FDR is distorted as a consequence.

In light of testing groups of null hypotheses in the literature (for example, Benjamini and Heller 2007; Siegmund et al. 2011), Brzyski et al. (2017) creatively tackled this problem

by grouping the SNPs into blocks based on the correlation and defining the FDR on the block level instead of the SNP level. Brzyski et al. (2017) proposed to screen the SNPs and select block representatives, and then apply a testing procedure to the test statistics of selected hypotheses at a more stringent FDR level. However, one notable drawback is that the effects of causal SNPs on correlated neighbors are not accounted for even if the dependence between representative SNPs is greatly decreased due to blocking.

We take on the challenge and quantify the effects of z values of causal SNPs on correlated neighbors. Then, we propose a search algorithm to locate tentative causal SNPs and compute block p-values by removing the effects of tentative causal SNPs outside a block. Hence, the block p-values are uniformly distributed for the blocks with no causal SNPs inside, and we are able to apply testing procedures for the FDR control that are only applicable to conventional settings where null hypotheses are independent. The classical BH procedure is used in our procedure.

Our procedure is able to control the FDR at a nominal level in different settings and yield more detection power than other candidate methods. Moreover, our procedure is quite flexible in the choice of testing procedures for the FDR control, as the effects of signal leakage are already accounted for. We may combine the block p-values with side information such as minor allele frequency and annotation data to further improve the detection power.

Chapter 4

Pleiotropy-Informed Conditional Local False Discovery Rate in GWAS with Shared Control

4.1 Introduction

Since genome-wide association studies (GWAS) were first introduced in 2005, they have become the standard method for identifying associations between single nucleotide polymorphisms (SNPs) and human diseases/traits. As of September 2016, more than $24,000$ associations have been reported from the GWAS catalog (Welter et al., 2013). Although GWAS have been successful in identifying a relatively large number of associations, the detection power is still limited. For complex human traits, the reported associations can only explain a small proportion of heritability, and this phenomenon is referred to as "missing heritability" (Manolio et al., 2009). For example, there are 128 SNPs identified to be associated with schizophrenia from a large study by the Psychiatric Genetics Consortium (PGC), but these SNPs only explain about 3% of the variation while the total variation due to genetic effects is estimated as 80% (Ripke et al., 2014).

In order to address the problem of missing heritability, many efforts have been made to increase the power of GWAS. A simple method is to increase the sample size, but this can be time-consuming and expensive. We can also obtain a larger sample size by a meta-analysis that combines multiple similar studies (Begum et al., 2012). However, a meta-analysis would require data access to other related studies in addition to the coordination between

different study groups. In summary, increasing the sample size directly or through meta-analysis can be economically or administratively challenging, and we focus on developing statistical methods to improve the detection power given the current sample size.

The common practice of GWAS is to compute marginal p-values for all the SNPs across the genome and apply a very stringent Bonferroni-corrected threshold to control the family-wise error rate (FWER). The use of FWER can be very conservative in the setting of large-scale multiple testing problems, such as GWAS. The detection power is further limited when complex phenotypes are of interest. Complex diseases/traits are often associated with many SNPs with small or modest effect sizes.

To explain more phenotypic variation, it would be of great value to integratively analyze GWAS data with the aid of additional information or covariates, for example annotation data and minor allele frequency. A genetically related auxiliary phenotype is of critical importance to identify associated genetic variants for the primary phenotype. Recent studies have shown that some SNPs are associated with multiple genetically related complex traits, and this phenomenon is known as pleiotropy. Pleiotropy exists commonly in genes, for example, among all the SNPs that are associated with some trait, about 16.9% of them are associated with more than one trait (Sivakumaran et al., 2011). These genetically related phenotypes have a similar genetic mechanism and shared associated SNPs. Therefore, we can leverage the pleiotropy to increase the probability of identifying weakly associated SNPs by jointly analyzing genetically related phenotypes.

Some researchers (for example, see Cotsapas et al. 2011; Sklar et al. 2011) have utilized the pleiotropy and reported more associated SNPs, however they still use the overly-stringent Bonferroni-corrected p-value threshold. Andreassen et al. (2013b) take advantage of the pleiotropy between schizophrenia and bipolar disorder and successfully identify more associated SNPs, by extending the false discovery rate (FDR; Benjamini and Hochberg 1995) to the conditional FDR in a two-phenotype framework. The FDR is defined as the expected proportion of false positives among all the rejections, and controlling the FDR is more powerful than controlling the FWER in large-scale multiple testing problems. For all SNPs, they compute and threshold the conditional FDR, which is the posterior probability that a SNP is null for the primary phenotype given that the p-values for both phenotypes are no greater than their respective observed p-values. However, Andreassen et al. (2013b) require that the two GWAS data have distinct case and control samples, which may not be the case for many GWAS studies. It is common for two related GWAS studies to share a part of the control sample as recruiting distinct samples to save cost. The shared control sample will induce a positive correlation between test statistics of two GWAS studies and make the statistical inference more challenging. For example, when we observe small p-values for null SNPs of one GWAS, the corresponding p-values for its related GWAS also

tend to be small. Liley and Wallace (2015) extend the estimation method of the conditional false discovery rate to allow the shared control by explicitly modeling the positive correlation under the null hypothesis. However, one noticeable drawback of the methods in Andreassen et al. (2013b) and Liley and Wallace (2015) is that there is no guarantee of the FDR control by thresholding the conditional FDR. Moreover, the correlation can be complicated as its effect may vary for combinations of null and non-null SNPs for two GWAS. In this chapter, we carefully model the correlation due to the shared control and estimate the conditional local FDR.

This chapter is organized as follows. In Section 2, we propose a method to model the z-values from the primary and auxiliary GWAS and estimate the conditional local FDR. In Section 3, we run the simulation under independent and dependent cases to illustrate the performance of our proposed procedures. In Section 4, we apply our method to the GWAS data of schizophrenia with auxiliary summary statistics of bipolar disorder. Finally, we conclude the chapter with some remarks and discussions in Section 5.

4.2 Methods

In this section, we first propose a method to estimate the conditional local FDR under a bivariate normal mixture model of z-values from the primary and auxiliary genome-wide association studies (GWAS). The z-values from shared control GWAS are not independent of each other, and our model takes into consideration the correlation arising from shared control. We then use an expectation-maximization (EM) algorithm to estimate the parameters in the mixture model. Then, we estimate the conditional local FDR for each hypothesis. It is well known that there may exist high correlations between neighboring SNPs (also known as linkage disequilibrium), and we apply the blocking and adaptive pruning procedures to mitigate the correlation effects.

4.2.1 Probabilistic Model of Test Statistics from GWAS

We conduct a GWAS study to identify the genomic loci that are associated with the phenotype of interest. Suppose we have collected the phenotypes Y and genotype scores X for a collection of m single nucleotide polymorphisms (SNPs) across the genome from a sample of n individuals. The phenotype of interest can be a continuous trait (e.g. height, blood pressure) or a discrete trait (i.e. disease and no disease). In this chapter, we focus on the case-control GWAS study, and the trait values Y take values in $\{0 = no\ disease, 1 =$

disease} as the disease status. The genotype scores X can have different supports in different genetic models. Under the commonly used additive model of inheritance, X is the number of minor alleles and takes values in $\{0, 1, 2\}$. Then, we naturally have the following logistic regression model,

$$log(\frac{P}{1-P}) = \alpha + X^T\beta, \tag{4.1}$$

where $P = \Pr(Y = 1|X)$ is the disease risk, α is the intercept, and β is a coefficient vector of the log odds ratios for SNPs. Other covariates such as population stratification may also have an influence on the disease risk, and they can easily be accommodated by the logistic model.

We are interested in identifying the SNPs that are associated with the phenotype of interest, and testing simultaneously

$$H_i : \beta_i = 0$$

for $i = 1, 2, \ldots, m$.

The number of genotyped SNPs is often much larger than the number of subjects in GWAS, and a joint model of all the SNPs cannot be fitted. Instead, a univariate logistic regression model can be fitted, and the summary test statistics are computed for each SNP. We shall use the z-values as our test statistics in the probabilistic model. Sometimes, z-values are missing in the GWAS results. If the p-value (P) and odds ratio (OR) are given, we can compute the z-value (Z) as

$$Z = sign(1 - OR)\Phi^{-1}(P/2),$$

where *sign* is the sign function and Φ is the cumulative distribution function (CDF) of standard normal distribution.

Note that in GWAS, SNPs are often highly corrected with their neighbors, and this is known as linkage disequilibrium. For simplicity, we first assume that SNPs are independent of each other and will tackle the problem of linkage disequilibrium in Section 4.2.3.

We first start with one GWAS study. From the asymptotic properties of maximum likelihood estimators for logistic regression, we have

$$\hat{\beta}|\beta \sim N(\beta, se(\hat{\beta})^2).$$

Then, the observed z-value for that SNP is computed as

$$Z = \frac{\hat{\beta}}{se(\hat{\beta})},$$

where se is the standard error, and with simple derivation, we have

$$Z|\beta \sim N(\frac{\beta}{se(\hat{\beta})}, 1).$$

We consider the effect size β for a SNP as a random variable and assume β follows a mixture distribution of a point mass at 0 with probability $p = \pi_0$ and a normal distribution with mean 0 and variance σ^2 with probability $p = 1 - \pi_0$, i.e.,

$$\beta \sim \begin{cases} 0, & p = \pi_0; \\ N(0, \sigma^2), & p = 1 - \pi_0. \end{cases}$$

Similar models have been considered in the literature, for example, see Liley and Wallace (2015) and Zhang et al. (2018).

Then, the marginal distribution of z-values is

$$Z \sim \begin{cases} N(0, 1), & p = \pi_0; \\ N(0, 1 + \frac{\sigma^2}{se(\hat{\beta})^2}), & p = 1 - \pi_0. \end{cases}$$

That is, the observed z-values follow a mixture of two normal distributions with mean 0 and different variances. It is known that for the SNPs across the genome, the standard error differs. We can approximate the standard errors for different SNPs with the mean or median of all the standard errors and denote the approximation as σ'. Then, we can write

$$Z \sim \begin{cases} N(0, 1), & p = \pi_0; \\ N(0, 1 + \sigma'^2), & p = 1 - \pi_0. \end{cases}$$

Now, we extend the above probabilistic model to two related GWAS 1 and 2, where GWAS 1 is the primary study and GWAS 2 is the auxiliary one. We aim to improve detection power for primary GWAS with the aid of auxiliary GWAS.

At a SNP, given the effect sizes β_1 and β_2, then the observed z-values jointly follow

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \Bigg| \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \sim MVN \left(\begin{pmatrix} \frac{\beta_1}{se(\hat{\beta}_1)} \\ \frac{\beta_2}{se(\hat{\beta}_2)} \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right),$$

where ρ is the correlation between the z-values. The primary z-values and auxiliary z-values are not independent and the correlation arises from the shared control sample in the two GWAS studies.

Integrating the joint conditional distribution of z-values over the distribution of (β_1, β_2), we can compute the joint marginal distribution of z-values as

$$f(Z_1, Z_2) = \pi_{00} f_{00}(Z_1, Z_2) + \pi_{01} f_{01}(Z_1, Z_2) + \pi_{10} f_{10}(Z_1, Z_2) + \pi_{11} f_{11}(Z_1, Z_2),$$

where π_{00}, π_{01}, π_{10} and π_{11} are the proportions of true status combinations for primary and auxiliary hypotheses H_{00}, H_{01}, H_{10} and H_{11} with 0 indicating null and 1 indicating alternative, and $f_{00}(Z_1, Z_2)$, $f_{01}(Z_1, Z_2)$, $f_{10}(Z_1, Z_2)$ and $f_{11}(Z_1, Z_2)$ defined below are the joint probability distribution functions (PDFs) of (Z_1, Z_2) of the four components,

$$f_{00}(Z_1, Z_2) = MVN\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \sigma_{00}\\ \sigma_{00} & 1\end{pmatrix}\right),$$

$$f_{01}(Z_1, Z_2) = MVN\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \sigma_{01}\\ \sigma_{01} & 1+\sigma_2^2\end{pmatrix}\right),$$

$$f_{10}(Z_1, Z_2) = MVN\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1+\sigma_1^2 & \sigma_{10}\\ \sigma_{10} & 1\end{pmatrix}\right),$$

$$f_{11}(Z_1, Z_2) = MVN\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1+\sigma_1^2 & \sigma_{11}\\ \sigma_{11} & 1+\sigma_2^2\end{pmatrix}\right),$$

where σ_1 and σ_2 are the values of σ' for the primary and auxiliary GWAS, and σ_{00}, σ_{01}, σ_{10} and σ_{11} are the covariances between the primary and auxiliary z-values for the four components. The induced correlation between test statistics under null has been investigated in the literature, and an explicit formula of σ_{00} is readily available (Lin and Sullivan 2009; Zaykin and Kozbur 2010). We need to develop an algorithm to estimate other parameters in the bivariate normal mixture model.

We choose the false discovery rate (FDR; Benjamini and Hochberg 1995) as the error rate. We can define the conditional local false discovery rate (CLfdr; Ferkingstad et al. 2008; Zablocki et al. 2017) as

$$\begin{aligned}\text{CLfdr}(z_1|z_2) &= \Pr(H_i \textit{ is true}|z_1, z_2)\\ &= \frac{\pi_{00} f_{00}(z_1, z_2) + \pi_{01} f_{01}(z_1, z_2)}{f(z_1, z_2)}.\end{aligned}$$

We can easily see that the conditional local FDR computes the posterior probability of primary null hypothesis being true, conditioning on the primary and auxiliary test statistics. The conditional local FDR was proved to be the optimal test statistics (Du et al., 2014) to form the oracle rejection region $S_{OR} = \{(z_1, z_2) : \text{CLfdr}(z_1, z_2) \leq C\}$, where $0 < C < 1$ is a

cut-off. That is, among all the rejection regions that can asymptotically control the FDR at some nominal level, the rejection region constructed by thresholding the conditional local FDR has the largest average number of rejections and the highest power.

Given the conditional local FDR for each hypothesis H_i, similar to Cai and Sun (2009), we use the following oracle procedure to control the FDR at level α,

1. Order all the hypothesis pairs by CLfdr and denote the increasing CLfdr values by $\text{CLfdr}_{(1)}, \ldots, \text{CLfdr}_{(m)}$;

2. Reject the k hypotheses with the smallest CLfdr values, where $k = max\{j : \frac{\sum_{i=1}^{j} \text{CLfdr}_{(i)}}{j} \leq \alpha\}$.

The above oracle procedure is formed by thresholding the optimal test statistics. It is the optimal procedure for the multiple testing problem with auxiliary statistics available.

4.2.2 An Expectation-Maximization Algorithm

The parameters in the four-component bivariate normal mixture models can be estimated by an expectation-maximization (EM) algorithm. The true statuses of primary and auxiliary hypotheses are unknown and latent variables. For simplicity of notations, we denote the vector of z-values as $\boldsymbol{Z}_i = (Z_{1i}, Z_{2i})$, the vector of proportion parameters as $\boldsymbol{\pi} = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$ and the vector of variance/covariance parameters as $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_{00}, \sigma_{01}, \sigma_{10}, \sigma_{11})$, the full data likelihood is given as follows,

$$L(\boldsymbol{\pi}, \boldsymbol{\sigma}) = \prod_{i=1}^{m} (\pi_{00} f_{00}(\boldsymbol{z}_i))^{H_{00i}} (\pi_{01} f_{01}(\boldsymbol{z}_i))^{H_{01i}} (\pi_{10} f_{10}(\boldsymbol{z}_i))^{H_{10i}} (\pi_{11} f_{11}(\boldsymbol{z}_i))^{H_{11i}},$$

and the log-likelihood is

$$l(\boldsymbol{\pi}, \boldsymbol{\sigma}) = \sum_{i=1}^{m} \pi_{00} f_{00}(\boldsymbol{z}_i) H_{00i} + \pi_{01} f_{01}(\boldsymbol{z}_i) H_{01i} + \pi_{10} f_{10}(\boldsymbol{z}_i) H_{10i} + \pi_{11} f_{11}(\boldsymbol{z}_i) H_{11i}.$$

The true statuses of hypotheses H_{00}, H_{01}, H_{10} and H_{11} are missing. We use an expectation-maximization (EM) algorithm to estimate the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\sigma}$. Details are given in Appendix.

After we have estimated the parameters in the model, we can further estimate the conditional local FDR as

$$\widehat{\text{CLfdr}}(Z_1, Z_2) = \frac{\hat{\pi}_{00}\hat{f}_{00}(Z_1, Z_2) + \hat{\pi}_{01}\hat{f}_{01}(Z_1, Z_2)}{\hat{f}(Z_1, Z_2)}.$$

In the adaptive procedure, we replace the conditional local FDR in the oracle procedure with its estimates.

4.2.3 Linkage Disequilibrium and Signal Leakage

Most FDR controlling procedures require independent test statistics, and our EM algorithm to estimate the parameters is designed to maximize the likelihood of independent observations. However, the linkage disequilibrium between SNPs implies that there exists strong dependence within neighboring SNPs in GWAS. Furthermore, the dependence can lead to the phenomenon of *signal leakage* that the test statistics of non-causal SNPs who are correlated with a causal SNP may show strong evidence of significance.

To address the issue of dependence, we take advantage of the available genotype and phenotype data for primary GWAS. We first divide the SNPs into blocks with hierarchical clustering as in Candes et al. (2018) with the Pearson correlation between genotype scores of SNPs as the similarity measure and the threshold $\rho = 0.5$ is used as the single-linkage cutoff. The multiple testing problem simplifies to testing whether a block contains any causal SNPs or not. Then, we select the SNP within a block that has the smallest marginal p-value as the block representative, and the z-values for block representatives are used as block test statistics. However, this selection step distorts the distribution of test statistics as it picks the most significant SNP in each block. We use a similar idea as in Candes et al. (2018) to split the sample into two subsets, say, 20% and 80%, respectively, and we use the 20% subset to select block representatives while using the other 80% to compute the test statistics.

The aforementioned procedure can greatly decrease the dependence between test statistics by testing relatively independent block representatives. However, testing block representatives may not be enough to mitigate the effects of signal leakage, especially when the effect size is large. Consider a causal SNP c and a non-causal SNP k, it has been shown that the expectation of non-causal z-value z_k is proportional to the product of causal z-value z_c and their correlation of genotype scores r_{kc} (Han et al. 2009; Hormozdiari et al.

2014). We propose an adaptive pruning procedure to further alleviate the effects of signal leakage,

Algorithm 3: Adaptive pruning procedure

1. Order all blocks by the absolute value of z-values of their representatives in a decreasing order, denoted as $z_1,\ z_2, \ldots,\ z_k$, where k is the number of blocks;
2. Start with block $i = 1$, and prune blocks $j > i$ that satisfies $|z_i r_{ij}| > c$, where r_{ij} is the Pearson correlation between the genotype scores of representatives of blocks i and j, and c is a cutoff;
3. Iterate i in an increasing order through the remaining blocks and repeat step 2 until no blocks can be pruned.

We can choose the pruning cutoff c based on the knowledge of how much mean shift is significant enough for a standard normal distribution, and we set $c = 0.2$ in this chapter. The conventional pruning only tests whether the correlation r between two SNPs is greater than some threshold and then keeps the SNP with higher MAF. Our adaptive pruning procedure performs better as it takes into account the effect size by testing $|rz|$ and has a higher probability to keep causal SNPs with the knowledge of marginal z-values and the aid of sample splitting technique.

After the blocking and pruning procedures based on the 20% sample of primary GWAS, we have a list of remaining block representatives and can use the other 80% sample to compute test statistics for primary GWAS. The test statistics of remaining block representatives for the auxiliary GWAS can be obtained from its summary statistics. Then we apply the adaptive FDR control procedure to the z-value pairs of remaining block representatives.

4.3 Simulations

In this section, we describe candidate methods and the simulation setting in both the independent and dependent cases, and present the simulation results to illustrate the FDR control and power of our proposed procedure.

4.3.1 Candidate Methods

We consider the following candidate methods in the independent case:

`Oracle`: The proportion parameters π_{00}, π_{01}, π_{10} and π_{11} are known from the number of causal SNPs and shared causal SNPs. The variance parameters σ_1 and σ_2 can be empirically estimated by the sample variance, and the covariance parameters σ_{00}, σ_{01}, σ_{10} and σ_{11} are estimated by the sample covariances. We then compute the conditional local FDR, and apply the oracle procedure.

`Adaptive`: Same as `Oracle`, except that we use the EM algorithm to estimate the parameters in the bivariate normal mixture model and then use the plug-in estimates of conditional local FDR.

`BH`: The linear step-up Benjamini-Hochberg (BH) procedure in Benjamini and Hochberg (1995) is applied to the p-values from the primary GWAS. It is used as a comparison to our procedures.

In the dependent case, we perform the blocking and adaptive pruning procedures in Section 4.2.3, and apply `Oracle`, `Adaptive` and `BH` to the test statistics of remaining block representatives.

4.3.2 Simulation Setting

We conduct the simulation study to examine the performance of our proposed procedure in both the independent and dependent cases. The logistic regression model in Equation (4.1) is used to generate the case-control GWAS data. We use the additive model of inheritance, and the genotype scores take values in $\{0, 1, 2\}$ to represent the number of minor alleles.

In the independent case, we generate independent genotype data. We first generate the minor allele frequencies (MAF) for m independent single nucleotide polymorphisms (SNPs),

$$MAF_i \sim Uniform(0.05, 0.5),$$

for $i = 1, 2, \ldots, m$. Then we generate the genotype scores for N individuals,

$$X_{ji} \sim Binomial(2, MAF_i),$$

for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, N$.

In the dependent case, we simulate genotype data with HapGen2 (Su et al., 2011) across 22 chromosomes with the European population in the 1000 Genomes project (Consortium et al., 2010) as reference data.

Next, we randomly select causal SNPs from the m SNPs for the primary disease and the auxiliary one. For simplicity, the numbers of causal SNPs for two diseases are assumed

to be equal, denoted by m_1. The level of pleiotropy is characterized by the proportion of shared causal SNPs in the m_1 causal SNPs. The effect sizes for causal SNPs of primary and auxiliary diseases are simulated as

$$\beta_1 \sim N(0, \sigma_1^2),$$
$$\beta_2 \sim N(0, \sigma_2^2),$$

where σ_1^2 and σ_2^2 are the variances of effect sizes for causal SNPs. The effect sizes for non-causal SNPs are 0.

The intercept α is chosen to indicate the level of prevalence of the disease. Then, for individual $j = 1, 2, \ldots, N$, we compute the probabilities P_{j1} and P_{j2} for having the primary and auxiliary diseases, respectively, by the logistic model in Equation (4.1). The disease status Y is generated as,

$$Y_{j1} \sim Bernoulli(P_{j1}),$$
$$Y_{j2} \sim Bernoulli(P_{j2}).$$

Again for simplicity, we assume the two GWAS studies have the same sample size n, and each GWAS has equal numbers of control and case subjects $n/2$. Specifically, we sample $n/2$ control and $n/2$ case subjects based on the disease status of the N individuals. The level of shared control is characterized by the proportion of shared control subjects in the n sample. We make sure no shared case subjects between the two GWAS is included in the sample, as it is not of interest in this project and may be rare in real applications.

Now, we have generated the genotype data in both independent and dependent cases, computed the disease status for the primary and auxiliary diseases, and obtained a sample of case and control subjects that features some level of shared control. We compute the z-values by fitting univariate logistic regression to each SNP, and the bivariate z-value pairs (z_{i1}, z_{i2}) are obtained, for $i = 1, 2, \ldots, m$. Then, we apply our proposed procedures to test their performance.

We set the number of SNPs $m = 10000$ and generate $m_1 = 100$ causal SNPs. We set the total sample size $n = 2000$, with 1000 case subjects and 1000 control subjects. The effect sizes of the casual SNPs for the primary and auxiliary diseases are simulated as

$$\beta_1 \sim N(0, 3^2),$$
$$\beta_2 \sim N(0, 2^2).$$

We set various levels of pleiotropy from 0 to 0.9 with an even spacing of 0.1, and use three levels of shared control 10%, 50% and 100%. The target FDR level is set as $\alpha = 0.1$. We compute the average of realized FDR level and the number of true positives over 200 replications.

4.3.3 Simulation Results

Simulation Results for the Independent Case

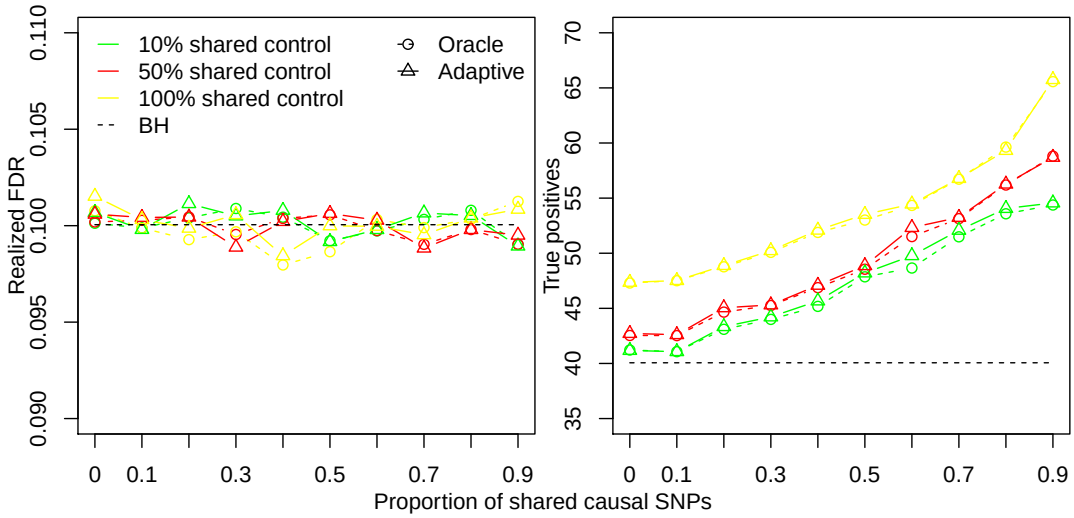

Figure 4.1: Realized FDR levels and true positives for various levels of shared control and pleiotropy (independent case)

Figure 4.1 shows the realized FDR levels (left panel) and the number of true positives (right panel) for various settings of shared control and pleiotropy in the independent case. From the left panel, we can see that `Oracle`, `Adaptive` and `BH` control the FDR around the target level, and the realized FDR levels are approximately in the range $(0.098, 0.102)$. Our procedures control the FDR level reasonably well. The right panel illustrates the detection power for the multiple testing problem and it depicts the average number of true positives for the procedures at different settings of shared control and shared causal SNPs. `BH` yields about 40 true positives, and our procedures achieve more power, especially when the level of shared control or the level of pleiotropy is higher. `Adaptive` generates a similar number of true positives as `Oracle`, and that means `Adaptive` approximates `Oracle` well. It can also be seen that our procedures generate more true positives as the level of shared control or pleiotropy increases. The auxiliary GWAS study can provide more additional information about the primary GWAS when they share more control samples or more causal SNPs. Hence, more true positives can be detected with the assistance of the auxiliary GWAS.

Simulation Results for the Dependent Case

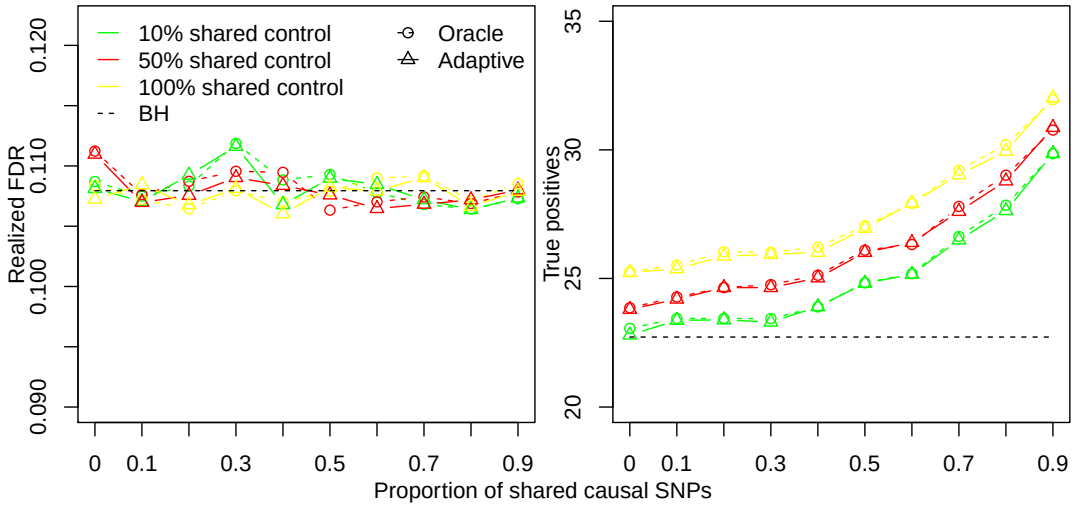

Figure 4.2: Realized FDR levels and true positives for various levels of shared control and pleiotropy (dependent case)

Figure 4.2 depicts the simulation results in the dependent case. We can see that the realized FDR levels of our procedures are roughly in the range $(0.107, 0.113)$ and the FDR is still reasonably controlled. The number of true positives shows a similar increasing trend as the level of shared control or pleiotropy increases. It can be seen that the procedures detect fewer true positives in the dependent case than in the independent case, and that is due to the fact that we performed blocking and pruning to reduce the dependence between hypotheses.

From the simulation results shown above, we can see that in both the independent and dependent cases our proposed procedure can control the realized FDR reasonably well at a nominal level, and yield more detection power than `BH`, especially when the levels of shared control or pleiotropy is high.

4.4 Application

In this section, we apply the proposed procedure to two related psychiatric diseases with schizophrenia as our primary GWAS and bipolar disorder as the auxiliary one. We

obtain access to the raw data of schizophrenia from dbGaP under accession number phs000021.v3.p2 and download the summary statistics for bipolar disorder from Psychiatric Genomics Consortium (PGC; https://www.med.unc.edu/pgc).

For the schizophrenia data, we use $1,404$ cases and $1,442$ controls from the European-American ancestry. The quality of phenotype data is guaranteed by a set of data collection methods and procedures. Case individuals are diagnosed by operational criteria, and control individuals are sampled from the population that is geographically and ethnically similar to cases. There are $702,603$ SNPs genotyped in the data, and common quality control procedures for genotype data are applied to the SNPs. See more details about quality control in Ripke et al. (2011).

We include the first five principal components as covariates in the logistic regression to account for the population structure, and compute the z-values for all SNPs. We apply the blocking and adaptive pruning procedures described in Section 4.2.3.

Due to the limited sample size, we are not able to set the target FDR at a reasonably low level. Instead we compare the goodness of test statistics $\widehat{\text{CLfdr}}$ and p-values by a receiver operating characteristic (ROC) curve in Figure 4.3. We treat the representative SNPs that fall in the 108 associated loci reported in Ripke et al. (2014) as associated SNPs and treat other SNPs as non-associated. We can clearly see that $\widehat{\text{CLfdr}}$ is a better ranking test statistic as its true positive rates are larger than those of p-values for the same false positive rates. Our procedure is likely to improve the detection power at a reasonable FDR level with a larger sample size.

4.5 Conclusion

It is common in GWAS that genetically related traits can share associated genetic variants, for instance see Sivakumaran et al. (2011) and Chambers et al. (2011) for reported genetic pleiotropy in human traits. Schizophrenia and bipolar disorder have some similar psychiatric symptoms (Craddock and Owen 2007; Vieta and Phillips 2007; Fischer and Carpenter Jr 2009), and the literature shows that they share some associated SNPs, for example, see Lichtenstein et al. (2009) and Consortium et al. (2009).

Several existing methods leverage on the genetic pleiotropy between related traits, however they mostly use the over-conservative family-wise error rate (FWER). The false discovery rate (FDR; Benjamini and Hochberg 1995) is a more powerful method for error rate control, and Andreassen et al. (2013b) creatively propose the conditional FDR as an advantageous way to consider the pleiotropy. Liley and Wallace (2015) extend the

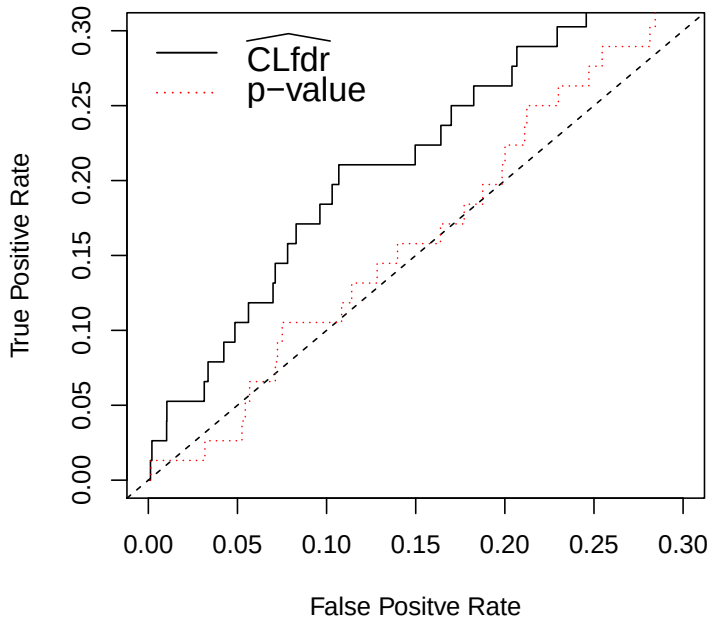

Figure 4.3: True positive rate (y axis) versus false positive rate (x axis) for $\widehat{\text{CLfdr}}$ and p-values

conditional FDR to allow for shared control in the sample, however, the FDR cannot be controlled properly by thresholding the conditional FDR.

In this chapter, we propose a bivariate normal mixture model for the z-values of primary and auxiliary GWAS and estimate the conditional local FDR. To tackle the problem of linkage disequilibrium and signal leakage, we apply the blocking procedure and the adaptive pruning procedure to select the representative SNPs that are relatively independent. Through simulation results, we show that the proposed procedure is able to reasonably control the FDR and yields more detection power than the Benjamini-Hochberg (BH) procedure. We also apply both procedures to schizophrenia and bipolar disorder, and show that $\widehat{\text{CLfdr}}$ is a better ranking test statistic and our procedure is likely to improve the detection power.

Chapter 5

Conclusions and Future Work

In this chapter, we briefly describe conclusions and possible future work.

In Chapter 2, we propose a non-parametric empirical Bayes method for the multiple testing problem with composite null hypotheses and discrete data, with a specific application in pharmacovigilance databases, and the proposed method outperforms other candidate methods. Our procedure estimates the local FDR via a non-parametric density estimate of odds ratio, and the simulation and application results show that it can control the FDR and achieve greater power than other methods that can control the FDR. However, the consistency of the estimate of the local FDR is not clear, and we need future work on the derivation.

In Chapter 3, the aim is to identify associated SNPs for complex continuous traits while controlling the FDR. The neighboring SNPs are often highly correlated, and the SNPs correlated to causal SNPs also have relatively small marginal p-values due to signal leakage. We quantify the effects of signal leakage by deriving the distribution of z-values of non-causal SNPs conditional on correlated causal SNPs, and compute adjusted p-values by removing the effects. The adjusted p-values for the null blocks with no causal SNPs within would follow uniform distribution, and conventional FDR controlling procedures such as the BH procedure would apply. It is of interest to extend the idea of accounting for signal leakage to case-control GWAS studies.

In Chapter 4, we aim to improve the detection power by integratively analyzing GWAS with the aid of additional information. Specifically, we assess the pleiotropy between related case-control GWAS studies with shared controls and shared risk variants, and propose a four-component normal mixture model for the z-value pairs of primary and auxiliary GWAS studies. We also suggest blocking and adaptive pruning procedures to mitigate the effects

of linkage disequilibrium and signal leakage. The pruning procedure would not be necessary if the effects of signal leakage could be accounted for properly as in Chapter 3, which is possible future work. Apart from a related study, annotation data is also advantageous to improve the detection power. It is believed that the SNPs that are functionally annotated perform a more important role and they are more likely to be associated SNPs, than those that are not annotated. For example, it is found that the annotated SNPs can explain more phenotypic variation in human height than those without known function in Yang et al. (2011). Integratively analyze GWAS with other additional information, for example annotation data, would also be useful to improve the power.

References

Agresti, A. (2002). Categorical data analysis. *A John Wiley and Sons, Inc. Publication, Hoboken, New Jersey, USA*.

Agresti, A. and Kateri, M. (2011). *Categorical data analysis*. Springer.

Ahmed, I., Dalmasso, C., Haramburu, F., Thiessard, F., Broët, P., and Tubert-Bitter, P. (2010). False discovery rate estimation for frequentist pharmacovigilance signal detection methods. *Biometrics*, 66(1):301–309.

Ahmed, I., Haramburu, F., Fourrier-Réglat, A., Thiessard, F., Kreft-Jais, C., Miremont-Salamé, G., Bégaud, B., and Tubert-Bitter, P. (2009). Bayesian pharmacovigilance signal detection methods revisited in a multiple comparison setting. *Statistics in Medicine*, 28(13):1774–1792.

Allen, H. L., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., Willer, C. J., Jackson, A. U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832.

Almenoff, J., Tonning, J. M., Gould, A. L., Szarfman, A., Hauben, M., Ouellet-Hellstrom, R., Ball, R., Hornbuckle, K., Walsh, L., Yee, C., et al. (2005). Perspectives on the use of data mining in pharmacovigilance. *Drug Safety*, 28(11):981–1007.

Andreassen, O. A., Djurovic, S., Thompson, W. K., Schork, A. J., Kendler, K. S., O'Donovan, M. C., Rujescu, D., Werge, T., van de Bunt, M., Morris, A. P., et al. (2013a). Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. *The American Journal of Human Genetics*, 92(2):197–209.

Andreassen, O. A., Thompson, W. K., Schork, A. J., Ripke, S., Mattingsdal, M., Kelsoe, J. R., Kendler, K. S., O'Donovan, M. C., Rujescu, D., Werge, T., et al. (2013b). Improved

detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS genetics*, 9(4):e1003455.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386.

Bate, A., Lindquist, M., Edwards, I., Olsson, S., Orre, R., Lansner, A., and De Freitas, R. M. (1998). A Bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology*, 54(4):315–321.

Begum, F., Ghosh, D., Tseng, G. C., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Research*, 40(9):3777–3784.

Benjamini, Y. and Bogomolov, M. (2014). Selective inference on multiple families of hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):297–318.

Benjamini, Y. and Heller, R. (2007). False discovery rates for spatial signals. *Journal of the American Statistical Association*, 102(480):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)*, 57(1):289–300.

Benjamini, Y., Yekutieli, D., et al. (2001). The control of the false discovery rate in multiple testing under dependency. *The annals of statistics*, 29(4):1165–1188.

Besag, J. and Clifford, P. (1991). Sequential monte carlo p-values. *Biometrika*, 78(2):301–304.

Bogdan, M., Van Den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015). Slope—adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3):1103.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62.

Brzyski, D., Peterson, C. B., Sobczyk, P., Candès, E. J., Bogdan, M., and Sabatti, C. (2017). Controlling the rate of GWAS false discoveries. *Genetics*, 205(1):61–75.

Cai, T. T. and Sun, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *Journal of the American Statistical Association*, 104(488).

Candes, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold: model-free knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Cantor, R. M., Lange, K., and Sinsheimer, J. S. (2010). Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *The American Journal of Human Genetics*, 86(1):6–22.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Chambers, J. C., Zhang, W., Sehmi, J., Li, X., Wass, M. N., Van der Harst, P., Holm, H., Sanna, S., Kavousi, M., Baumeister, S. E., et al. (2011). Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nature genetics*, 43(11):1131.

Chen, X., Doerge, R. W., and Heyse, J. F. (2018). Multiple testing with discrete data: Proportion of true null hypotheses and two adaptive FDR procedures. *Biometrical Journal*, 60(4):761–779.

Clayton, D. (2012). snpStats: SnpMatrix and XSnpMatrix classes and methods. *R package*.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, 10(4):417–451.

Consortium, . G. P. et al. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.

Consortium, I. S. et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748.

Cotsapas, C., Voight, B. F., Rossin, E., Lage, K., Neale, B. M., Wallace, C., Abecasis, G. R., Barrett, J. C., Behrens, T., Cho, J., et al. (2011). Pervasive sharing of genetic effects in autoimmune disease. *PLoS genetics*, 7(8):e1002254.

Craddock, N. and Owen, M. J. (2007). Rethinking psychosis: the disadvantages of a dichotomous classification now outweigh the advantages. *World Psychiatry*, 6(2):84.

Devlin, B. and Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29(2):311–322.

Döhler, S., Durand, G., Roquain, E., et al. (2018). New FDR bounds for discrete and heterogeneous tests. *Electronic Journal of Statistics*, 12(1):1867–1900.

Du, L., Zhang, C., et al. (2014). Single-index modulated multiple testing. *The Annals of Statistics*, 42(4):1262–1311.

DuMouchel, W. (1999). Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *The American Statistician*, 53(3):177–190.

DuMouchel, W. and Pregibon, D. (2001). Empirical bayes screening for multi-item associations. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 67–76. ACM.

Efron, B. (2004a). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104.

Efron, B. (2004b). *Selection and estimation for large-scale simultaneous inference*. Citeseer.

Efron, B. (2010). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Efron, B. (2014). *The Bayes deconvolution problem*.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23(1):70–86.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160.

European Medicines Agency (2012). Guideline on good pharmacovigilance practices (GVP), module VI – management and reporting of adverse reactions to medicinal products. *Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/06/WC500129135.pdf.[Last accessed on 2012 Dec 30]*.

Evans, S., Waller, P. C., and Davis, S. (2001). Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and drug safety*, 10(6):483–486.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Ferkingstad, E., Frigessi, A., Rue, H., Thorleifsson, G., Kong, A., et al. (2008). Unsupervised empirical Bayesian multiple testing with external covariates. *The Annals of Applied Statistics*, 2(2):714–735.

Fischer, B. A. and Carpenter Jr, W. T. (2009). Will the kraepelinian dichotomy survive dsm-v? *Neuropsychopharmacology*, 34(9):2081.

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(3):499–517.

Gilbert, P. B. (2005). A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):143–158.

Griffin, J. E. and Brown, P. J. (2011). Bayesian hyper-lassos with non-convex penalization. *Australian & New Zealand Journal of Statistics*, 53(4):423–442.

Guo, S. W. and Thompson, E. A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics*, pages 361–372.

Gupta, S. K. (2008). Tenofovir-associated fanconi syndrome: review of the FDA adverse event reporting system. *AIDS Patient Care and STDs*, 22(2):99–103.

Gur, R. H. H. (2011). False discovery rate controlling procedures for discrete tests. *arXiv preprint arXiv:1112.4627*.

Habiger, J. D. (2015). Multiple test functions and adjusted p-values for test statistics with discrete distributions. *Journal of Statistical Planning and Inference*, 167:1–13.

Han, B., Kang, H. M., and Eskin, E. (2009). Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genetics*, 5(4):e1000456.

Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Classic papers in genetics. Prentice-Hall, Inc.: Englewood Cliffs, NJ*, pages 60–62.

Heyse, J. F. (2011). A false discovery rate procedure for categorical data. *Resent Advances in Biostatistics: False Discovery Rates, Survival Analysis, and Related Topics*, pages 43–58.

Hill, W. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38(6):226–231.

Hochberg, J. and Tamhane, A. C. (1987). Multiple comparison procedures. Technical report, John Wiley & Sons,.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802.

Hoffman, G. E., Logsdon, B. A., and Mezey, J. G. (2013). PUMA: a unified framework for penalized multiple regression analysis of GWAS data. *PLoS computational biology*, 9(6):e1003101.

Hoggart, C. J., Whittaker, J. C., De Iorio, M., and Balding, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS genetics*, 4(7):e1000130.

Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508.

Kim, S. A., Cho, C.-S., Kim, S.-R., Bull, S. B., and Yoo, Y. J. (2017). A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated SNPs. *Bioinformatics*, 34(3):388–397.

Korte, A. and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant methods*, 9(1):29.

Kulinskaya, E. and Lewin, A. (2009). On fuzzy familywise error rate and false discovery rate procedures for discrete distributions. *Biometrika*, 96(1):201–211.

Lancaster, H. O. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association*, 56(294):223–234.

Lei, L. and Fithian, W. (2018). Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):649–679.

Lewontin, R. (1964). The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics*, 49(1):49.

Liang, K. (2017). Covariate assisted large-scale multiple testing. *Annals of Applied Statistics*, 42(5):409–419.

Liang, K. and Nettleton, D. (2012). Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):163–182.

Lichtenstein, P., Yip, B. H., Björk, C., Pawitan, Y., Cannon, T. D., Sullivan, P. F., and Hultman, C. M. (2009). Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *The Lancet*, 373(9659):234–239.

Liley, J. and Wallace, C. (2015). A pleiotropy-informed Bayesian false discovery rate adapted to a shared control design finds new disease associations from GWAS summary statistics. *PLoS genetics*, 11(2):e1004926.

Lin, D.-Y. and Sullivan, P. F. (2009). Meta-analysis of genome-wide association studies with overlapping subjects. *The American Journal of Human Genetics*, 85(6):862–872.

Lindquist, M. (2008). VigiBase, the WHO global ICSR database system: basic facts. *Drug Information Journal*, 42(5):409–419.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747.

Martin, R. and Tokdar, S. T. (2009). Asymptotic properties of predictive recursion: robustness and rate of convergence. *Electronic Journal of Statistics*, 3:1455–1472.

Martin, R. and Tokdar, S. T. (2012). A nonparametric empirical Bayes framework for large-scale multiple testing. *Biostatistics*, 13(3):427–439.

Miller, R. G. (1981). *Simultaneous statistical inference*. Springer.

Newton, M. A. (2002). On a nonparametric recursive estimator of the mixing distribution. *Sankhyā: The Indian Journal of Statistics, Series A*, 64:306–322.

of the Psychiatric Genomics Consortium, S. W. G. et al. (2018). Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes. *Cell*, 173(7):1705–1715.

Padilla, O. H. M., Polson, N. G., and Scott, J. G. (2015). A deconvolution path for mixtures. *arXiv preprint arXiv:1511.06750*.

Portenoy, R. K., Farrar, J. T., Backonja, M.-M., Cleeland, C. S., Yang, K., Friedman, M., Colucci, S. V., and Richards, P. (2007). Long-term use of controlled-release oxycodone for noncancer pain: results of a 3-year registry study. *The Clinical Journal of Pain*, 23(4):287–299.

Pounds, S. and Cheng, C. (2006). Robust estimation of the false discovery rate. *Bioinformatics*, 22(16):1979–1987.

Privitera, G. J. (2011). *Statistics for the behavioral sciences*. Sage.

Ripke, S., Neale, B. M., Corvin, A., Walters, J. T., Farh, K.-H., Holmans, P. A., Lee, P., Bulik-Sullivan, B., Collier, D. A., Huang, H., et al. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421.

Ripke, S., Sanders, A. R., Kendler, K. S., Levinson, D. F., Sklar, P., Holmans, P. A., Lin, D.-Y., Duan, J., Ophoff, R. A., Andreassen, O. A., et al. (2011). Genome-wide association study identifies five new schizophrenia loci. *Nature genetics*, 43(10):969.

Sabatti, C., Service, S. K., Hartikainen, A.-L., Pouta, A., Ripatti, S., Brodsky, J., Jones, C. G., Zaitlen, N. A., Varilo, T., Kaakinen, M., et al. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics*, 41(1):35.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633.

Siegmund, D., Zhang, N., and Yakir, B. (2011). False discovery rate for scanning statistics. *Biometrika*, 98(4):979–985.

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J. G., Zgaga, L., Manolio, T., Rudan, I., McKeigue, P., Wilson, J. F., and Campbell, H. (2011). Abundant pleiotropy in human complex diseases and traits. *The American Journal of Human Genetics*, 89(5):607–618.

Sklar, P., Ripke, S., Scott, L. J., Andreassen, O. A., Cichon, S., Craddock, N., Edenberg, H. J., Nurnberger Jr, J. I., Rietschel, M., Blackwood, D., et al. (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near odz4. *Nature genetics*, 43(10):977.

Spencer, C. C., Su, Z., Donnelly, P., and Marchini, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics*, 5(5):e1000477.

Srba, J. (2014). Pharmacovigilance: Spontaneous reporting systems.

Stefansson, H., Ophoff, R. A., Steinberg, S., Andreassen, O. A., Cichon, S., Rujescu, D., Werge, T., Pietiläinen, O. P., Mors, O., Mortensen, P. B., et al. (2009). Common variants conferring risk of schizophrenia. *Nature*, 460(7256):744.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, 27(16):2304–2305.

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(479):901–912.

Sun, W. and McLain, A. C. (2012). Multiple testing of composite null hypotheses in heteroscedastic models. *Journal of the American Statistical Association*, 107(498):673–687.

Szarfman, A., Machado, S. G., and O'neill, R. T. (2002). Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the us fda's spontaneous reports database. *Drug Safety*, 25(6):381–392.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tokdar, S. T., Martin, R., and Ghosh, J. K. (2009). Consistency of a recursive estimate of mixing distributions. *The Annals of Statistics*, 37(5A):2502–2522.

van Puijenbroek, E. P., Bate, A., Leufkens, H. G., Lindquist, M., Orre, R., and Egberts, A. C. (2002). A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiology and drug safety*, 11(1):3–10.

Vermeer, N. S., Straus, S. M., Mantel-Teeuwisse, A. K., Domergue, F., Egberts, T. C., Leufkens, H. G., and De Bruin, M. L. (2013). Traceability of biopharmaceuticals in spontaneous reporting systems: a cross-sectional study in the FDA Adverse Event Reporting System (FAERS) and EudraVigilance databases. *Drug Safety*, 36(8):617–625.

Vieta, E. and Phillips, M. L. (2007). Deconstructing bipolar disorder: a critical review of its diagnostic validity and a proposal for dsm-v and icd-11. *Schizophrenia Bulletin*, 33(4):886–892.

Visscher, P. M. (2008). Sizing up human height variation. *Nature Genetics*, 40(5):489.

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of GWAS discovery. *The American Journal of Human Genetics*, 90(1):7–24.

Visscher, P. M., Hill, W. G., and Wray, N. R. (2008). Heritability in the genomics era—concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255.

Weinberg, W. (1908). ber den nachweis der vererbung beim menschen. *Jahres. Wiertt. Ver. Vaterl. Natkd.*, 64:369–382.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2013). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1):D1001–D1006.

WHO et al. (2002). The importance of pharmacovigilance.

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., Mora, S., et al. (2013). Discovery and refinement of loci associated with lipid levels. *Nature genetics*, 45(11):1274.

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

Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., De Andrade, M., Feenstra, B., Feingold, E., Hayes, M. G., et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics*, 43(6):519.

Zablocki, R. W., Levine, R. A., Schork, A. J., Xu, S., Wang, Y., Fan, C. C., Thompson, W. K., et al. (2017). Semiparametric covariate-modulated local false discovery rate for genome-wide association studies. *The Annals of Applied Statistics*, 11(4):2252–2269.

Zaykin, D. V. and Kozbur, D. O. (2010). P-value based analysis for shared controls design in genome-wide association studies. *Genetic epidemiology*, 34(7):725–738.

Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942.

Zhang, Y., Qi, G., Park, J.-H., and Chatterjee, N. (2018). Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nature genetics*, 50(9):1318.

Zhou, H., Sehl, M. E., Sinsheimer, J. S., and Lange, K. (2010). Association screening of common and rare genetic variants by penalized regression. *Bioinformatics*, 26(19):2375.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

APPENDICES

Appendix A

Appendix in Chapter 2

A.1 Proof of Theorem 2

For ease of notation, for any set A, we write,

$$p(A) = \sum_{x \in A} p(x),$$
$$p_0(A) = \sum_{x \in A} p_0(x),$$
$$p_1(A) = p(A) - p_0(A).$$

It is obvious that $\alpha < C$ as $\mathrm{Fdr}(S_{OR})$ is the conditional expectation of $\mathrm{fdr}(x)$ given $x \in S_{OR}$.

Note that Fdr is discrete, and a randomization technique can be applied such that Fdr can reach any level $\alpha \in (0, \pi_0]$. From the rejection region S, we can construct a new rejection region S_1 by adding all the hypotheses with the next greater fdr value than the greatest fdr value in S, and it is straightforward that $\mathrm{Fdr}(S) < \mathrm{Fdr}(S_1)$. We can also add a fraction $\beta \in (0, 1)$ of all the hypotheses with the next greater fdr value, and this rejection region will yield an Fdr level $\alpha \in (\mathrm{Fdr}(S), \mathrm{Fdr}(S_1))$. This construction continues until we obtain a rejection region S' such that $S \subset S'$ and $\mathrm{Fdr}(S') = \alpha$. It remains to prove that $p_1(S) \leq p_1(S_{OR})$.

For every x,

$$I(x \in S')\{1 - \mathrm{fdr}(x)/C\} \leq I(x \in S_{OR})\{1 - \mathrm{fdr}(x)/C\},$$

as the left-hand side is smaller than or equal to 0 and the right-hand side is equal to 0 when $x \notin S_{OR}$, and the left-hand side is obviously smaller than or equal to the right-hand side when $x \in S_{OR}$.

Taking expectation on both sides,

$$\begin{aligned}&\sum_x I(x \in S')\{1 - \text{fdr}(x)/C\}p(x)\\ &\leq \sum_x I(x \in S_{OR})\{1 - \text{fdr}(x)/C\}p(x),\end{aligned}$$

where $p(x)$ is the probability mass function of X.

Note that we can write the left-hand side as

$$\begin{aligned}&\sum_x I(x \in S')\{1 - \text{fdr}(x)/C\}p(x)\\ &= \sum_x I(x \in S')p(x) - \sum_x I(x \in S')\text{fdr}(x)/Cp(x)\\ &= p(S') - \sum_{x \in S'} \text{fdr}(x)p(x)/C\\ &= p(S') - E\{\text{fdr}(x)|x \in S'\}p(S')/C\\ &= p(S')\{1 - \text{Fdr}(S')/C\}.\end{aligned}$$

Similarly, the right-hand side can be simplified.

We can get

$$p(S')\{1 - \text{Fdr}(S')/C\} \leq p(S_{OR})\{1 - \text{Fdr}(S_{OR})/C\}.$$

Note that $\text{Fdr}(S') = \text{Fdr}(S_{OR}) = \alpha$, and $1 - \alpha/C > 0$. Therefore, we can get $p(S') \leq p(S_{OR})$. Moreover, $p_1(S') = p(S')(1 - \alpha)$ and $p_1(S_{OR}) = p(S_{OR})(1 - \alpha)$, and it leads to $p_1(S') \leq p_1(S_{OR})$. We also have $p_1(S) \leq p_1(S')$ as $S \subset S'$, therefore $p_1(S) \leq p_1(S_{OR})$.

A.2 Predictive Recursion

Consider the following mixing density estimation problem, $X_1, \ldots, X_m$ are independently distributed with the probability mass function,

$$p(x) = \int_\Theta f(x|\theta)g(\theta)\mu(d\theta),$$

where x is an observable data point, $f(x|\theta)$ is a known sampling distribution, $g(\theta)$ is the mixing distribution of parameter θ, and μ is a σ-finite measure on Θ. Newton (2002) proposed a recursive algorithm that keeps updating the density estimate $g(\theta)$ of parameter θ.

Predictive Recursion (PR) Algorithm. Start with an initial estimate of g, denoted by g_0, given a sequence of weights $w_1, \ldots, w_m \in (0, 1)$, then for $i = 1, \ldots, m$, the ith density estimate of θ can be obtained as

$$g_i(\theta) = (1 - w_i)g_{i-1}(\theta) + w_i \frac{f(x|\theta)g_{i-1}(\theta)}{\int_\Theta f(x|\theta')g_{i-1}(\theta')\mu(d\theta')},$$

and the ith marginal probability mass function estimate of X is

$$p_i(x) = \int_\Theta f(x|\theta)g_i(\theta)\mu(d\theta).$$

Similar to Martin and Tokdar (2012), we set the weights $w_i = (i + 1)^{-0.67}$ for $i = 1, \ldots, n$.

A.3 p-values based procedure

Here we present the details of the p-value based procedure, which is compared in simulation studies. We estimate the flattened proportion in the center of the distribution of p-values as $\pi_0\pi_{0*}$. Because of the large proportion of empty cells, this estimate can be quite small. So we treat all p-values greater than 0.5 as coming from null hypotheses, and for any p-value threshold γ, we can estimate the FDR as

$$\widehat{\text{FDR}}(\gamma) = \begin{cases} \widehat{\pi_0\pi_{0*}}\gamma/\hat{F}(\gamma) & \text{if } \gamma \leq 0.5; \\ [\widehat{\pi_0\pi_{0*}}0.5 + \#(0.5 < p_{ij} \leq \gamma)/m]/\hat{F}(\gamma) & \text{if } \gamma > 0.5. \end{cases}$$

Appendix B

Appendix in Chapter 3

B.1 Distribution of z-values of causal and non-causal SNPs

First, we consider z-values from marginal linear regression for SNP $j = 1, 2, \ldots, M$. For simplicity, we center X_j and Y to have mean 0 and it will not affect the inference of b_j. The maximum likelihood estimate is

$$\begin{aligned}\hat{b}_j &= Cov(X_j, Y)/Var(X_j) \\ &= \frac{1}{(n-1)S_j^2} X_j^T Y,\end{aligned}$$

where S_j is the standard deviation of X_j.

Under model (3.1), the expectation of $\hat{b}_j$

$$\begin{aligned}E(\hat{b}_j) &= \frac{1}{(n-1)S_j^2} X_j^T E(Y) \\ &= \frac{1}{(n-1)S_j^2} X_j^T (\beta_0 \mathbf{1} + X_c \beta_c) \\ &= \frac{1}{(n-1)S_j^2} \sum_{k \in \mathcal{C}} (n-1) S_j S_k r_{jk} \beta_k \\ &= \frac{1}{S_j} \sum_{k \in \mathcal{C}} S_k r_{jk} \beta_k,\end{aligned}$$

where r_{jk} is the Pearson correlation coefficient between X_j and X_k.

The variance for $\hat{b}_j$ is

$$\begin{aligned} V(\hat{b}_j) &= \frac{\sum_{i=1}^{n} X_{ij}^2 \sigma^2}{(n-1)^2 S_j^4} \\ &= \frac{\sigma^2}{(n-1)S_j^2}. \end{aligned}$$

The covariance between $\hat{b}$ for SNPs j and k is

$$\begin{aligned} Cov(\hat{b}_j, \hat{b}_k) &= \sigma^2 \frac{1}{(n-1)S_j^2} X_j^T X_k \frac{1}{(n-1)S_k^2} \\ &= \frac{r_{jk}\sigma^2}{(n-1)S_j S_k}, \end{aligned}$$

and the correlation is

$$\begin{aligned} Cor(\hat{b}_j, \hat{b}_k) =& \frac{Cov(\hat{b}_j, \hat{b}_k)}{\sqrt{V(\hat{b}_j)V(\hat{b}_k))}} \\ &= r_{jk}. \end{aligned}$$

We can compute the z-value for SNP j,

$$\begin{aligned} Z_j &= \frac{\hat{b}_j}{sd(\hat{b}_j)} \\ &= \frac{X_j^T Y}{(n-1)S_j^2} \Big/ \sqrt{\frac{\sigma^2}{(n-1)S_j^2}} \\ &= \frac{X_j^T Y}{\sqrt{n-1}\sigma S_j}. \end{aligned}$$

The expectation of Z_j is

$$\begin{aligned} E(Z_j) &= \frac{1}{S_j} \sum_{k \in C} S_k r_{jk} \beta_k \Big/ \frac{\sigma}{\sqrt{n-1} S_j} \\ &= \frac{\sqrt{n-1}}{\sigma} \sum_{k \in C} S_k r_{jk} \beta_k, \end{aligned}$$

and the variance is

$$\begin{aligned} V(Z_j) &= \frac{\sum_{i=1}^{n} X_{ij}^2 \sigma^2}{(n-1)\sigma^2 S_j^2} \\ &= 1. \end{aligned}$$

That says, for a causal SNP c and non-causal SNP k, the joint distribution of z-values is

$$\begin{pmatrix} Z_k \\ Z_c \end{pmatrix} \sim MVN \left(\begin{pmatrix} \frac{\sqrt{n-1}}{\sigma} S_c r_{kc} \beta_c \\ \frac{\sqrt{n-1}}{\sigma} S_c \beta_c \end{pmatrix}, \begin{pmatrix} 1 & r_{kc} \\ r_{kc} & 1 \end{pmatrix} \right).$$

Then, the distribution of Z_k conditioning on Z_c is

$$\begin{aligned} Z_k | Z_c &\sim N(\frac{\sqrt{n-1}}{\sigma} S_c r_{kc} \beta_c + r_{kc}(Z_c - \frac{\sqrt{n-1}}{\sigma} S_c \beta_c), 1 - r_{kc}^2) \\ &\sim N(r_{kc} Z_c, 1 - r_{kc}^2). \end{aligned}$$

B.2 Simulation results with Brzyski et al. (2017)'s blocking procedure and different definitions of true null blocks

Brzyski et al. (2017) propose a sequential blocking procedure. Among the un-grouped SNPs, their blocking procedure sequentially finds the SNP with the smallest marginal p-value as a new block representative, and groups all the SNPs that have correlations with the representative greater than some threshold ρ into the new block. Brzyski et al. (2017) define the null hypothesis corresponding to a block to be true if the block representative SNP has correlations smaller than 0.3 with all causal SNPs. There could be several blocks whose representatives have correlations higher than 0.3 with one causal SNP, and those blocks are considered to be non-null corresponding to one causal SNP. This definition of true null hypotheses may distort the FDR and power. Instead, we define the null hypothesis related to a block to be true if the block does not contain any causal SNP.

Here, we present the simulation results using the blocking procedure proposed by Brzyski et al. (2017), and we use both their definition of true null blocks and ours when evaluating the FDR and true positives. The target FDR level is 0.1, and the results are based on 100 iterations.

Figures B.1–B.2 show the realized FDR and true positives for $\rho = 0.3$ and $\rho = 0.5$ when our definition of true null hypotheses is used. We can see that our procedures are able to control the FDR for all the numbers of signals, and `GeneSLOPE` has a liberal FDR control. Our procedures also yield more numbers of true positives than `GeneSLOPE`.

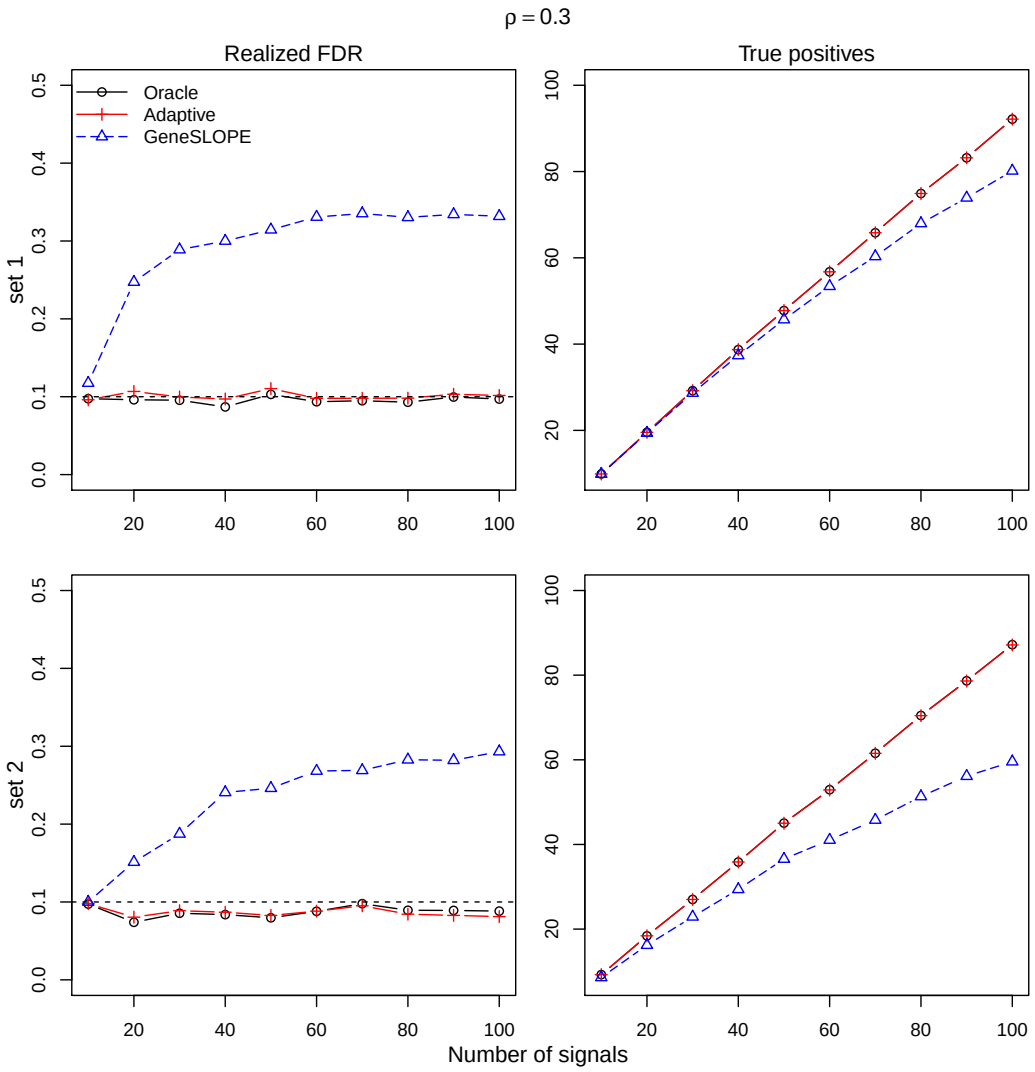

Figure B.1: Realized FDR levels and true positives for various numbers of signals ($\rho = 0.3$ and true null hypotheses as the blocks that contain any causal SNP)

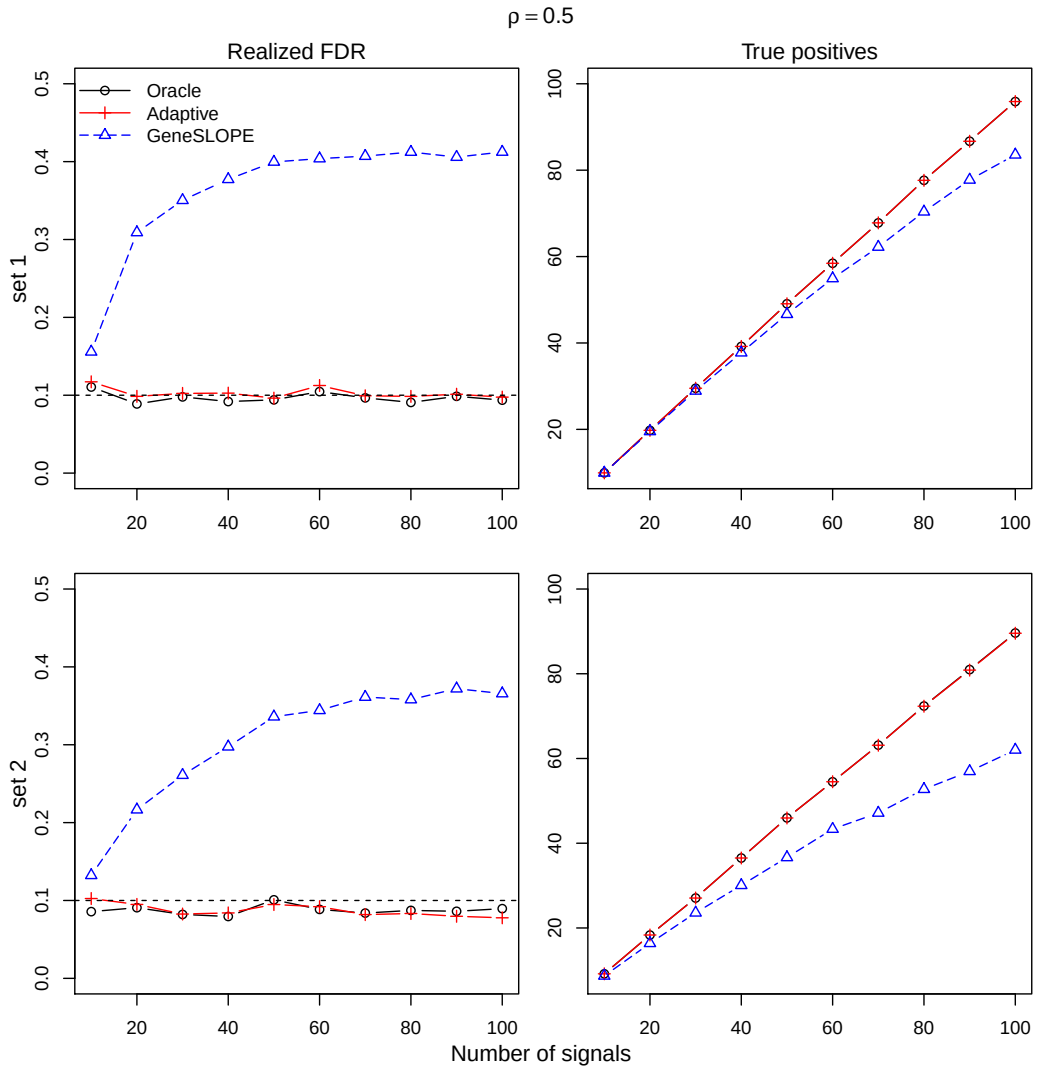

Figure B.2: Realized FDR levels and true positives for various numbers of signals ($\rho = 0.5$ and true null hypotheses as the blocks that contain any causal SNP)

Figures B.3–B.4 show the realized FDR and true positives when the definition of true null hypotheses in Brzyski et al. (2017) is used. We can see that our procedures can control the FDR at the target level while `GeneSLOPE` fails in the control. `GeneSLOPE` has slightly more numbers of true positives than our procedures in set 1, especially for $\rho = 0.5$. However, we can easily see that in Figure B.2 the true positives for `GeneSLOPE` are greater than the number of simulated signals, and this verifies that more than one

block is considered to be non-null for one causal SNP and the FDR and true positives are distorted. Our procedures still yield more true positives in set 2, and they are able to detect more signals with small effect sizes.

The simulation results above show that our procedures are able to control the FDR using their blocking procedure and both definitions of true null hypotheses, while `GeneSLOPE` fails in the control in all the cases. Our procedures yield more power than `GeneSLOPE` when our definition of true null hypotheses is used. When the definition of true null hypotheses in Brzyski et al. (2017) is used, the FDR and true positives are distorted, and our procedures can detect more signals with small effect sizes.

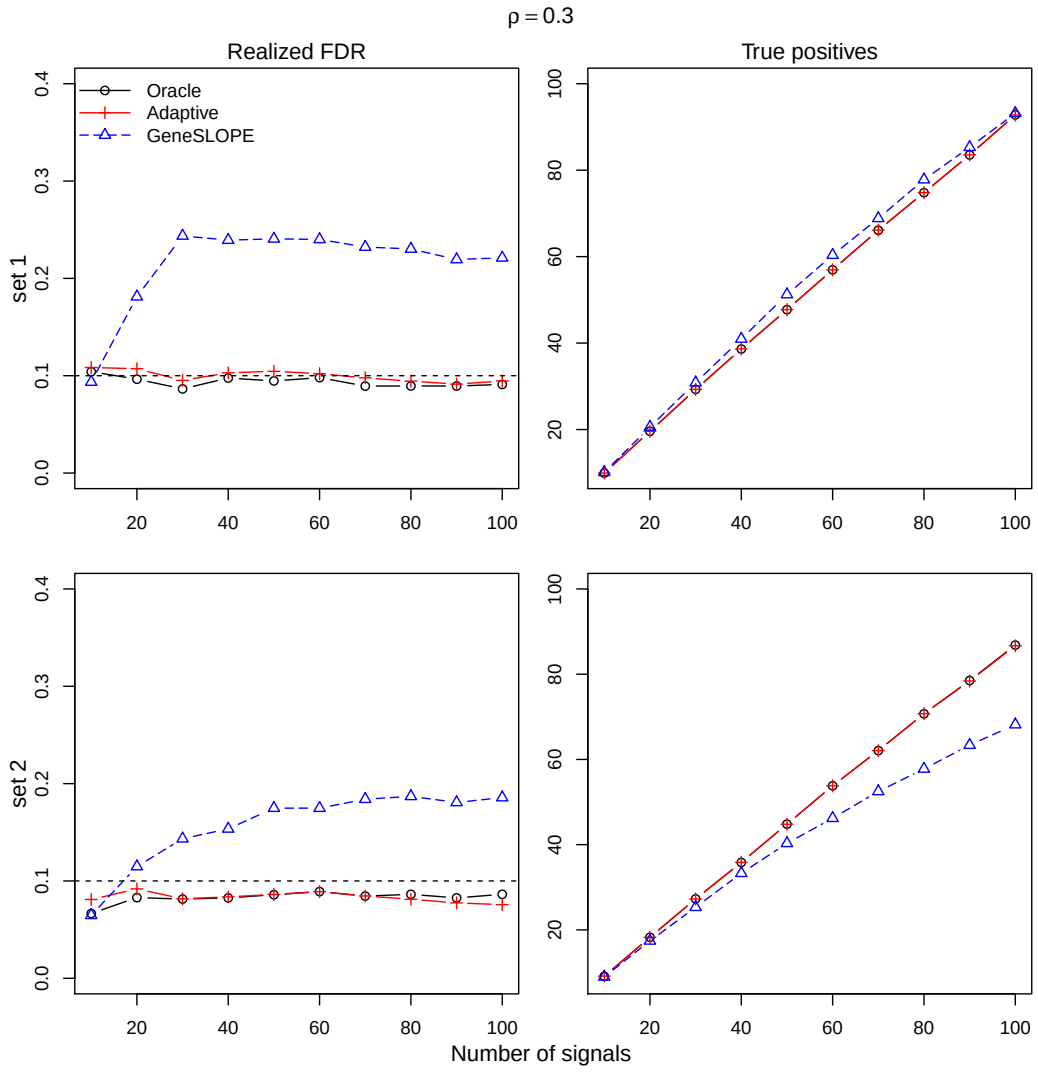

Figure B.3: Realized FDR levels and true positives for various numbers of signals ($\rho = 0.3$ and true null hypotheses defined as the blocks whose representative SNPs have correlations higher than 0.3 with any causal SNP)

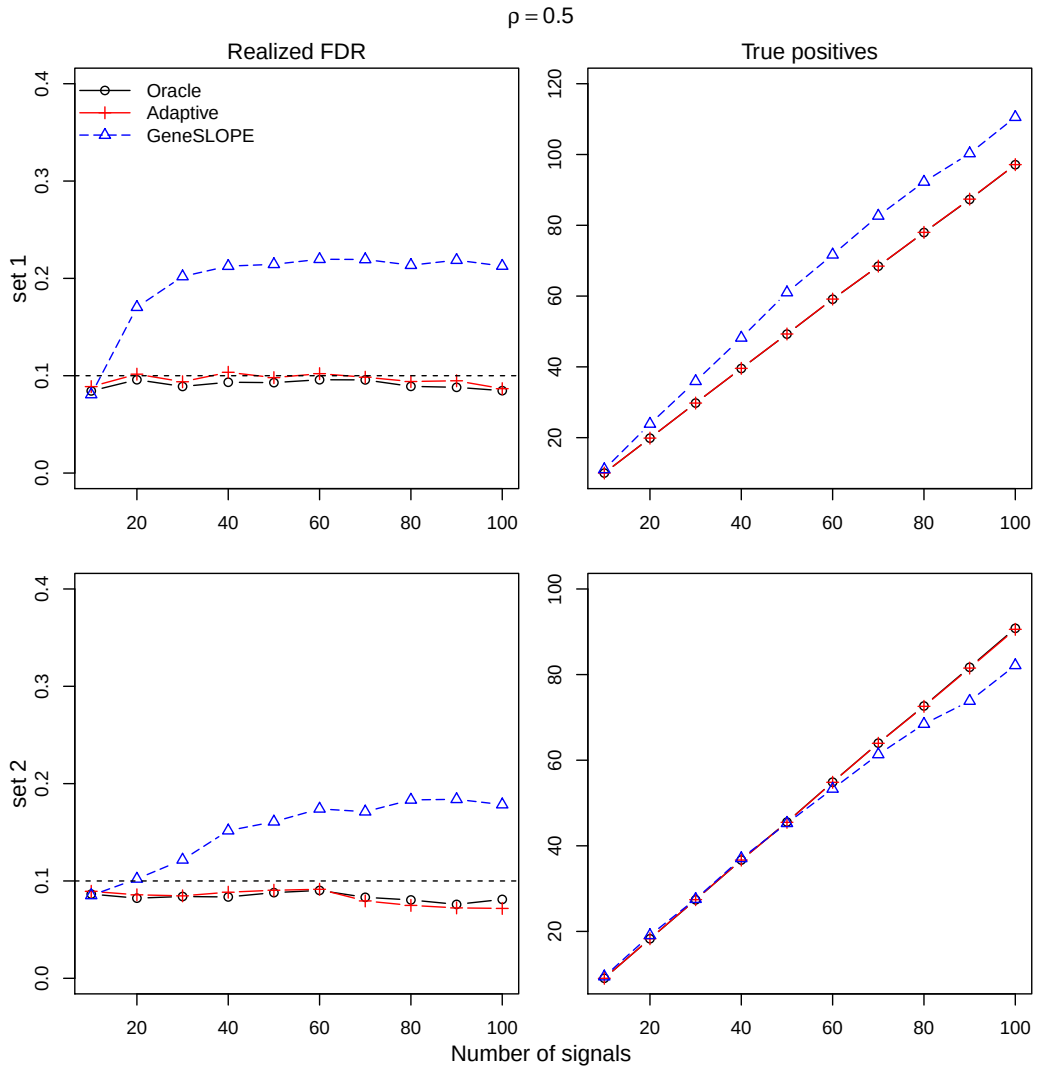

Figure B.4: Realized FDR levels and true positives for various numbers of signals ($\rho = 0.5$ and true null hypotheses defined as the blocks whose representative SNPs have correlations higher than 0.3 with any causal SNP)

Appendix C

Appendix in Chapter 4

C.1 EM algorithm

The EM algorithm recursively updates the estimates of parameters $\boldsymbol{\pi}$ and $\boldsymbol{\sigma}$ to maximize the likelihood, and alternates an expectation (E) step and an maximization (M) step until some convergence criterion is reached. For any iteration r, the E step computes the component weights w for each $\boldsymbol{Z}_i$ in each component, given the estimates of parameters from the last iteration $\hat{\boldsymbol{\pi}}^{(r-1)}$ and $\hat{\boldsymbol{\sigma}}^{(r-1)}$, as

$$
\begin{aligned}
w_{00i} &= E(H_{00i}|(\boldsymbol{z}_i, \hat{\boldsymbol{\pi}}^{(r-1)}, \hat{\boldsymbol{\sigma}}^{(r-1)})) = \hat{\pi}_{00}^{(r-1)} \hat{f}_{00}^{(r-1)}(\boldsymbol{z}_i)/\hat{f}^{(r-1)}(\boldsymbol{z}_i), \\
w_{01i} &= E(H_{01i}|(\boldsymbol{z}_i, \hat{\boldsymbol{\pi}}^{(r-1)}, \hat{\boldsymbol{\sigma}}^{(r-1)})) = \hat{\pi}_{01}^{(r-1)} \hat{f}_{01}^{(r-1)}(\boldsymbol{z}_i)/\hat{f}^{(r-1)}(\boldsymbol{z}_i), \\
w_{10i} &= E(H_{10i}|(\boldsymbol{z}_i, \hat{\boldsymbol{\pi}}^{(r-1)}, \hat{\boldsymbol{\sigma}}^{(r-1)})) = \hat{\pi}_{10}^{(r-1)} \hat{f}_{10}^{(r-1)}(\boldsymbol{z}_i)/\hat{f}^{(r-1)}(\boldsymbol{z}_i), \\
w_{11i} &= E(H_{11i}|(\boldsymbol{z}_i, \hat{\boldsymbol{\pi}}^{(r-1)}, \hat{\boldsymbol{\sigma}}^{(r-1)})) = \hat{\pi}_{11}^{(r-1)} \hat{f}_{11}^{(r-1)}(\boldsymbol{z}_i)/\hat{f}^{(r-1)}(\boldsymbol{z}_i),
\end{aligned}
$$

and $\hat{f}^{(r-1)}(\boldsymbol{z}_i) = \hat{\pi}_{00}^{(r-1)} \hat{f}_{00}^{(r-1)}(\boldsymbol{z}_i) + \hat{\pi}_{01}^{(r-1)} \hat{f}_{01}^{(r-1)}(\boldsymbol{z}_i) + \hat{\pi}_{10}^{(r-1)} \hat{f}_{10}^{(r-1)}(\boldsymbol{z}_i) + \hat{\pi}_{11}^{(r-1)} \hat{f}_{11}^{(r-1)}(\boldsymbol{z}_i)$. In the estimates of σ_1 and σ_2, we guarantee the estimates are greater than 0 by taking the maximum.

The M step then updates the estimates of parameters $\boldsymbol{\pi}$ and $\boldsymbol{\sigma}$ by maximizing the log likelihood given the weights. The explicit formula of new estimates are given as below,

$$
\hat{\pi}_{00}^{(r)} = \sum_{i=1}^{m} w_{00i}/m,
$$

$$\hat{\pi}_{01}^{(r)} = \sum_{i=1}^{m} w_{01i}/m,$$

$$\hat{\pi}_{10}^{(r)} = \sum_{i=1}^{m} w_{10i}/m,$$

$$\hat{\pi}_{11}^{(r)} = \sum_{i=1}^{m} w_{11i}/m,$$

$$\hat{\sigma}_1^{(r)} = max(1e-4, \frac{\sum_{i=1}^{m}(w_{10i}+w_{11i})z_{i1}^2}{\sum_{i=1}^{m}(w_{10i}+w_{11i})} - 1)^{1/2},$$

$$\hat{\sigma}_2^{(r)} = max(1e-4, \frac{\sum_{i=1}^{m}(w_{01i}+w_{11i})z_{i2}^2}{\sum_{i=1}^{m}(w_{01i}+w_{11i})} - 1)^{1/2},$$

$$\hat{\sigma}_{00}^{(r)} = \frac{\sum_{i=1}^{m} w_{00i}z_{i1}z_{i2}}{\sum_{i=1}^{m} w_{00i}} - \frac{\sum_{i=1}^{m} w_{00i}z_{i1}}{\sum_{i=1}^{m} w_{00i}} \frac{\sum_{i=1}^{m} w_{00i}z_{i2}}{\sum_{i=1}^{m} w_{00i}},$$

$$\hat{\sigma}_{01}^{(r)} = \frac{\sum_{i=1}^{m} w_{01i}z_{i1}z_{i2}}{\sum_{i=1}^{m} w_{01i}} - \frac{\sum_{i=1}^{m} w_{01i}z_{i1}}{\sum_{i=1}^{m} w_{01i}} \frac{\sum_{i=1}^{m} w_{01i}z_{i2}}{\sum_{i=1}^{m} w_{01i}},$$

$$\hat{\sigma}_{10}^{(r)} = \frac{\sum_{i=1}^{m} w_{10i}z_{i1}z_{i2}}{\sum_{i=1}^{m} w_{10i}} - \frac{\sum_{i=1}^{m} w_{10i}z_{i1}}{\sum_{i=1}^{m} w_{10i}} \frac{\sum_{i=1}^{m} w_{10i}z_{i2}}{\sum_{i=1}^{m} w_{10i}},$$

$$\hat{\sigma}_{11}^{(r)} = \frac{\sum_{i=1}^{m} w_{11i}z_{i1}z_{i2}}{\sum_{i=1}^{m} w_{11i}} - \frac{\sum_{i=1}^{m} w_{11i}z_{i1}}{\sum_{i=1}^{m} w_{11i}} \frac{\sum_{i=1}^{m} w_{11i}z_{i2}}{\sum_{i=1}^{m} w_{11i}}.$$

We iterate the E step and M step until the log-likelihood does not change much. Details of the EM algorithm are given as follows,

1. Initialize the parameters:
   set the proportions $\hat{\pi}_{00}^{(0)} = 0.8$, $\hat{\pi}_{01}^{(0)} = \hat{\pi}_{10}^{(0)} = 0.09$, $\hat{\pi}_{11}^{(0)} = 0.02$;
   set the variances $\hat{\sigma}_1^{(0)} = \hat{\sigma}_2^{(0)} = 1$;
   set the covariances $\hat{\sigma}_{00}^{(0)} = \hat{\sigma}_{01}^{(0)} = \hat{\sigma}_{10}^{(0)} = \hat{\sigma}_{11}^{(0)} = 0.1$;
   set the tolerance level $tol = 1e-4$, and the likelihood difference $d = 1$;
   set the maximum iterations $maxit = 1e4$;
   set the iteration $r = 1$;

2. for iteration $r < maxit$ and log-likelihood difference $df > tol$:

   (a) The E step: compute the weights w_{00i}, w_{01i}, w_{10i} and w_{11i} for each $\boldsymbol{z}_i = (z_{i1}, z_{i2})$ in the four components, given $\hat{\boldsymbol{\pi}}^{(r-1)}$ and $\hat{\boldsymbol{\sigma}}^{(r-1)}$;

(b) The M step: update the estimate of parameters $\hat{\boldsymbol{\pi}}^{(r)}$ and $\hat{\boldsymbol{\sigma}}^{(r)}$;

(c) Compute the difference of log-likelihood from iterations r and $r-1$,

$$d = |l^{(r)} - l^{(r-1)}|;$$

(d) If $d \leq tol$, then stop; else set $r = r + 1$ and repeat 2.(a)–(d).

The above algorithm still tries to estimate the correlation σ_{00} between primary and auxiliary z-values under null, and the estimate is very close to the formula given in (Lin and Sullivan 2009; Zaykin and Kozbur 2010) from simulations.