INVESTIGATIONS ON GENOMIC META-ANALYSIS: IMPUTATION FOR INCOMPLETE DATA AND PROPERTIES OF ADAPTIVELY WEIGHTED FISHER'S METHOD

by

Shaowu Tang

MS in Biostatistics, University of Pittsburgh, 2010

PhD in Mathematics, Jacobs University Bremen, Germany, 2005

BS in Applied Mathematics, Wuhan University, China, 1993

Submitted to the Graduate Faculty of

the Department of Biostatsitcs

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Shaowu Tang

It was defended on

April 10, 2014

and approved by

George C. Tseng, ScD

Associate Professor

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Daniel E. Weeks, PhD

Professor

Department of Human Genetics

Graduate School of Public Health

University of Pittsburgh

Jong-Hyeon Jeong, PhD

Professor

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Eleanor Feingold, PhD

Professor

Department of Human Genetics

Graduate School of Public Health

University of Pittsburgh

Abdus S. Wahed, PhD

Associate Professor

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Dissertation Director: **George C. Tseng**, ScD

Associate Professor

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

INVESTIGATIONS ON GENOMIC META-ANALYSIS: IMPUTATION FOR INCOMPLETE DATA AND PROPERTIES OF ADAPTIVELY WEIGHTED FISHER'S METHOD

Shaowu Tang, PhD

University of Pittsburgh, 2014

Abstract:

Microarray analysis to simultaneously monitor expression activities in thousands of genes has become a routine experiment in biomedical research during the past decade. The microarray expression data generated by high throughput experiments may consist of thousands of variables and therefore pose great challenges to researchers in a wide variety of statistical and computational issues. A commonly encountered problem by researchers is to detect genes differentially expressed between two or more conditions and is the major concern of this thesis.

In the first part of the thesis, we consider imputation of incomplete data in transcriptomic meta-analysis. In the past decade, a tremendous amount of expression profiles are generated and stored in the public domain and information integration by meta-analysis to detect differentially expressed (DE) genes has become popular to obtain increased statistical power and validated findings. Methods that combine p-values have been widely used in such a genomic setting, among which the Fisher's, Stouffer's, minP and maxP methods are the most popular ones. In practice, raw data or p-values of DE evidence of the entire genome are often not available in a subset of genomic studies that are to be combined. Instead, only the detected DE gene lists under certain p-value threshold (e.g. DE genes with p-value< 0.001) are reported in journal publications. The truncated p-value information voided the aforementioned meta-analysis methods and researchers are forced to apply less efficient vote counting

method or naïvely drop the studies with incomplete information. In the thesis, effective imputation methods were derived for such situations with partially censored p-values. We developed and compared three imputation methods – mean imputation, single random imputation and multiple imputation – for a general class of evidence aggregation methods of which Fisher, Stouffer and logit methods are special examples. The null distribution of each method was analytically derived and subsequent inference and genomic analysis frameworks were established. Simulations were performed to investigate the type I error and power for univariate case and the control of false discovery rate (FDR) for (correlated) gene expression data. The proposed methods were also applied to several genomic applications in prostate cancer, major depressive disorder (MDD), colorectal cancer and pain research.

In the second part, we investigate statistical properties of adaptively weighted (AW) Fisher's method. The traditional Fisher's method assigns equal weights to each study, which are sim- ple in nature but can not always achieve high power for a variety of alternative hypothesis settings. Intuitively more weight should be assigned to the studies with higher power to detect the difference between different conditions. The AW-Fisher's method, where the best binary 0/1 weights are determined by minimizing the p-value of the weighted test statistics, was proposed in Li and Tseng (2011). By using the order statistics technique, the searching space for adaptive weights reduces to linear complexity instead of exponential, which reduce the computational complexity dramatically, and a closed form is derived to compute the p-values for combining two studies, and an importance sampling algorithm is proposed to evaluate the p-values for combining more than two studies. Theoretical properties of the AW-Fisher's method such as consistency and asymptotical Bahadur optimality (ABO) are also investigated. Simulations will be performed to verify the asymptotical Bahadur optimality of the AW-Fisher and compare the performance of AW-Fisher and Fisher's methods.

Meta-analysis of multiple genomic studies increases the statistical power of biomarker detection and therefore the work in this thesis could improve public health by providing more effective methodologies for biomarker detection in the integration of multiple genomic studies when the information is incomplete or when different hypothesis settings are tested.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to express my deepest gratitude to my advisor, Dr. George C. Tseng for introducing me to the exiting field of genomic meta-analysis, and for his positive guidance throughout my research. His stimulating comments and arguments have been a constant source of inspiration.

I am also very thankful for Dr. Yongseok Park's valuable support. Without his help, the second problem can't be solved so perfectly.

There are some special persons whom I can never thank enough for the valuable expertise they shared with me, and also for their warmth and friendship: Dr. Jong-Hyeon Jeong had frequently given me the benefit of his advice and pertinent comments throughtout my research in survival analysis. Dr. Eleanor Feingold, Dr. Abdus S. Wahed and Dr. Daniel E. Weeks have given me good comments for my thesis writing and career advice.

I would like to thank all my group members Ying Ding, Xingbin Wang, Lun-Ching Chang, Rui Chen, Serena Liao, Masaki Lin, Silvia Liu, SungHwan Kim and Tianzhou Ma for providing me kind help in various projects.

Finally, I would like to thank my wife, Hong Qu. Without her support, I would never have succeeded.

1.0 INTRODUCTION

The rapid development of high-throughput experimental technology in the past decade has made the generation of genomic data increasingly affordable. This results in the rapid accumulation of experimental data in the public domains. The Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) is one example that is the largest public database to store gene expression data.

Among the vast amounts of gene expression data stored in the public domain, it is common that many of them were generated to test the same or similar hypothesis for the same disease. Since a single individual study in general only contains a limited number of samples, the statistical power of the test is relatively low and the generalizability of the conclusions is often unreliable. In order to improve the statistical power of the tests and provide validated conclusions, it is very common in practice that researchers attempt to combine information across different, independent studies. This is done using a class of meta-analysis methods that are particularly useful in microarray data analysis.

In this thesis, I will emphasize on applying meta-analysis to microarray data. The Chapter 1 is outlined as follows. In section 1.1 I briefly review the microarray data analysis. In section 1.2, univariate meta-analysis methods and microarray meta-analysis are introduced. In section 1.3 several complementary hypothesis settings are introduced, and several important questions are posed that will be answered in this thesis and the structure of the dissertation is outlined.

1.1 MICROARRAY DATA ANALYSIS

In the past decades, microarray technology has become one of the most important and powerful tools that many researchers use to monitor genome wide expression levels of genes. In general a microarray may contain thousands of genes for a limited number of samples. Commonly used statistical methods for microarray data analysis include class comparison, multiple testing, class discovery, class prediction and pathway analysis and so on. Among them, the most popular application is to compare the expression of a set of genes for different conditions (for instance, cases versus controls).

Unlike the traditional epidemiological problems, microarrays monitor gene expression for thousands of genes simultaneously. The standard data structure of a set of microarray data are a series of rectangular matrices in which the rows represent the expression of genes and the columns represent samples. Therefore, one can express the microarrays by y_{gsk}, where y_{gsk} denotes the gene expression for the gth gene in the sth sample of the kth study for $g = 1, \cdots, G; s = 1, \cdots, S$ and $k = 1, \cdots, K$. Usually samples are identified by a clinical variable r_{sk} indicating their classes. Thus, for a given study k, $r_{sk} \in \{0, 1\}$ represents a two-class comparison problem and $r_{sk} \in \{1, \cdots, S\}$ leads to a multi-class comparison problem.

Microarray meta-analysis usually refers to combining multiple transcriptomic studies for detecting differentially expressed (DE) genes (or biomarkers) across two or more conditions (e.g., case and control) with statistical significance and/or biological significance (e.g., fold change). For DE gene detection, hypothesis testing (such as two-sample t-test) is performed per gene. Since multiple hypothesis tests are performed, the problem of multiple comparisons should be addressed. For example, N tests generate an average of αN significant genes or biomakers at significance level α by chance. Therefore, false discovery rate (FDR) should be controlled for microarray analysis. A widely used procedure to control FDR is the B-H method proposed by [Benjamini and Hochberg, 1995].

1.2 META-ANALYSIS AND MICROARRAY META-ANALYSIS

Meta-analysis refers to systematic methods that integrate information from different, independent studies by using statistical techniques. Although the name of "meta-analysis" was invented by Glass in 1976 [Glass 1976], some of the techniques of meta-analysis can be traced back to a long time before that. Pearson performed the first meta-analysis in 1904 to summarize the correlation coefficients across studies of typhoid vaccination (Pearson 1904). Tippett (1931), Fisher (1948), and Wilkinson (1951) also proposed methods for to combine p-values. Today, meta-analysis is widely used in epidemiology and the field of medical research.

In meta-analysis, two major types of statistical techniques have been used: combining effect sizes and combining p-values.

1.2.1 Combining effect sizes

In the methods of combining effect sizes, the fixed effect model and random effect model are most popular [Cooper et al., 2009]. These methods are usually more straightforward and powerful to directly synthesize information of the effect size estimates and should be used in priority when the effect sizes are well-defined and comparable across different studies.

1.2.1.1 Fixed effects model In fixed effect model, one assumes that there is one true effect size θ and all the differences in observed effects are due to sampling error. In other words, the fixed effect model can be written as

$$T_k = \theta + \epsilon_k \text{ with } \epsilon_k \sim \ \mathrm{N}(0, \sigma_k^2),$$

where T_k is the observed effect size of study k. So each effect size T_k estimates a single mean effect θ, and differs from this mean effect by sampling error ϵ_k.

1.2.1.2 Random effects model

In fixed effect models, the true effect size is assumed to the same in all studies, which in many applications is implausible. More generally, in random effect models, each effect size is assumed to differ from the underlying population mean θ, due to both sampling error and the underlying population variance, i.e., the random effect model can be writen as

$$T_k = \theta + \epsilon_k + \zeta_k \text{ with } \epsilon_k \sim \text{ N}(0, \sigma_k^2), \zeta_k \sim \text{ N}(0, \tau^2).$$

1.2.2 Combining p-values

P-value combination methods are good alternatives of effect size combination methods when the effect sizes are not directly comparable across different studies. The well-known p-value combination methods include Fisher's method [Fisher, 1948], Stouffer's method [Stouffer, 1949], minP method [Tippett, 1931] and maxP method [Wilkinson, 1951].

These methods can be divided further into two classes: evidence aggregation methods and order-statistics based methods. The maxP and minP methods are two commonly used order-statistics based meta-analysis methods, since they use the order statistics of the observed p-values as their test statistics. On the contrary, Fisher and Stouffer methods are among the most popular evidence aggregation meta-analysis methods, in which the test statistics are defined as the sum of selected transformations of p-values for each individual study, i.e, the evidence is aggregated when new studies are included into the analysis. In this section, we assume that the null hypothesis is $H_0 : \cap_{k=1}^{K}\{\theta_k = 0\}$, where θ_k is the true effect size of study k.

1.2.2.1 Evidence aggregation methods

For evidence aggregation methods, given a set of p-values $\{p_1, \cdots, p_K\}$, the test statistic is defined as

$$T = \sum_{k=1}^{K} T_k := \sum_{k=1}^{K} F_X^{-1}(p_i),$$

where $F_X(\cdot)$ is the cumulative distribution function (CDF) of some random variable X.

In theory, any random variable X can be picked up to form a combining p-values method. However, only those Xs such that the null distribution of T is simple under the null hypothesis H_0 are selected. In this thesis, we focus on three popular special cases:

1. **Fisher's method**: When $X \sim \chi^2_2$, $T_k = F_X^{-1}(p_k) = -2\log(p_k)$.
2. **Stouffer's method**: When $X \sim \mathrm{N}(0,1)$, $T_k = F_X^{-1}(p_k) = \Phi^{-1}(p_k)$.
3. **Logit method**: When $X \sim \mathrm{Logistic}(0,1)$, $T_k = -\log\frac{p_i}{1-p_i}$ and $\sqrt{\frac{3}{\pi^2}\frac{5K+4}{K(5K+2)}}T \sim t_{5K+4}$ approximately (Hedges and Olkin 1985) under null hypothesis. For $K \geq 5$, it has been further approximated $\sqrt{\frac{3}{\pi^2}\frac{5K+4}{K(5K+2)}}T \sim \mathrm{N}(0,1)$.

1.2.2.2 Order-statistic based methods

Given a set of p-values $\{p_k\}_{k=1}^K$, let $\{p_{(k)}\}_{k=1}^K$ be its ordered version. Then for order-statistic based methods, the order statistic is selected as the test statistic (Song and Tseng 2014), i.e.,

$$T := p_{(r)} \sim \ \mathrm{Beta}(r, K-r+1) \text{ for } 1 \leq r \leq K.$$

Obviously minP and maxP are special cases with $r = 1$ and $r = K$ respectively.

1.2.3 Microarray meta-analysis

When multiple microarray studies are available, meta-analysis can be used to increase the statistical power for DE gene detection. Most meta-analytic methods for microarray studies are based on extensions of the univariate meta-analysis methods used for traditional medical research. Rhode was the first one to apply the conventional Fisher's method for combining multiple microarray studies [Rhode 2002]. In this thesis I will focus on the methods of combining p-values. Since these test statistics have simple analytical forms of null distributions, they are easy to apply to the genomic setting. Recall that given study k, suppose an appropriate test statistic T_k is selected for comparison $\{r_{sk}, 1 \leq s \leq S_k\}$ and the resulting p-values for each gene g (denoted as p_{gk}) can be derived from the observed expression intensities, then for each fixed g, the conventional meta-analysis methods can be applied to $\{p_{gk}\}_{k=1}^K$ for information integration. The final p-values obtained for each gene $\{p_g\}_{g=1}^G$ will be adjusted by the B-H method to control the FDR, and the DE genes can be detected at different FDR thresholds.

1.3 COMPLEMENTARY HYPOTHESIS SETTINGS

In meta-analysis one needs to combine independent p-values from a set of hypothesis tests. Given K individual hypotheses $H_{0k} : \theta_k = 0$ for $k = 1, \cdots, K$, the joint null hypothesis is defined as $H_0 : \cap_{k=1}^{K}\{\theta_k = 0\}$. Obviously H_0 is true only if all the effect sizes are 0 and false when at least one effect size is non-zero. It has been shown that there is no uniformly most powerful test, and some tests may be more powerful than others when some specific alternative hypotheses are true.

Two commonly encountered hypothesis settings are defined at what follows:

$$HS_A : H_0 : \bigcap_{k=1}^{K}\{\theta_k = 0\} \text{ versus } H_A : \bigcap_{k=1}^{K}\{\theta_k \neq 0\},$$
$$HS_B : H_0 : \bigcap_{k=1}^{K}\{\theta_k = 0\} \text{ versus } H_A : \bigcup_{k=1}^{K}\{\theta_k \neq 0\}.$$

In order for the null hypothesis to be false, in HS_A, the alternative hypothesis is the intersection event that effect sizes of all K studies are non-zero (i.e., the effect sizes in all studies are zero), while HS_B pursues non-zero effects in one or more studies (the alternative hypothesis is the union event instead of intersection in HS_A). Obviously, HS_A is more stringent and more desirable to identify consistency across all studies if the combined studies are homogeneous. HS_B, however, is useful when heterogeneity in effect sizes is expected.

HS_A and HS_B are closely related to two often-asked biological questions in genomic studies: "Which genes are significant in one or more data sets?" and "Which genes are significant in all data sets?". It is easy to know that the maxP method targets on HS_A and all other combining p-value methods target on HS_B.

Note that since in practice it is *a priori* unknown which individual null hypotheses are false, it is difficult for researchers to select appropriate hypothesis test with high power. In order to find a test which can achieve good power properties across such uncertainty, a new complementary hypothesis setting HS_r is defined as

$$HS_r : H_0 \text{ versus } H_r : \bigcap_{k=1}^{r}\{\theta_k \neq 0\} \text{ and } \bigcup_{k=r+1}^{K}\{\theta_k = 0\}.$$

In order the null hypothesis is false, at least r effect sizes should be non-zero.

1.4 SCOPE OF THE THESIS

In this thesis, I will focus on methods of combining p-values, which in turn implies that in order to utilize the methods, the p-value of each study is available in advance. Although this is generally true in conventional meta-analysis, it is not unusual that in many genomic studies the raw data are unavailable and only a partial DE gene list is reported with a given p-value threshold [Griffth et. al., 2006]. Therefore, for some gene g, only the range of p_{gk} is available, i.e., whether the gene is differentially expressed at a given p-value threshold α_k. The naïve methods which drop either the genes or the studies with incomplete information are not plausible, because they neglect the rich information contained in the truncated data. Therefore, there are practical needs to develop meta-analysis approaches that can efficiently combine truncated p-value information. One solution is to impute the truncated p-values before applying conventional meta-analysis. In this thesis, three imputation methods - the mean imputation, the single random imputation and the multiple imputation - are applied. In chapter 2, I investigate the imputation of truncated p-values for evidence aggregation meta-analysis methods.

When integrating multiple genomic studies, the expression patterns of genes may vary in a study specific manner. Li and Tseng proposed an adaptively weighted Fisher's method (AW-Fisher) to uncover the altered gene expression pattern across different studies [Li and Tseng, 2011], in which they started with the weighted statistic $U_g(w_g) = -\sum_{k=1}^{K} w_{gk} \log(p_{gk})$, where $w_{gk} \in \{0, 1\}$ is the weighted assigned to the kth study and $w_g = (w_{g1}, \cdots, w_{gK})$. Denoting by $p_U(u_g(w_g))$ the corresponding p-value, the adaptively weighted statistic is defined as the minimal p-value among all possible weights:

$$V_g^{AW} = \min_{w_g} p_U(u_g(w_g)) \text{ and } \hat{w_g} = \arg\min_{w_g} p_U(u_g(w_g)).$$

The resulting weight $\hat{w_g}$ reflects a natural biological interpretation of whether or not a study contributes to the statistical significance of a gene.

Recall that the number of studies where the null hypotheses are false is unknown *a priori*, the proposed AW-Fisher's method can maintain good power properties across such an uncertainty. In chapter 3, the AW-Fisher's method is generalized to a class of evidence aggregation meta-analysis methods and some properties such as the linear searching complexity, the asymptotical consistency of the weights and the asymptotic Bahadur optimality of the proposed tests will be investigated.

2.0 IMPUTATION OF TRUNCATED P-VALUES FOR EVIDENCE AGGREGATION META-ANALYSIS METHODS AND ITS GENOMIC APPLICATION

2.1 INTRODUCTION AND MOTIVATION

Microarray analysis to monitor expression activities in thousands of genes simultaneously has become routine in biomedical research during the past decade. The rapid development in biological high-throughput technology results in a tremendous amount of experimental data and many datasets are available from public domains such as Gene Expression Omnibus (GEO) and ArrayExpress. Since most microarray studies have relatively small sample sizes and limited statistical power, integrating information from multiple transcriptomic studies using meta-analysis techniques is becoming popular. Microarray meta-analysis usually refers to combining multiple transcriptomic studies for detecting differentially expressed (DE) genes (or candidate markers). DE gene analysis identifies genes differentially expressed across two or more conditions (e.g., cases and controls) with statistical significance and/or biological significance (e.g., fold change). Microarray meta-analysis in many situations refers to performing traditional meta-analysis techniques on each gene repeatedly and then controlling the false discovery rate (FDR) to adjust p-values for multiple comparison (Borovecki et al 2005; Cardoso et al. 2007; Pirooznia et al. 2007; Segal et al. 2004). Fisher's method (Fisher 1931) was the first meta-analysis technique introduced in microarray data analysis in 2002 (Rhodes et al. 2002), followed by Tippett's minimum p-value method in 2003 (Moreau et al. 2003). Subsequently many meta-analysis approaches have been used in this field, including extensions of existing meta-analysis techniques and novel methods to encompass the chal-

lenges presented in the genomic setting (Choi et al. 2003, Choi et al. 2007, Moerau et al. 2003, Owen 2009, Li and Tseng 2011, and see the review paper Tseng et. al. 2012).

To combine findings from multiple research studies, one needs to know either the effect size or the p-value for each study. Since the differences in data structures and statistical hypotheses across multiple studies may make the direct combination of effect sizes impossible or the result suspicious, combining p-values from multiple studies is often more appealing. One major category of combining p-value methods are evidence aggregation methods, which utilize summation of certain transformations of p-values as their test statistics and evidence will aggregate when new studies are included. Among evidence aggregation methods, Fisher's method is the most well-known, in which the test statistic is defined as $T^{Fisher} = -2\sum_{k=1}^{K}\log(p_k)$, where K is the number of independent studies to be combined and p_k is the p-value of individual study $k, 1 \leq k \leq K$. Under the null hypothesis of no effect size in all studies and assuming that studies are independent and models for assessing p-values are correctly specified, T^{Fisher} follows a chi-square distribution with degrees of freedom $2K$. Fisher's method has been popular due to its simplicity and some theoretical properties, including admissibility under Gaussian assumption (Birnbaum 1954 & 1955) and asymptotically Bahadur optimality (ABO) under equal non-zero effect sizes across studies (Littel and Folk, 1971). Some variations of Fisher's methods were proposed by using unequal weights or a trimmed version of Fisher's test statistic (Olkin and Saner, 2001). Another widely used evidence aggregation method is the Stouffer's method, in which the test statistic is defined as $T^{Stouffer} = \sum_{k=1}^{K}\Phi^{-1}(p_k)$ (Stouffer 1949), where $\Phi(\cdot)$ is the inverse CDF of standard normal distribution.

In order to combine p-values, all p-values across studies should be known. In genomic applications, however, raw data and thus p-values are often not available and usually only a list of statistically significant DE genes (p-value less than a threshold) is provided in the publication (Griffith et. al., 2006). Although many journals and funding agencies have encouraged or enforced data sharing policies, the situation has only improved moderately. Many researchers are still concerned about data ownership, and researchers whose studies are sponsored by

private funding are not obligated to share data in the public domain. For example, in Chan et. al (2007), publications of 23 colorectal cancer versus normal gene expression profiling studies were collected to perform meta-analysis to identify consistently reported candidate disease-associated genes. However only one raw dataset is available from the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/, GSE3294) and most other papers only provided a list of DE genes (and their p-values) under a pre-specified p-value threshold. A second motivating example comes from a microarray meta-analysis study for pain research (LaCroix-Fralish 2011), in which 20 microarray studies of pain models were collected to detect the gene signature and patterns of pain conditions. Among the 20 studies, only one raw dataset was available on the author's website and all the other papers reported the DE gene lists under different thresholds.

In these two motivating examples (details to be shown in Section 4.1 and 4.2), the incomplete data forced researchers to either drop studies with incomplete p-values or apply the convenient vote counting method (Hedges and Olkin, 1980). Dropping studies with incomplete information greatly reduces statistical power and is obviously not applicable in the two motivating examples since only one study was available with complete data. The conventional vote counting procedure is well-known as flawed and low-powered (McCarley et al., 2001). Ioannidis et al., (2009) attempted to reproduce 18 microarray studies published in Nature Genetics during 2005-2006. Interestingly, only two were "in principle" replicated, six "partially" replicated and ten could not be reproduced. This result illustrates well the wide-spread difficulty of obtaining raw data or reproducing published results in the field. Therefore, developing methods to efficiently combine studies with truncated p-value information is an important problem in microarray meta-analysis.

In this chapter, we assume that $K = K_1 + K_2$ studies are to be combined. In K_1 studies, the raw gene expression data matrix and sample annotations are available and the complete p-values p_{gi} ($1 \leq g \leq G$ for genes and $1 \leq i \leq K_1$) can be reproduced for meta-analysis. For the remaining K_2 studies, either the raw data or annotation are not available. Only incomplete information of a DE gene list (under p-value threshold α_i for study i) is provided

from the Journal publication. In this situation, the available information is an indicator function $\mathbb{1}_{\{p_{gi} \leq \alpha_i\}}$ to represent whether the p-value of gene g in study i is smaller than α_i or not. We outline the chapter structure as the following. In Section 2.2, a general class of evidence aggregation meta-analysis methods under univariate scenario are investigated for the mean imputation, the single random imputation and the multiple imputation methods respectively, in which the exact or approximate null distributions are derived under the null hypotheses and the results are shown for three popular special cases of Fisher, Stouffer and Logit methods. Simulations for Fisher and Stouffer methods are performed in Section 2.3.1 to demonstrate the correct control of type I errors and the power of different imputation methods are compared with naïve methods and complete cases in the univariate meta-analysis scenario. In Section 2.3.2 simulations of expression profiles were performed to compare performance of different methods. Simulations were further performed in Section 2.3.3 using 8 major depressive disorder (MDD) and 7 prostate cancer studies where raw data were completely available and the true best performance (complete case) could be obtained. In Section 2.4 the proposed methods were applied to the two motivating examples. In Section 2.4.1 the proposed methods were applied to 7 colorectal cancer studies, where the raw data are available only in 3 studies and the rest of 4 studies only have DE gene lists under different p-value thresholds. In Section 2.4.2, the proposed methods were applied to 11 microarray studies of pain conditions, where no raw data was available.In Section 2.4.3, we developed an unconventional application of the proposed methods to facilitate the large computational and data storage needs in a liquid association meta-analysis. Discussions and conclusions are included in Section 2.5.

2.2 METHODS AND INFERENCES

2.2.1 Evidence aggregation meta-analysis methods

Here we consider a general class of univariate evidence aggregation meta-analysis methods (for gene g fixed), in which the test statistics are defined as the sum of selected transformations of p-values for each individual study. Without loss of generality, assuming that $F_X(\cdot)$ is the cumulative distribution function (CDF) of a random variable X, the test statistic T is defined as

$$T = \sum_{i=1}^{K} T_k := \sum_{k=1}^{K} F_X^{-1}(p_k), \tag{2.2.1}$$

where p_k is the p-value of study k.

Theoretically X can be any continuous random variable. However, in practice, X is usually selected such that the test statistic T follows a simple distribution. For instance, when $X \sim \chi^2_2$, it holds $T \sim \chi^2_{2K}$ (Fisher's method) and $T \sim \mathrm{N}(0, K)$ holds, provided $X \sim \mathrm{N}(0, 1)$ (Souffer's method).

The hypothesis that corresponds to testing the homogeneous effect sizes of K studies by evidence aggregation methods is a union-intersection test (UIT) (Roy 1953):

$$H_0 : \bigcap_{k=1}^{K} \{\theta_k = 0\} \text{ versus } H_A : \bigcup_{k=1}^{K} \{\theta_k \neq 0\}. \tag{2.2.2}$$

In this paper, we focus on two popular special cases:

1. Fisher's method (Fisher 1931): $T_k = -2\log(p_k)$, i.e., $X \sim \chi^2_2$. Under null hypothesis, $T \sim \chi^2_{2K}$ if studies are independent.
2. Stouffer's method (Stouffer 1949): $T_k = \Phi^{-1}(p_k)$, i.e., $X \sim \mathrm{N}(0, 1)$. Under null hypothesis, $T \sim \mathrm{N}(0, K)$ if studies are independent.

Another example is the logit method (Hedges and Olkin, 1985), where $T_k = -\log(\frac{p_k}{1-p_k})$. But since this method is rarely used in practice, we will not examine it further here. To apply the evidence aggregation meta-analysis methods mentioned above, all the p-values should be observed. However, in genomic applications, it often happens that p-values of some studies are truncated and only their ranges are reported. Two naïve methods are commonly used to overcome this situation: vote counting method or the available-case method which only combines studies with observed p-values. The available-case method discards rich information contained in the studies with truncated p-values, and therefore the statistical power is reduced. Hedges and Olkin (1980) showed that the power of vote counting converges to 0 when many studies of moderate effect sizes are combined and therefore the vote counting method should be avoided whenever possible. In this section, three imputation methods - mean imputation, single random imputation and multiple imputation method - are proposed and investigated to combine studies with truncated p-values and the corresponding null distributions are derived analytically, respectively. We first define some notations.

Assume that K independent studies are to be combined and $p_1, \cdots, p_K$ are the corresponding p-values. Without loss of generality, assume that all the p-values are available in the the first K_1 studies and only the indicator function of DE evidence are reported in the other K_2 studies.

Define a pair $(c_i, x_i),\ i = 1, \cdots, K$ for each study, in which c_i is the "censoring" indicator satisfying

$$c_i := \begin{cases} 0, & \text{if } p_i \text{ is observed(i.e.}, 1 \le i \le K_1), \\ 1, & \text{if } p_i \text{ is censored(i.e.}, K_1 + 1 \le i \le K), \end{cases} \tag{2.2.3}$$

and x_i is the final observed values which is defined as

$$x_i := \begin{cases} p_i, & \text{if } c_i = 0, \\ \mathbb{1}_{\{p_i < \alpha_i\}}, & \text{if } c_i = 1, \end{cases} \tag{2.2.4}$$

where α_i is the p-value threshold for study i $(K_1 + 1 \le i \le K_1 + K_2 = K)$. For each $i = 1, 2, \cdots, K$, one can impute the missing value by $\tilde{p}_i$:

$$\tilde{p}_i = p_i \cdot \mathbb{1}_{\{c_i=0\}} + [q_i \cdot \mathbb{1}_{\{x_i=1\}} + r_i \cdot \mathbb{1}_{\{x_i=0\}}] \cdot \mathbb{1}_{\{c_i=1\}}$$

with $q_i \in (0, \alpha_i)$, and $r_i \in [\alpha_i, 1)$. Section 2.2-2.4 develop three imputation methods for selection of q_i and r_i.

2.2.2 Mean imputation method

The simplest imputation method is the mean imputation method, in which $q_i = \frac{\alpha_i}{2}$ and $r_i = \frac{1+\alpha_i}{2}$. Then the test statistic $\tilde{T}$ for truncated data satisfies

$$\tilde{T} = \sum_{i=1}^{K} \tilde{T}_i = \sum_{i=1}^{K} F_X^{-1}(\tilde{p}_i) = \sum_{i=1}^{K_1} F_X^{-1}(p_i) + \sum_{j=1}^{K_2} F_X^{-1}(\tilde{p}_{K_1+j}) = A + \sum_{j=1}^{K_2} B_j, \tag{2.2.5}$$

with $A = \sum_{i=1}^{K_1} F_X^{-1}(p_i)$ and

$$B_j = F_X^{-1}(\tilde{p}_{K_1+j}) = F^{-1}(\frac{\alpha_{K_1+j}}{2}) \cdot \mathbb{1}_{\{p_{K_1+j} < \alpha_{K_1+j}\}} + F^{-1}(\frac{1+\alpha_{K_1+j}}{2}) \cdot \mathbb{1}_{\{p_{K_1+j} \geq \alpha_{K_1+j}\}} \tag{2.2.6}$$

for $j = 1, \cdots, K_2$. Recall that under null hypothesis, the random variable A satisfies $A \sim \chi^2_{2K_1}$ for the Fisher's method and $A \sim \mathrm{N}(0, K_1)$ for the Stouffer's method. Obviously B_j follows a Bernoulli distribution.

The results can be summarized into the following theorem:

Theorem 2.2.1. *For $j = 1, 2, \cdots, K_2$ and given t, by defining*

$$b_j = F_X^{-1}(\frac{\alpha_{K_1+j}}{2}) - F_X^{-1}(\frac{1+\alpha_{K_1+j}}{2}) \text{ and } c = \sum_{j=1}^{K_2} F_X^{-1}(\frac{1+\alpha_{K_1+j}}{2}), \tag{2.2.7}$$

it holds

$$\mathbb{P}(\tilde{T} \leq t) = \sum_{(j_1, \cdots, j_{K_2}) \in \{0,1\}^{K_2}} \Pi_{i=1}^{K_2} \alpha_{K_1+i}^{j_i} (1 - \alpha_{K_1+i})^{1-j_i} F_A(t - c - \sum_{i=1}^{K_2} j_i b_i), \tag{2.2.8}$$

where $F_A(\cdot)$ is the CDF of A. Given the CDF, the expected values of test statistic $\tilde{T}$ under null distributions can be calculated as follows.

1. *For the Fisher's method, it holds*

$$\mathbb{E}(\tilde{T}) = 2K_1 - 2\sum_{j=1}^{K_2}[\alpha_{K_1+j}\log(\frac{\alpha_{K_1+j}}{2}) + (1-\alpha_{K_1+j})\log(\frac{1+\alpha_{K_1+j}}{2})],$$

while the expectation of the original T *is* $\mathbb{E}(T) = 2K_1 + 2K_2 = 2K$.

2. *For the Stouffer's method, it holds*

$$\mathbb{E}(\tilde{T}) = \sum_{j=1}^{K_2}[\alpha_{K_1+j}\Phi^{-1}(\frac{\alpha_{K_1+j}}{2}) + (1-\alpha_{K_1+j})\Phi^{-1}(\frac{1+\alpha_{K_1+j}}{2})],$$

while the expectation of the original T *is* $\mathbb{E}(T) = 0$.

Proof. Note that in this case, for $j = K_1+1, \cdots, K$, it holds

$$B_j = F_X^{-1}(\frac{\alpha_j}{2}) \cdot \mathbb{1}_{\{p_i<\alpha_j\}} + F_X^{-1}(\frac{1+\alpha_j}{2}) \cdot \mathbb{1}_{\{p_i\geq\alpha_j\}}. \tag{2.2.9}$$

Let $Y_j \sim \text{Bernoulli}(\alpha_j)$. Since $p_i \sim \text{Uniform}(0,1)$ under null hypothesis, it holds

$$B_j = [F_X^{-1}(\frac{\alpha_j}{2}) - F_X^{-1}(\frac{1+\alpha_j}{2})]Y_j + F_X^{-1}(\frac{1+\alpha_j}{2}) = b_jY_j + c_j, \tag{2.2.10}$$

and therefore

$$\tilde{T} = A + \sum_{j=1}^{K_2} b_jY_j + c \text{ with } c = \sum_{j=1}^{K_2} c_j. \tag{2.2.11}$$

For given t, it holds

$$\mathbb{P}(\tilde{T} \leq t) = \mathbb{P}(A + \sum_{i=1}^{K_2} b_i Y_i + c \leq t) \tag{2.2.12}$$

$$= \sum_{(j_1,\cdots,j_{K_2})\in\{0,1\}^{K_2}} \mathbb{P}(A + \sum_{i=1}^{K_2} b_i Y_i + c \leq t | Y_1 = j_1, \cdots, Y_{K_2} = j_{K_2}) \mathbb{P}(Y_1 = j_1, \cdots, Y_{K_2} = j_{K_2})$$

$$= \sum_{(j_1,\cdots,j_{K_2})\in\{0,1\}^{K_2}} \Pi_{i=1}^{K_2} \alpha_i^{j_i} (1-\alpha_i)^{1-j_i} \mathbb{P}(A \leq t - c - \sum_{i=1}^{K_2} j_i b_i)$$

$$= \sum_{(j_1,\cdots,j_{K_2})\in\{0,1\}^{K_2}} \Pi_{i=1}^{K_2} \alpha_i^{j_i} (1-\alpha_i)^{1-j_i} F_A(t - c - \sum_{i=1}^{K_2} j_i b_i),$$

where $F_A(\cdot)$ is the CDF of A.

Note that there are 2^{K_2} terms summation in the right hand side of Equ. (2.8), which may cause severe computing problem when K_2 is large. However, when some α_i are equal, the formula can be simplified. Without loss of generality, assume there are $r \geq 1$ different p-value thresholds $\{\beta_1, \cdots, \beta_r\}$ such that

$$\sum_{j=1}^{K_2} \mathbb{1}_{\{\alpha_{K_1+j}=\beta_1\}} = n_1, \cdots, \sum_{j=1}^{K_2} \mathbb{1}_{\{\alpha_{K_1+j}=\beta_r\}} = n_r \text{ and } \sum_{l=1}^{r} n_l = K_2, \tag{2.2.13}$$

then by defining $f(j; n_l, \beta_l) := C_j^{n_l} \beta_l^j (1-\beta_l)^{n_l - j}$ for $j = 0, \cdots, n_l$ and $l = 1, \cdots, r$, the formula can be simplified as

$$\mathbb{P}(\tilde{T} \leq t) = \sum_{j_1=0}^{n_1} \cdots \sum_{j_r=0}^{n_r} \Pi_{l=1}^{r} f(j_l; n_l, \beta_l) F_A(t - c - \sum_{l=1}^{r} j_l (F_X^{-1}(\frac{\beta_l}{2}) - F_X^{-1}(\frac{1+\beta_l}{2}))). \tag{2.2.14}$$

Therefore, the summation is reduced from 2^{K_2} terms to $\Pi_{l=1}^{r}(n_l + 1)$ terms.

From the above theorem one concludes that $\tilde{T}$ is a biased estimator of the original T. This motivates the following two stochastic imputation methods.

2.2.3 Single random imputation method

It is well-known that the mean imputation method will underestimate the variance of $\{p_{K_1+j}\}_{j=1}^{K_2}$ (Little and Rubin 2002). Furthermore, Theorem 2.2.1 shows that the test statistic $\tilde{T}$ from the mean imputation method is a biased estimator of the original T. To avoid this problem, one can replace the mean by randomly simulating q_i and r_i from Uniform$(0, \alpha_i)$ and Uniform$(\alpha_i, 1)$ respectively.

Recall that for $j = 1, \cdots, K_2$, $B_j = F_X^{-1}(\tilde{p}_{K_1+j})$. The next theorem states that $B_j \sim X$ holds under the null hypothesis, i.e., B_j and X follow the same distribution.

Theorem 2.2.2. *For $j = 1, 2, \cdots, K_2$, it holds*

$$B_j \sim X. \tag{2.2.15}$$

Proof. We show that for given t

$$\begin{aligned}
\mathbb{P}(B_i \le t) &= \mathbb{P}(F_X^{-1}(\tilde{p}_i) \le t) = \mathbb{P}(\tilde{p}_i \le F_X(t)) \\
&= \mathbb{P}(x_i = 1) \cdot \mathbb{P}(\tilde{p}_i \le F_X(t) | x_i = 1) + \mathbb{P}(x_i = 0) \cdot \mathbb{P}(\tilde{p}_i \le F_X(t) | x_i = 0) \\
&= \alpha_i \mathbb{P}[q_i \le F_X(t)] + (1 - \alpha_i)\mathbb{P}[r_i \le F_x(t)] \\
&= \begin{cases} \alpha_i \cdot \frac{F_X(t)}{\alpha_i} = F_X(t), & \text{if } t \in (-\infty, F_X^{-1}(\alpha_i)], \\ \alpha_i + (1 - \alpha_i) \cdot \frac{F_X(t) - \alpha_i}{1 - \alpha_i} = F_X(t), & \text{if } t \in (F_X^{-1}(\alpha_i), \infty) \end{cases} \\
&= F_X(t),
\end{aligned} \tag{2.2.16}$$

which implies that

$$B_i \sim X. \tag{2.2.17}$$

The following corollary is a simple consequence of the above theorem.

Corollary 2.2.3. *For the single random imputation method, the following facts hold for* $\tilde{T}$*:*

1. *For Fisher's method, it holds* $B_j \sim \chi^2_2$ *and* $\tilde{T} \sim \chi^2_{2K}$.

2. *For Stouffer method, it holds* $B_j \sim N(0,1)$ *and* $\tilde{T} \sim N(0,K)$.

Therefore, in this case, $\tilde{T}$ is a unbiased estimator of T.

2.2.4 Multiple imputation method

Although the single random imputation method allows the use of standard complete-data meta-analysis methods, it cannot reflect the sampling variability from one random sample. The multiple imputation method (MI) overcomes this disadvantage (Little and Rubin 2002). In MI, each missing value is imputed D times. Therefore $\{\tilde{T}^l\}_{l=1}^D$ is a sequence of test statistics which are defined as

$$\tilde{T}^l = \sum_{i=1}^{K} F_X^{-1}(\tilde{p}_i^l) = A + \sum_{j=1}^{K_2} B_j^l, \quad \text{for } l = 1, \cdots, D \tag{2.2.18}$$

with

$$q_i^l \sim \text{Uniform}(0, \alpha_i) \text{ and } r_i^l \sim \text{Uniform}(\alpha_i, 1). \tag{2.2.19}$$

The test statistic is defined as the average $\overline{T} = \frac{1}{D}\sum_{l=1}^{D} \tilde{T}^l$ which satisfies,

$$\begin{aligned}
\overline{T} &= A + \sum_{j=1}^{K_2}[(\frac{1}{D}\sum_{l=1}^{D} F_X^{-1}(q^l_{K_1+j})) \cdot \mathbb{1}_{\{p_{K_1+j} < \alpha_{K_1+j}\}} + (\frac{1}{D}\sum_{l=1}^{D} F_X^{-1}(r^l_{K_1+j})) \cdot \mathbb{1}_{\{p_{K_1+j} \geq \alpha_{K_1+j}\}}] \\
&= A + \sum_{j=1}^{K_2}[(\frac{1}{D}\sum_{l=1}^{D} W_j^l) \cdot \mathbb{1}_{\{p_{K_1+j} < \alpha_{K_1+j}\}} + (\frac{1}{D}\sum_{l=1}^{D} V_j^l) \cdot \mathbb{1}_{\{p_{K_1+j} \geq \alpha_{K_1+j}\}}] \\
&= A + \sum_{j=1}^{K_2}[\overline{W}_j \cdot \mathbb{1}_{\{p_{K_1+j} < \alpha_{K_1+j}\}} + \overline{V}_j \cdot (1 - \mathbb{1}_{\{p_{K_1+j} < \alpha_{K_1+j}\}})] = A + \sum_{j=1}^{K_2} Z_j.
\end{aligned}$$

Since $Z_j = \overline{W}_j$ with probability α_{K_1+j} and $Z_j = \overline{V}_j$ with probability $1 - \alpha_{K_1+j}$, Z_j is a mixture distribution of $\overline{W}_j$ and $\overline{V}_j$ and therefore $\overline{T} - A$ is a mixture distribution of $\{\overline{W}_j, \overline{V}_j, j = 1, \cdots, K_2\}$.

Note that W_j^l and V_j^l are independent and identically distributed (i.i.d) for fixed j. Denote by $(\mu_{W_j}, \sigma^2_{W_j}), (\mu_{V_j}, \sigma^2_{V_j})$ the mean and variance of W_j^l and V_j^l respectively. By the central limit theorem one concludes that for large enough $D > 0$ it holds

$$\overline{W}_j = (\frac{1}{D}\sum_{l=1}^{D} W_j^l) \sim \mathrm{N}(\mu_{W_j}, \frac{\sigma^2_{W_j}}{D}), \text{ and } \overline{V}_j = (\frac{1}{D}\sum_{l=1}^{D} V_j^l) \sim \mathrm{N}(\mu_{V_j}, \frac{\sigma^2_{V_j}}{D}).$$

Then the following theorem holds.

Theorem 2.2.4. *For $(j_1, \cdots, j_{K_2}) \in \{0,1\}^{K_2}$, by defining $U(j_1, \cdots, j_{K_2}) = \sum_{i=1}^{K_2}(j_i\overline{W}_i + (1-j_i)\overline{V}_i)$ which satisfies*

$$U(j_1, \cdots, j_{K_2}) \sim \ N[\sum_{i=1}^{K_2}(j_i\mu_{W_i} + (1-j_i)\mu_{V_j}), \frac{1}{D}\sum_{i=1}^{K_2}(j_i\sigma^2_{W_i} + (1-j_i)\sigma^2_{V_j})], \tag{2.2.20}$$

then for sufficiently large D, it holds approximately that

$$\mathbb{P}(\overline{T} \le t) = \sum_{(j_1,\cdots,j_{K_2})\in\{0,1\}^{K_2}} \Pi_{i=1}^{K_2}\alpha_i^{j_i}(1-\alpha_i)^{1-j_i}\mathbb{P}(A + U(j_1, \cdots, j_{K_2}) \le t). \tag{2.2.21}$$

The detailed notations are left to Section 2.2.5.

Similar to the mean imputation method, the formula can be simplified when some p-value thresholds are equal, i.e.,

$$\mathbb{P}(\overline{T} \le t) = \sum_{j_1=0}^{n_1} \cdots \sum_{j_r=0}^{n_r} \Pi_{l=1}^{r} f(j_l; n_l, \beta_l)\mathbb{P}(A + U(j_1, \cdots, j_r) \le t), \tag{2.2.22}$$

with $U(j_1, \cdots, j_r) = \sum_{l=1}^{r}(j_l F_X^{-1}(q_l) + (n_l - j_l)F_X^{-1}(r_l)), q_l \sim \text{Uniform}(0, \beta_l)$ and $r_l \sim \text{Uniform}(\beta_l, 1)$.

2.2.5 Some parameters in theorem 2.2.4 for the Stouffer's and Fisher's methods

2.2.5.1 Stouffer's method It is easy to obtain that

$$\begin{aligned}\mu_{W_i} &= \int_0^{\alpha} \frac{1}{\alpha}\cdot\Phi^{-1}(t)dt = \frac{1}{\alpha}\int_{-\infty}^{\Phi^{-1}(\alpha)} u d\Phi(u) = -\frac{1}{\alpha\sqrt{2\pi}}e^{-\frac{[\Phi^{-1}(\alpha)]^2}{2}}, \\ \mu_{V_i} &= \int_{\alpha}^{1} \frac{1}{1-\alpha}\cdot\Phi^{-1}(t)dt = \frac{1}{1-\alpha}\int_{\Phi^{-1}(\alpha)}^{\infty} u d\Phi(u) = \frac{1}{(1-\alpha_i)\sqrt{2\pi}}e^{-\frac{[\Phi^{-1}(\alpha)]^2}{2}}.\end{aligned} \tag{2.2.23}$$

and

$$\begin{aligned}\sigma^2_{W_i} &= 1 - \frac{\Phi^{-1}(\alpha)}{\alpha\sqrt{2\pi}}e^{-\frac{[\Phi^{-1}(\alpha)]^2}{2}} - \frac{1}{2\pi\alpha^2}e^{-[\Phi^{-1}(\alpha)]^2}, \\ \sigma^2_{V_i} &= 1 + \frac{\Phi^{-1}(\alpha)}{(1-\alpha)\sqrt{2\pi}}e^{-\frac{[\Phi^{-1}(\alpha)]^2}{2}} - \frac{1}{2\pi(1-\alpha)^2}e^{-[\Phi^{-1}(\alpha)]^2}.\end{aligned} \tag{2.2.24}$$

2.2.5.2 Fisher's method Similarly it holds

$$\begin{aligned}\mu_{W_i} &= \int_0^{\alpha} \frac{1}{\alpha}(-2\ln(t))dt = 2[1-\ln\alpha], \\ \mu_{V_i} &= \int_{\alpha}^{1} \frac{1}{1-\alpha}(-2\ln(t))dt = 2 + \frac{2\alpha}{1-\alpha}\ln(\alpha),\end{aligned} \tag{2.2.25}$$

and

$$\begin{aligned}\sigma^2_{W_i} &= \mathbb{E}(W_i^2) - \mu^2_{W_i} = 4, \\ \sigma^2_{V_i} &= \mathbb{E}(V_i^2) - \mu^2_{V_i} = 4 - \frac{4\alpha}{(1-\alpha)^2}\ln^2\alpha.\end{aligned} \tag{2.2.26}$$

2.3 SIMULATION RESULTS

Below we evaluate the proposed imputation methods using type I error and power in simulated data under univariate meta-analysis scenario in Section 2.3.1. In Section 2.3.2, we extend to simulated microarray data with correlated gene structure and assess the performance using the numbers of detected DE genes and false discovery rate (FDR). In Section 2.3.3, the proposed methods were applied to two real microarray datasets.

2.3.1 Control of type I error and power analysis for univariate meta-analysis

In this subsection we perform simulations to study the type I errors and statistical power of the proposed methods at various nominal levels of α. Two scenarios were used to investigate the type I errors and powers respectively at what follows:

(I) Investigation on type I errors: two random samples $\{X_i\}_{i=1}^{50}$ and $\{Y_i\}_{i=1}^{50}$ were both drawn from $\mathrm{N}(0,1)$;

(II) Investigation on powers: random samples $\{X_i\}_{i=1}^{50} \sim \mathrm{N}(0,1)$ and $\{Y_i\}_{i=1}^{50} \sim \mathrm{N}(0.3,1)$.

T-test was used to compare the means of the two samples and the p-values were generated. A vector of p-values with 8 entries, $(p_1, \cdots, p_8)$, was generated by repeating the process eight times and among them 5 were selected for p-value truncation. The observed data after truncation were (x_i, c_i) where $x_i = p_i$ for $c_i = 0$ $(1 \leq i \leq 3)$ and $x_i = \mathbb{1}_{\{p_i<\alpha\}}$ for $c_i = 1$ $(4 \leq i \leq 8)$. The three proposed imputation methods (mean imputation, single random imputation and multiple imputation) and available-case method were applied to the truncated data and the corresponding p-values were calculated by the closed form solutions forms given in Section 2.2. 1000 iterations were carried out and the proportion of cases that gave significant difference in means between two samples was evaluated at a given nominal level $\alpha = 0.05, 0.10$ or 0.15, which gave the empirical estimation for the type I error and powers respectively. For a full comparison, we also compared to the result using all eight raw p-values without truncation. The simulations were repeated 20 times for scenario I and scenario II to obtain the means and standard errors of the Type I errors and statistical power respectively.

The error bar plots of the simulation results for Type I errors and statistical power are presented in Figure 2.1 and 2.2 respectively, where D is set to be 100 in multiple imputation method. Figure 2.1 shows that all methods controlled their type I errors accurately for different significant level α. This is expected, because for all the methods we have derived their exact (or good approximated) null distributions. Figure 2.2 shows that the three imputation methods were more powerful than the available cases method, since they incorporated truncated information neglected by the available-case method. Among the imputation methods,

the single random imputation method has the lowest power. It took only one random draw which could be a bad guess of the underlying true value. The mean imputation method imputed by the mean value. Although it underestimated the variation in the imputation, the bias seemed relatively small and it offered better power than the single random imputation method. Multiple imputation method and mean imputation method achieved similar statistical power at various significant level α. Comparing to the complete case situation, we surprisingly found that mean imputation and multiple imputation methods could recover most of the statistical power that the complete case situation could achieve (e.g., for $\alpha = 0.05$ in Fisher's method, power = 0.807 and 0.804 for mean and multiple imputation compared to power = 0.887 for complete cases and power = 0.577 for available-case method.) This is particularly notable since large amount of p-value information has been truncated (5 out of 8 studies) in this simulation.

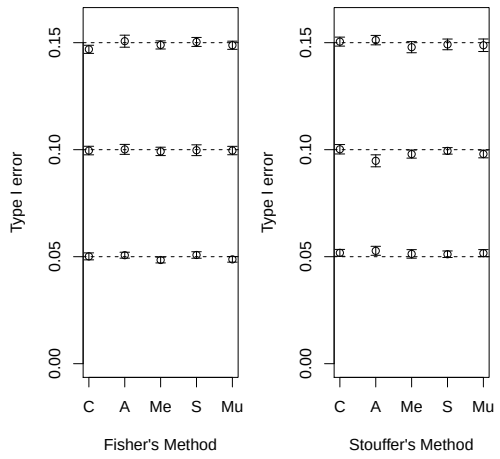

Figure 1: Type I error analysis. C: complete cases; A: available-case; Me: mean-imputation; S: single-imputation; Mu: multiple imputation when $\alpha = 5\%, 10\%$ and 15%.

In order to investigate the performance of multiple imputation method across different imputation number D, we repeated the two simulation scenarios for multiple imputation method with $D = \{20, 40, 60, 80, 100, 120, 140, 160, 180\}$ at significance level $\alpha = 0.05$. Similarly the

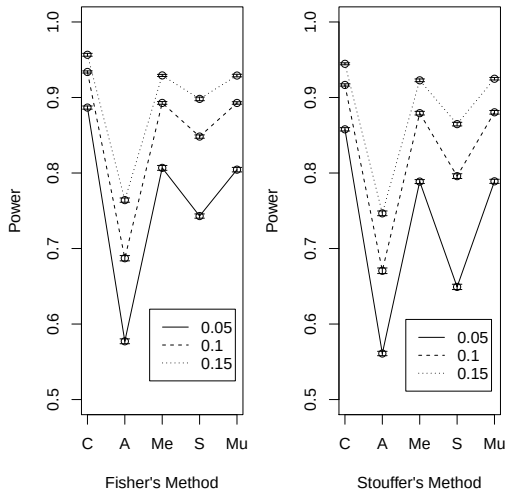

Figure 2: Power analysis. C: complete cases; A: available-case; Me: mean-imputation; S: single-imputation; Mu: multiple imputation when $\alpha = 5\%, 10\%$ and 15%.

error bar plots of the simulation results for Type I errors and powers are presented in Figure 2.3 and 2.4 respectively. Figure 2.3 demonstrates that Type I error for both methods can be well-controlled for moderately large D. In Figure 2.4 the horizontal dashed lines represent the power obtained by setting $D = 1000$, where for Fisher's method it is 0.805 and for Stouffer's method it is 0.797. Stouffer's method appears to converge faster than Fisher's method probably because its truncated transformation of W_j^l and V_j^l is closer to a Gaussian distribution. However, since the null distributions of the proposed multiple imputation method were derived based on central limit theorem, we recommend to set D being at least 50 for good approximation.

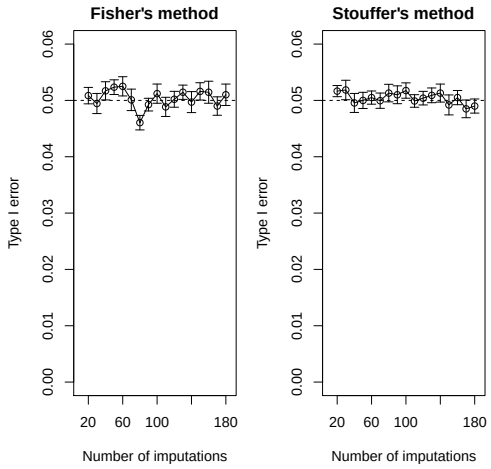

Figure 3: Type I error analysis at $\alpha = 0.05$ for different numbers of imputation D.

2.3.2 Simulated expression profiles

To evaluate performance of the proposed imputation methods in the genomic setting, we simulated expression profiles with correlated gene structure and variable effect sizes as follows.

Step 1 Randomly sample gene cluster labels of $10,000$ genes ($C_g \in \{0, 1, 2, \cdots, C\}$ and $1 \leq g \leq G$), such that $C = 200$ clusters each containing 20 genes are generated ($\sum_g \mathbb{1}(C_g = c) = 20, \forall 1 \leq c \leq C = 200$) and the remaining $6,000$ genes are unclustered genes ($\sum_g \mathbb{1}(C_g = 0) = 6,000$).

Step 2 For any cluster c ($1 \leq c \leq C$) in study k ($1 \leq k \leq K$), sample $\Sigma'_{ck} \sim W^{-1}(\Psi, 60)$, where $\Psi = 0.5I_{20\times 20} + 0.5J_{20\times 20}, W^{-1}$ denotes the inverse Wishart distribution, I is the identity matrix and J is the matrix with all the entries being 1. Set vector σ_{ck} as the square roots of the diagonal elements in Σ'_{ck}. Calculate Σ_{ck} such that $\sigma_{ck}\Sigma_{ck}\sigma_{ck}^T = \Sigma'_{gk}$.

Step 3 Denote by $g_1^{(c)}, \cdots, g_{20}^{(c)}$ as the indices for genes in cluster c. In other words, $C_{g_j^{(c)}} = c$, where $1 \leq c \leq 200$ and $1 \leq j \leq 20$. Sample expression of clustered genes by $(X'_{g_1^{(c)}nk}, \cdots, X'_{g_{20}^{(c)}nk})^T \sim MVN(0, \Sigma_{ck})$, where $1 \leq n \leq N = 100$ and $1 \leq k \leq K = 10$.

Sample expression for unclustered genes $X'_{gnk} \sim \mathrm{N}(0,1)$ for $1 \leq n \leq N$ and $1 \leq k \leq K$ if $C_g = 0$.

Step 4 Sample effect sizes μ_{gk} from Unif$(0.1, 0.5)$ for $1 \leq g \leq 1,000$ as DE genes and set $\mu_{gk} = 0$ for $1,001 \leq g \leq G$ as non-DE genes.

Step 5 For the first 50 control samples, $X_{gnk} = X'_{gnk}$ $(1 \leq g \leq G, 1 \leq n \leq N/2 = 50, 1 \leq k \leq K)$. For cases, $Y_{gnk} = X'_{g(n+50)k} + \mu_{gk}$ $(1 \leq g \leq G, 1 \leq n \leq N/2 = 50, 1 \leq k \leq K)$.

In the simulated datasets, $K = 10$ studies with $G = 10,000$ genes were simulated. Within each study, there were $\frac{N}{2} = 50$ cases and 50 controls. The first $1,000$ genes were DE in all 10 studies with effect sizes randomly simulated from a uniform distribution on $(0.1, 0.5)$ respectively, and the remaining $9,000$ were non-DE genes. Therefore, the DE genes have different effect sizes in different studies and the averaged mean effect size is $(0.1 + 0.5)/2 = 0.3$. In each study, 200 gene clusters existed, each containing 20 genes. The correlation structure within each cluster was simulated from an inverse Wishart distribution.

In the simulations, we performed a two sample t-test for each gene in each study and then combined the p-values using the imputation methods proposed in this paper. For simplicity, we viewed the p-values from the last 5 studies as truncated with thresholds $(\alpha_1, \cdots, \alpha_5) = (0.001, 0.001, 0.01, 0.01, 0.05)$ respectively. In most genomic meta-analysis, researchers often use conventional permutation analysis by permuting sample labels to compute the p-values to preserve gene correlation structure. However, such a nonparametric approach is not applicable in our situation, since raw data are not available in some studies. In order to control the false discovery rate (FDR), we examined Benjamini-Hochberg (B-H) method (Benjamini and Hochberg, 1995) and Benjamini-Yekutieli (B-Y) method (Benjamini and Yekutieli, 2001) separately. The number of DE genes detected at nominal FDR rate 5% were recorded and the true FDR rates were computed for each meta-analysis method by

$$\mathrm{FDR} = \frac{\sum_g \mathbb{1}(\text{gene g detected with } g \geq 1001)}{\#\{\text{genes detected}\}}.$$

In the multiple imputation method, $D = 50$ was selected. Simulations were repeated for 50 times and the mean and standard errors of numbers of DE genes controlled by BH and BY

methods and their true FDR are reported in Table 1. The results showed that the FDRs were controlled well for B-H correction but rather conservative for B-Y correction (the true FDR of B-Y is only 1/10 of B-H at nominal $FDR = 5\%$). This is consistent with the previous observation that the B-Y adjustment tends to be over-conservative since it guards against any type of correlation structure (Benjamini and Yekutieli, 2001). As a result, the BH correction will be used for all applications hereafter. The simulation results showed consistently that imputation methods had higher statistical power than the available-case method, and the mean imputation and multiple imputation methods outperform single random imputation method with similar performance. Surprisingly, the ratio of detected DE genes compared to complete case increased from 41.6% in available case (263.5/632.9) to 80.4% in mean imputation (508.6/632.9) using Fisher's method. The improvement is even more significant using Stouffer's method (from 41.8% to 86.7%), while at the same time the true FDRs were controlled at similar level for all methods. The result shows that imputation methods successfully utilize the incomplete p-value information to greatly recover the detection power.

We further examined the situation when gene dependence structure does not exist (i.e. Steps 1-3 were skipped and $X'_{gnk} \sim \mathrm{N}(0,1)$). Table 2.2 shows the true Type I error control under nominal significance level 5% (i.e.
True type I error $= \frac{\sum_{g=1,001}^{10,000} \mathbb{1}(\text{gene g is detected at significance level 0.05})}{9{,}000}$). The result shows adequate type I error control and confirms the validity of the closed form or approximated formula of different imputation methods in Section 2.
To investigate the impact of D on the performance of multiple imputation method, simulations were performed for $D \in \{20, 30, 50, 100, 150, 200, 250, 300, 500\}$. The result is shown in Figure 2.4 which demonstrates that the performance of multiple imputation method is quite robust for different number of imputation D. We use $D = 50$ throughout this thesis.

2.3.3 Simulation from complete real datasets

In this subsection, the proposed methods were applied to two real microarray datasets, including 7 prostate cancer studies (Gorlov 2009) and 8 major depressive disorder (MDD)

studies (Wang et al., 2012)). The details are summarized in Table 2.3. For each dataset, about half of the studies (four for MDD and three for prostate cancer) were randomly selected with p-value truncation threshold 0.05. Five methods including complete data, available-case, single random imputation, mean imputation and multiple imputation methods were applied to the datasets with the simulated incomplete data to impute by Stouffer's and Fisher's methods respectively. The generated p-values were corrected by the B-H method and the simulation was repeated for 50 times. Figure 1 shows boxplots of the numbers of differentially expressed (DE) genes at FDR = 1% for different methods in MDD and FDR = 0.5% for prostate cancer data. The result in Figure 2.6 indicates similar conclusion that multiple imputation and mean imputation methods detect more DE genes than the available-case method and single random imputation method. In the MDD example, very few DE genes (average of 16 and 83 for Fisher and Stouffer respectively) were detected using the available-case method if half of the studies have truncated p-values. The mean and multiple imputation methods greatly improved the detection sensitivity. About 95.2% (Fisher) and 96.3% (Stouffer) of DE genes detected by the mean imputation method overlapped with DE genes detected by complete data analysis in MDD and about 94.7% (Fisher) and 88.1% (Stouffer) of DE genes detected by the mean imputation method overlapped with DE genes detected by complete data analysis in prostate cancer, showing the ability of imputation methods to recover DE gene detection power.

2.4 APPLICATIONS

2.4.1 Application to colorectal cancer

In the first motivating example, we followed Chan et. al. (2007) and attempted to collect 23 colorectal cancer versus normal gene expression profiling studies. Raw data were available in only one study (Bianchini 2006) and other 4 studies containing more than 100 DE genes were included in our analysis. We searched the GEO database and identified two additional new

studies (Jiang et. al. 2008 and Bellot et. al. 2012). The seven studies under analysis were summarized in Table 3. After gene-matching, $6,361$ genes overlapped in all three studies with raw data. The available-case method, the mean imputation method, the single random imputation method and the multiple imputation method were applied for the seven studies for the Fisher and Stouffer methods respectively and the results were reported in Table 2.4. For the single random imputation method and multiple imputation method, the analyses were repeated 50 times and the mean and standard error of the number of DE genes detected were reported under FDR control by the BH method. The results demonstrate that for various FDR thresholds, the mean imputation method and the multiple imputation method detected more DE genes than the available-case method and the single random imputation method, which was consistent with previous findings in simulations. Under FDR = 0.01% control, Fisher and Stouffer mean imputation detected 2.07 (1183/571) and 10.35 (383/37) times of DE genes than those by available-case method, respectively.

Table 1: Simulation results for correlated data matrix at nominal FDR=5%

		Fisher		Stouffer	
	Method/Mean(s.e.)	No. DE	True FDR	No. DE	True FDR
BH	Complete cases	632.9(32.5)	0.043(0.0013)	518.6(36.2)	0.046(0.0015)
	available-case	263.5(37.4)	0.048(0.0076)	216.8(35.3)	0.064(0.022)
	Mean imputation	508.6(35.1)	0.046(0.0016)	449.8(36.2)	0.047(0.0022)
	Single imputation	408.9(35.7)	0.043(0.0018)	293.9(32.6)	0.045(0.0027)
	Multiple imputation	509.2(35.0)	0.045(0.0015)	463.8(35.7)	0.050(0.0019)
BY	Complete cases	354.0(34.4)	0.0041(0.00083)	261.7(33.9)	0.0036(0.00097)
	available-case	102.4(21.9)	0.0047(0.0012)	82.8(20.6)	0.0029(0.00096)
	Mean imputation	234.5(32.1)	0.0037(0.00074)	203.8(30.8)	0.0034(0.00073)
	Single imputation	164.0(27.3)	0.0057(0.0014)	113.5(22.3)	0.0039(0.0015)
	Multiple imputation	235.3(32.0)	0.0037(0.00075)	216.1(30.9)	0.0050(0.0010)

Table 2: Type I error control for independent data matrix at nominal significance level 5%

	Fisher	Stouffer
Complete cases	0.050(0.00031)	0.050(0.00037)
available-case	0.050(0.00035)	0.050(0.00033)
Mean imputation	0.050(0.00031)	0.050(0.00033)
Single imputation	0.050(0.00032)	0.051(0.00032)
Multiple imputation	0.050(0.00031)	0.051(0.00031)

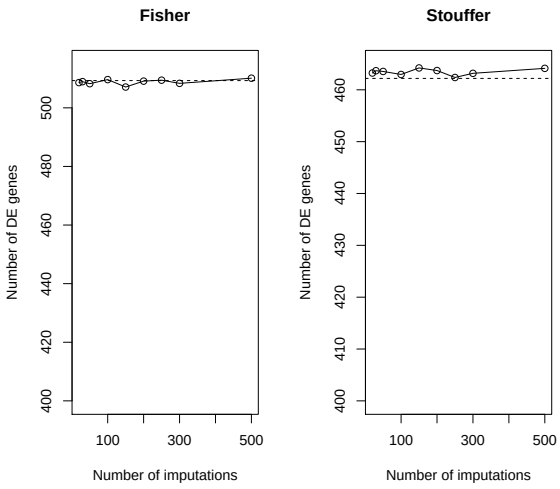

Figure 4: Number of DE genes at significance level 0.05 by multiple imputation method with different numbers of imputation D. The dashed lines represent the theoretical asymptotic power obtained by setting $D = 1000$.

Table 3: Detailed data sets description

	Author	Year	Platform	Sample Size (Case/Controls)	Source
Prostate Cancer Studies (Normal v.s Primary) (6940 genes)	Welsh	2001	HG-U95A	34(25/9)	public.gnf.org/cancer/
	Singh	2002	HG-U95Av2	102(52/50)	www.broad.mit.edu
	Lapointe	2004	cDNA	103(62/41)	GSE3933
	Yu	2004	HG-U95Av2	83(65/18)	GSE6919
	Varambally	2005	HG-U133 Plus 2	13(7/6)	GSE3325
	Wallace	2008	HG-U133A2	89(69/20)	GSE6956
	Nanni	2006	HG-U133A	30(23/7)	GSE3868
MDD Studies (7570 genes)	MD1_AMY	2009	HG-U133 Plus 2	28(14/14)	Dr. Sibille
	MD1_ACC	2009	HG-U133 Plus 2	32(16/16)	Dr. Sibille
	MD3_ACC	2009	HumanHT-12	44(22/22)	Dr. Sibille
	MD2_ACC_M	2010	HG-U133 Plus 2	18(9/9)	Dr. Sibille
	MD2_ACC_F	2010	HG-U133 Plus 2	26(13/13)	Dr. Sibille
	MD2_DLPFC_M	2010	HG-U133 Plus 2	28(14/14)	Dr. Sibille
	MD2_DLPFC_F	2010	HG-U133 Plus 2	32(16/16)	Dr. Sibille
	MD3_AMY	2009	HumanHT-12	42(21/21)	Dr. Sibille

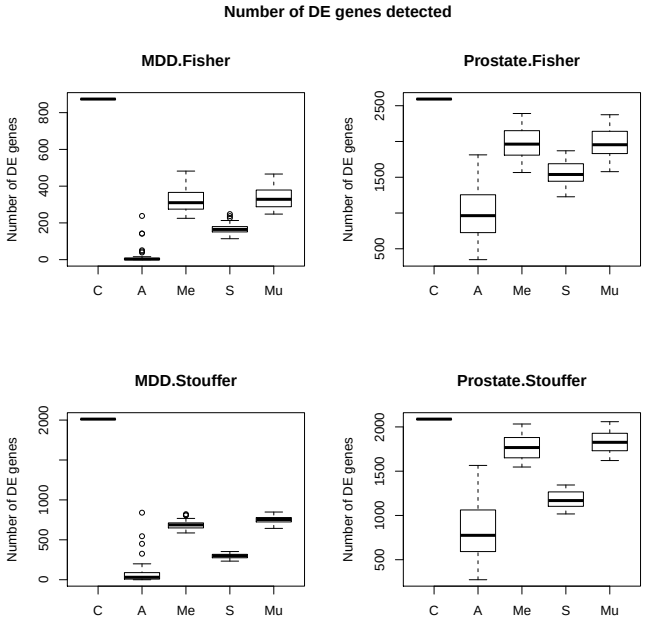

Figure 5: Number of DE genes detected by Fisher's or Stouffer's method. C: complete data; A: available-case; Me: mean-imputation; S: single-imputation; Mu: multiple imputation

Table 4: Seven colorectal cancer versus normal tissue expression profiling studies included in analysis

Study	No. of samples	No. of genes	Raw data availability	No. of DE genes	No. of overlapped DE genes	p-value threshold
*Bianchini*_2006	24	7403	GSE3294	-	-	-
*Bellot*_2012	17	18191	GSE24993	-	-	-
*Jiang*_2008	48	18197	GSE10950	-	-	-
*Grade*_2007	103	21543	-	1950	635	1e-7
*Croner*_2005	33	22283	-	130	47	0.006
*Kim*_2004	32	18861	-	448	143	0.001
*Bertucci*_2004	50	8074	-	245	97	0.009

Table 5: Summary of results for colorectal cancer

	Fisher				Stouffer			
FDR	Available	Mean	Single	Multiple	Available	Mean	Single	Multiple
1%	2587	2855	2172.4(2.90)	2785.4(2.93)	1318	1675	668.4(3.96)	1616.0(2.10)
0.1%	1472	1874	1265.6(2.34)	1805.7(1.50)	299	709	252.7(1.93)	680.5(1.12)
0.01%	571	1183	748.4(1.89)	1138.6(2.00)	37	383	102.5(1.65)	366.7(0.69)

2.4.2 Application to pain research

The second motivating example comes from the meta-analysis of 20 microarray studies of pain to detect the patterns of pain (LaCroix-Fralish, 2011). The original meta-analysis utilized DE gene lists from each study under different threshold criteria from p-value, FDR or fold change and identified 79 "statistically significant" genes that appeared in the DE gene lists of four or more studies. The vote counting method essentially lost tremendous amount of information with flawed statistical inference. When we attempted to repeat the meta-analysis, raw data of only one of the 20 studies ($Barr\_2005$) could be found. The old platform used in that study, however, contained only 792 genes and had to be excluded from further meta-analysis. In the remaining 19 studies, 11 studies contained DE gene lists under various p-value thresholds (marked bold in Table 2.6) and were included in our application. In other words, this example contained exclusively only studies with truncated p-values. Table 2.7 shows the result of three imputation methods. Fisher and Stouffer identified 280 and 45 genes under 5% FDR control, respectively. Note that the original meta-analysis tested the 79 genes using an overall binomial test and the statistical significance was controlled at an overall p-value level, not at a gene-specific FDR level. As a result, DE gene lists from the new imputation methods are theoretically more powerful and accurate.

To validate the finding, we used the Gene Functional Annotation tool from the DAVID Bioinformatics Resources website (http://david.abcc.ncifcrf.gov). DAVID applied a modified Fisher's exact test to evaluate the association between the DE gene lists and pathways. Functional annotation of the 280 DE genes from the Fisher's mean imputation method identified 208 pathways at FDR= 5%, among which selected important pain-related pathways were grouped into five major biological categories and displayed in Table 6. In contrast, the 79 genes from vote counting identified only 14 pathways, of which the expected pain-related pathways under the categories of inflammation and of differentiation, development and projection are missing (see Table 2.8). The pathway enrichment q-values after multiple comparison control of the "280 gene list" were very significant, while those of the "79 gene list" were not. Since the p-value calculation from Fisher's exact test can be impacted by the

DE gene size, we further compared the enrichment odd-ratios of genes in the pathway versus in the DE gene list. Still the enrichment odds-ratios of the "280 gene list" were generally much higher than those for the "79 gene list", showing stronger pain functional association from the Fisher's mean imputation method.

2.4.3 Application to a three-way association method (liquid association)

In the literature, it has long been argued that positively correlated expression profiles are likely to encode functionally related proteins. Liquid association (LA) analysis (Li 2002) is an advanced three-way co-expression analysis beyond the traditional pairwise correlations. For any triplet of genes X, Y and Z, the LA score $LA(X, Y|Z)$ measures the effect that expression of Z to control on and off of the co-expression between X and Y. For example, high expression of Z turns on positive correlation between X and Y while when expression of Z is low, X and Y are negatively or non-correlated. The theory in Li (2002) simplified the calculation of the LA score to a linear order of sample size and made the genome-wide computation barely feasible. Suppose we want to combine K studies of the liquid association, liquid association p-values of all triplets in all $K = 10$ studies have to be stored for meta-analysis. When the number of genes $G = 1,000$, the number of p-values to be stored is $G \cdot C_2^{G-1} \cdot K = 4.99GB$. For a reasonable $G = 20,000$ genome-wide analysis, storage size for all p-values quickly increases to $39.99TB$. One may argue that univariate (i.e. triplet by triplet) meta-analysis may be applied repeatedly to avoid the need of storing all p-value results. There are many other genomic meta-analysis situations when this may not be feasible. For example, in GWAS meta-analysis under a consortium collaboration, raw genotyping data cannot be shared for privacy reasons and only the derived statistics or p-values can be transferred for meta-analysis. Below we describe how imputation methods can help circumvent the tremendous data storage problem.

We performed a small scale of analysis on 566 DE genes previously reported from the meta-analysis of the eight MDD studies used in Section 3.2 (Wang et al., 2012). The total number

of possible triplets $(X, Y|Z)$ was $90,180,780$. By setting up p-value threshold at 0.001, we only needed to store exact p-values for $2,094,123$ ($\sim$ 2.32%) triplets and the remaining were truncated as considered in this paper. Since we also needed to store the truncation index information, we only needed to store $2 \times 2.32\% = 4.64\%$ of the information and the compression ratio was 95.36%. To investigate the loss of information by the truncation, Figure 2.7 indicates meta-analysis p-values (at $-\log(p)$ scale) from Fisher's method using full data and Fisher mean imputation method using truncated data. The result shows high concordance in the top significant triplets, which are the major targets of this exploratory analysis. Among the top 1000 triplets detected by Fisher's method using complete p-value information, 83.7% of them were also identified by the top 1000 by Fisher mean imputation. The remaining 163 triplets were still in top ranks (rank between 1199 and 4763) using truncated data in the result of Fisher mean imputation. This result suggests good potential of applying data truncation to preserve the most informative information and performing imputation to approximate the finding of the top targets when meta-analysis of "big data" is needed. The compression ratio may further increase by a more stringent truncation threshold but the performance may somewhat decline as a trade-off.

2.5 DISCUSSION AND CONCLUSION

When combining multiple genomic studies by p-value combination methods, the raw data are often not available and only the ranges of p-values are reported for some studies in genomic applications. This is especially true for microarray meta-analysis since owners of many microarray studies tend not to publish their data in the public domain. This incomplete data issue is often encountered when one attempts to perform a large-scale microarray meta-analysis. If raw data are not available, two naïve methods - vote counting method and available-case method - are commonly used. Since these two methods completely or largely neglect the information contained in the truncated p-values, and their statistical power is generally low. In this chapter, we proposed three imputation methods for a general class of

evidence aggregation meta-analysis methods to combine independent studies with truncated p-values: mean imputation, single random imputation and multiple imputation methods. For each proposed imputation method, the null distribution was derived analytically for the Fisher and Stouffer methods. Theoretical results showed that the test statistics from the single random imputation and the multiple imputation methods were unbiased, while those for mean imputation method were biased. Simulations were performed for the imputed Fisher method and imputed Stouffer method. The simulation results showed that type I errors were well-controlled for all methods, which was consistent with our theoretical derivation. Compared to the naive available-case method, all the imputation methods achieved higher statistical powers, and the mean imputation and the multiple imputation methods recovered much of the power that the complete cases method achieved even when half of the studies had truncated p-values. Furthermore, Supplementary Figure 1 showed that the power of the multiple imputation method was robust to the number of imputation D. Although small to moderate D provided good results, we recommend choosing D being larger than 50 to guarantee that central limit theorem can approximate well. Applications to two motivating examples in colorectal cancer and pain conditions showed that both mean imputation and multiple imputation performed among the best in terms of detection sensitivity and biological validation by pathway analysis.

In regression-type missing-data imputation methods, the null distribution of the error term is unknown and is assumed to be normally distributed with equal variance, a setting in which multiple imputation method usually outperforms mean imputation in practice and in theory (Little and Rubin 2002), particularly because mean imputation underestimates the true variance. However, our simulation results demonstrated that the power of the two methods were quite similar. Two reasons may contribute to this result. First, although the test statistic from the mean imputation method is biased and neglects the variation of truncated p-values, its p-value can be computed accurately when the null distribution is derived analytically. Second and more importantly, we find that the test statistic of mean imputation is in fact $F_X^{-1}(\mathbb{E}(p))$, while for sufficiently large D, the test statistic of multiple imputation converges to $\mathbb{E}(F_X^{-1}(p))$ in distribution. It is easy to show that these two quantities are very close

to each other for a small range of p, provided $F_X^{-1}(\cdot)$ is smooth. Since $F_X^{-1}(\cdot)$ is infinitely differentiable for the Fisher and Stouffer methods, and the small p-value range in $(0, \alpha)$ are particularly of interest to us, it is not surprising that the mean imputation method and multiple imputation method perform similarly. Since the mean imputation method achieved almost the same power as the multiple imputation method with less computational complexity, it is more appealing and is recommended for microarray meta-analysis, where the imputed meta-analysis method is performed repeatedly for thousands of genes. In this paper only the evidence aggregation meta-analysis methods are investigated and further work will be needed to extended these results to order statistic based methods such as minP and maxP.

Note that although the truncated p-value issue discussed in this chapter may appear similar to the problem of "publication bias", it is fundamentally different. Publication bias refers to the fact that a study with a large positive treatment effect is more likely to be published than a study with a relatively small treatment effect, resulting in bias if one only considers published studies. Denote by $p_1, p_2, \cdots, p_N$ the p-values of all conducted studies that should have been collected. Only a subset of likely more significant p-values $p_1, p_2, \cdots, p_n$ are observed. Under this setting, N is unknown and $p_{n+1}, \cdots, p_N$ are unknown as well. Since the number of missing publications is unknown, Duval and Tweedie proposed the "Trim and Fill" method to identify and correct for funnel plot asymmetry arising from publication bias (Duval and Tweedie, 2000a and 2000b), in which an estimate of the number of missing studies is provided and an adjusted treatment effect is estimated by performing a meta-analysis including the imputed studies. For the truncated p-value problem we consider here, the total number of studies, the number of studies with truncated p-values and the p-value truncation thresholds are all known. Therefore, investigation of the imputation of truncated p-values in meta-analysis is different from the traditional "publication bias" problem and has not been studied in the meta-analysis literature, to the best of our knowledge.

In this chapter, the methods we developed mainly target on microarray meta-analysis but the issue can happen frequently in other types of genomic meta-analysis (e.g. GWAS; Begurn et. al. 2012). In section 2.4.3, we demonstrated an unconventional application of our methods

to meta-analysis of liquid association. Due to the large number of triplets tested in the three-way association, the needed p-value storage is huge. By preserving only the most informative data by truncation, the storage burden is greatly alleviated and our imputation methods help approximate and recover the top meta-analysis targets with little power loss. In an on-going project, we also attempt to combine multiple genome-wide eQTL results via meta-analysis. In eQTL, regression analysis is used to investigate the association of a SNP genotyping and a gene expression. It is impractical to store all genome-wide eQTL p-values as the storage space required is too large ($25,000$ genes $\times$ $2,000,000$ SNPS $= 5 \times 10^{10}$ p-values). A practical solution is to record only the eQTL p-values smaller than a threshold (say 10^{-4}) for meta-analysis, which leads to the same statistical setting as discussed in this paper. In another project, we combine results from multiple ChIP-seq peak calling algorithms to develop a meta-caller. Since each peak caller algorithm can only report the top peaks with p-values smaller than a certain p-value threshold, we again encounter the same truncated p-value problem in meta-analysis. As more and more complex genomic data are generated and the need for meta-analysis increases, we expect the imputation methods we propose in this chapter will find even more applications in the future.

Table 6: Eleven pain-relevant microarray studies included in the analysis

Study	Species	No. of DE genes (% of total)	type of threshold	threshold
*Ko*_2002	Rat	42(0.5%)	fold change	2-fold
Costigan_2002	Rat	197(2.2%)	p-value	0.05
*Xiao*_2002	Rat	117(1.8%)	fold change	2-fold
Wang_2002	Rat	166(2.4%)	p-value	0.05
Sun_2002	Rat	44(0.6%)	p-value	0.05
*Bonilla*_2002	Mouse	13(0.2%)	fold change	2-fold
*Kubo*_2002	Mouse	53(0.6%)	fold change	2-fold
Valder_2003	Rat	139(2.0%)	p-value	0.05
Yang_2004	Rat	169(2.6%)	p-value	0.05
*Ren*_2005	Rat	31(15.1%)	FDR	0.05
*Barr*_2005	Rat	47(3.7%)	FDR	0.05
Nesic_2005	Rat	36(0.1%)	p-value	0.05
Rodriguez_2006	Rat	40(1%)	p-value	0.00047
LaCroix-Fralish_2006	Rat	805(17.1%)	p-value	0.01
Geranton_2007	Rat	74(0.2%)	p-value	0.05
Griffin_2007	Rat	96(1.1%)	p-value	0.01
Yukhananrov_2008	Rat	798(2.6%)	p-value	0.01
*Nishida*_2008	Rat	51(0.3%)	fold change	2-fold
*Levin*_2008	Rat	195(1.3%)	fold-change	3-fold

Table 7: Summary of results for patterns of pain

	Fisher	Stouffer
Mean	280	45
Single	57.04 (1.6228)	16.44(0.8605)
Multiple	280.36(0.8105)	77.56(0.6616)

Table 8: Summary of pathway analysis by DAVID

		280 DE (Fisher's mean imputation)			79 DE (Vote counting)		
Category	Pathway ID	pval	qval	odds ratio	pval	qval	odds ratio
Differentiation,	**GO** : 0030182 ∼	5.6e-6	0.0006	3.1	0.26	0.95	1.6
development and	**GO** : 0045664 ∼	1.6e-5	0.0011	4.7	0.37	0.98	1.9
projection	**GO** : 0048666 ∼	2.5e-6	0.0003	3.6	0.24	0.94	1.7
	GO : 0051960 ∼	6.5e-6	0.0006	4.2	0.29	0.96	1.9
	GO : 0031175 ∼	1.6e-5	0.0012	3.7	0.27	0.96	1.8
	GO : 0042995 ∼	3.6e-11	3.2e-9	3.5	0.033	0.47	1.9
	GO : 0043005 ∼	3.0e-11	3.4e-9	4.3	0.043	0.51	2.0
	GO : 0030030 ∼	1.6e-5	0.0012	3.3	0.24	0.94	1.7
Response to stimuli	**GO** : 0009611 ∼	3.8e-10	2.8e-7	4.3	2.7e-5	0.016	3.6
	GO : 0009719 ∼	3.2e-8	1.7e-5	3.4	0.35	0.97	1.3
	GO : 0048584 ∼	7.9e-8	2.5e-5	4.9	0.0049	0.34	3.6
	GO : 0032101 ∼	1.1e-5	0.001	4.8	0.043	0.71	2.8
Immune	**GO** : 0050778 ∼	4.2e-7	7.6e-5	5.9	0.018	0.57	4.0
	GO : 0002684 ∼	1.9e-6	0.0003	4.4	0.0009	0.13	4.2
	GO : 0006956 ∼	3.0e-5	0.0016	11.5	0.011	0.46	8.4
	GO : 0002478 ∼	1.3e-6	0.00022	19.0	0.00098	0.12	10.64
Inflammation	**GO** : 0002673 ∼	1.4e-6	0.0002	14.1	0.19	0.93	3.8
	GO : 0002526 ∼	7.1e-06	0.0007	6.7	0.012	0.48	4.4
	GO : 0050727 ∼	1.9e-5	0.0012	6.9	0.17	0.92	2.8
	GO : 0006954 ∼	1.5e-5	0.0012	4.1	0.001	0.11	3.8

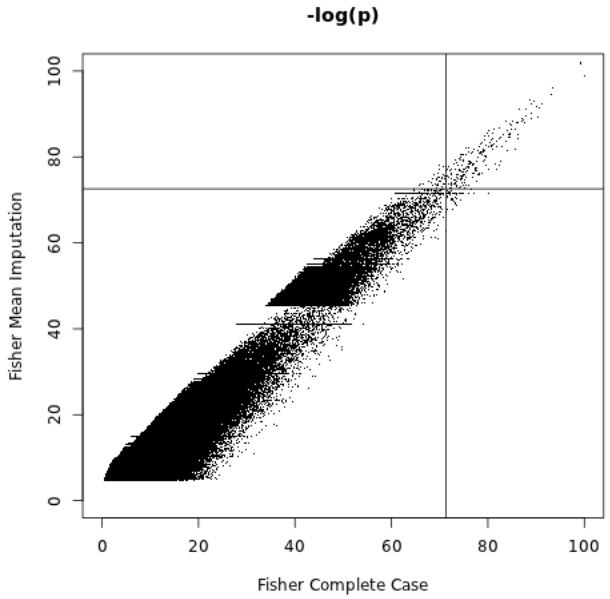

Figure 6: $-\log(p)$ comparison of the mean imputation method using truncated data with the complete case method using complete data. Vertical line: $x = 71.3$. Horizontal line: $y = 72.58$. Points right to vertical line are top $1,000$ triplets detected by Fisher's complete case method, and points above to horizontal line are top $1,000$ triplets detected by Fisher's mean imputation method

3.0 ON ADAPTIVE WEIGHTING FOR P-VALUE COMBINATION META-ANALYSIS

3.1 INTRODUCTION OF META-ANALYSIS

3.1.1 Genomic meta-analysis

Meta-analysis techniques have been widely used to combine results from multiple clinical or genomic studies which test the same or a similar hypothesis. To combine findings from multiple research studies, one can either combine the effect sizes directly or combine the p-values. Since the differences in data structures and statistical hypotheses across multiple studies may make the direct combination of effect sizes impossible or the result suspicious, combining p-values from multiple studies is often more appealing in omics applications. The frequently used p-value combination methods include the Fisher's method (Fisher, 1925), the Stouffer's method (Stouffer et al., 1949), the logit method (Lancaster 1961) and minP and maxP methods (Tippett 1931;Wilkinson 1951).

With the availability of tremendous genomic data in public domain, there is increased interest in combining multiple studies by meta-analysis techniques. In general, the genomic meta-analysis technique involves firstly performing combining p-values procedures for tens of thousands of biomarkers simultaneously and then controlling the false discovery rate (FDR) to correct the p-values for multiple comparisons. Since performing meta-analysis for large

amount of biomarkers simultaneously is computationally demanding, there is an urgent need to have a fast computation algorithm to reduce the computing cost. Furthermore, in order to control the FDR well, the p-values should be evaluated by a closed form solution or approximated accurately.

In the meta-analysis of genomic studies, in order to make a stronger scientific statement, it is more important and practical to identify those genes which are differentially expressed in a consistent pattern across multiple studies, i.e., to detect signals in the majority of studies. However, most combining p-value methods are targeting on the gain of statistical power by pooling together independent studies and producing a single combined p-value to show that at least one hypothesis is false, i.e., they are designed to test the hypothesis setting HS_B. Therefore, those combining p-values procedures can not handle the gene-specific heterogeneity, which makes it impossible to make a stronger scientific statement about the biological findings.

The problem first gained attention in fMRI research (Friston, Penny and Glaser 2005) and many other authors have tried to address this problem since then. For example, Song and Tseng (2013) proposed the rth ordered p-value (rOP) method to test the alternative hypothesis that the signal presents in at least a given percentage of studies. Li and Ghosh (2014) proposed a class of meta-analysis methods based on summaries of weighted ordered p-values (WOP). Li and Tseng (2011) proposed an adaptively weighted Fisher's method (AW-Fisher's) for gene expression data analysis to test the conjunction of null hypotheses against the alternative that at least one is false, where the weight of each study takes the value 0 or 1 and therefore only the subset of studies yielding the most significant results were selected for further analysis. Since expression of some important biomarkers may be altered in a study-specific manner, 0/1 weights reflect a natural biological interpretation of whether or not a study contributes to the statistical significance of a gene. Similar ideas such as AW-FEM and AW-Bayesian approach were applied to GWAS meta-analysis (Han and Eskin 2012; Bhattacharjee et al. 2012), where only the effect sizes in a subset of studies were assumed to be non-zero in alternative hypotheses (Flutre et al., 2013). In addition to producing a

single combined p-value, the AW-Fisher's method also returns a vector of weights indicating which studies are contributing to the significance of the signal. Therefore, although the AW-Fisher's method doesn't target on testing for partial conjunction hypotheses, one still can make a stronger scientific statement regarding how many studies share a consistent pattern when the null hypothesis is rejected.

3.1.2 Adaptively weighted Fisher's method

In this section we assume there are K studies to be combined and the effect size and corresponding p-value of study k is $\{\theta_k, p_k\}$, and the alternative hypothesis setting being dealt with in this section is defined as

$$HS_B : H_0 : \bigcap_{k=1}^{K} \{\theta_k = 0\} \text{ versus } H_B : \bigcup_{k=1}^{K} \{\theta_k \neq 0\}. \tag{3.1.1}$$

For Fisher's method, the test statistic T^{Fisher} is $T^{Fisher} := -2\sum_{k=1}^{K} \log(p_k)$, which follows a chi-square distribution with degree freedom of $2K$ (i.e., $T^{Fisher} \sim \chi^2_{2K}$) under the null hypothesis. Note that the Fisher's method only returns a p-value indicating whether the test of null hypothesis is statistical significant or not at a given significance level α, but gives no indication which studies contribute to the statistical significance. Although the AW-Fisher's method is designed to test hypothesis setting HS_B too, the returned adaptive weights indicate which studies have contributed to the statistical significance.

Figure 3.1 shows the heatmaps of gene expression for DE genes identified by Fisher's and AW-Fisher's methods for three tissue mouse datasets. The heatmap of Fisher's method gives no indication which studies contribute to the statistical significance, but the adaptive weights of AW-Fisher's method grouped together the genes sharing the same gene expression pattern. Therefore, the AW-Fisher's method can deal with gene-specific heterogeneity, which makes it very appealing in genomic data analysis.

Now by defining the vector of true weights W^* as

$$W^* = (w_1^*, \cdots, w_K^*)^T \text{ with } w_k^* = \mathbb{1}(\theta_k \neq 0) \text{ for } k = 1, \cdots, K, \tag{3.1.2}$$

then only the studies with nonzero effect size are included in the test statistic T^{W^*} and the tests are more powerful than $T^{\mathbf{1}}$. However, W^* is unknown and can only be approximated. Given a vector of weights $W \neq \mathbf{0}$, it holds $T^W \sim \chi^2_{2\sum_{k=1}^K w_k}$ and therefore the corresponding p-value p^W is

$$p^W = \mathbb{P}(\chi^2_{2\sum_{k=1}^K w_k} \geq -2\sum_{k=1}^K w_k \log(p_k)). \tag{3.1.3}$$

The estimate of W^* and the corresponding test statistic of AW-Fisher's method are defined as (Li & Tseng 2011)

$$\hat{W} := \text{argmin}_{W\neq\mathbf{0}}\{p^W\} \text{ and } T^{AW} := -\log(\min_{W\neq\mathbf{0}}\{p^W\}). \tag{3.1.4}$$

Here we show how to find the best weights by a simple toy example. Assume three studies are to be combined and the corresponding p-values of gene 1 are $(1, 1, 0.001)$. Table 3.1 summarizes every nonzero vector of weights and their corresponding test statistics and p-values. It is easy to know that $(0, 0, 1)$ are the best weights, since the p-value with respect to these weights is the most significant.

3.1.3 Open questions of AW-Fisher's method in Li and Tseng (2011)

It has been shown in Li & Tseng (2011) that AW-Fisher is admissible and have better power in a wide range of alternative hypotheses compared to minP, maxP and Fisher's methods. However, there remains several important theoretical and computational questions. As described in Li and Tseng (2011), in order to determine the best weights, theoretically all but nonzero vector of weights need to be searched and compared (i.e., $2^K - 1$), which is very time-consuming in genomic studies when the number of the studies to be combined is large. For example, if there are p-values of $10,000$ genes across 20 studies to be combined, the total number of searches for the best weights will be $10,000 \times (2^{20} - 1) = 10,485,750,000$. Furthermore, since the null distribution of the test statistic was not provided in Li and Tseng (2011), all the p-values were obtained by permutation tests, which further increased the computational burden and the p-values are not accurate and inefficient. For example, assuming there are 10,000 genes, in order to achieve precision to 10^{-12}, one needs to perform 10^8 permutations test, which generally is impossible in practice.

In addition to computing issues, there are also open theoretical questions. The goal of the AW-Fisher's method is to assign weights 1 to the studies with non-zero effect size and 0 to the studies with zero effect size. Therefore one question arises naturally that if the adaptive weights are asymptotically consistent? Since Fisher's method is ABO under a simplified Gaussian circumstance, one also wants to know if AW-Fisher is ABO.

In the sequel the solutions of two computing problems are shown in Section 3.2 and the answers to two theoretical questions are shown in Section 3.3.

3.2 SOLUTIONS TO TWO COMPUTING PROBLEMS

In this section the solutions to two open computational problems in Li & Tseng (2011) are discussed. A fast algorithm of searching the adaptive weights is provided in Section 3.2.1 and an importance sampling technique is proposed to obtain accurate p-values for $K \geq 3$ in Section 3.2.2.

3.2.1 Fast searching of the adaptive weights

Note that the searching space $\Omega = \{W : W \neq \mathbf{0}\}$ contains $2^K - 1$ vectors of weights and therefore searching the whole space Ω to find the adaptive weights $\hat{W}$ becomes very expensive when K is large. The situation becomes more severe when the AW-Fisher's method is applied to genomic data, in which the same procedure will be repeatedly performed for thousands of biomarkers simultaneously. In this section, based on the ordered p-values $\{p_{(i)}\}_{i=1}^K$, we proposed a fast algorithm to find $\hat{W}$ by searching only K vectors of weights instead of searching the whole exponential space. To this end, let's firstly rewrite Ω as $\Omega = \bigcup_{k=1}^K \Omega_k$ with $\Omega_k = \{W : \sum_{j=1}^K w_j = k\}$ for $k = 1, 2, \cdots, K$, then it holds $\hat{W} = \min_{W \in \Omega}\{p^W\} =$

$\min_{1\le k\le K}\{\min_{W\in\Omega_k}\{p^W\}\}$.

Lemma 3.2.1. *Denote* $\{p_{(1)},\cdots,p_{(K)}\}$ *as the ordered version of* $\{p_1,\cdots,p_K\}$*, then for* $k=1,2,\cdots,K$*, it holds*

$$-2\sum_{i=1}^{k}\log(p_{(i)})=\max_{W\in\Omega_k}\{-2\sum_{i=1}^{K}w_i\log(p_i)\}.$$

Proof. The proof is trivial.

Without loss of generality, denote by W^k the vector of weights such that

$$-2\sum_{j=1}^{K}w_j^k\log(p_j)=-2\sum_{j=1}^{k}\log(p_{(j)}),$$

then Lemma 3.2.1 demonstrates that the test statistic involving the first k ordered p-values will generate the most significant p-value in Ω_k, therefore in each Ω_k, only one vector of weights W^k has to be considered for further searching, which implies that instead of searching the whole space Ω, it is enough to search only K vectors of weights to find the adaptive weights $\hat{W}$, i.e.,

Corollary 3.2.2. T^{AW} *satisfies*

$$T^{AW}=-\log(\min_{1\le k\le K}\{p^{W^k}\})\ \textit{and}\ \hat{W}=\arg\min_{1\le k\le K}\{p^{W^k}\}. \tag{3.2.1}$$

The proposed fast algorithm contains two steps: firstly sorting K p-values and then searching K vectors of weights. The complexity of sorting a vector of K p-values varies from $\mathcal{O}(K)$ to $\mathcal{O}(K\log(K))$ and to $\mathcal{O}(K^2)$ by different sorting algorithms, and the complexity of searching K vectors of nonzero weights is of the order $\mathcal{O}(K)$. Therefore, the fast searching algorithm proposed in this section reduces the computational complexity from $\mathcal{O}(2^K)$ to at most $\mathcal{O}(K^2)$, which saves a lot of computing time when K is large (see Table 3.2 for comparison).

3.2.2 Computation of $\mathbb{P}(T^{AW} > -\log(t))$

Recall that $T^{AW} = -\log(t)$, where $t = \min_{1\le k\le K}\{p^{W_k}\} = p^{\hat{W}}$. Let $t_j = \exp(-\chi^{-2}_{2j}(t)/2)$ and $T_j = \prod_{i=1}^{j} P_{(i)}$, then

$$\mathbb{P}(T^{AW} > -\log(t)) = \mathbb{P}(\cup_{j=1}^{K} T_j < t_j). \tag{3.2.2}$$

Since the joint distribution of $P_{(i)}, i = 1,\dots,K$ is $f(p_1,\dots,p_K) = K!, 0 \le P_{(1)} \le P_{(2)} \le \cdots \le P_{(K)} \le 1$, it is possible to estimate $P(T^{AW} > -\log(t))$ analytically. Without loss of generality, let's denote by A_j the event $\{T_j < t_j\}$, then $\mathbb{P}(T^{AW} > -\log(t))$ can be rewritten as

$$\mathbb{P}(T^{AW} > -\log(t)) = \mathbb{P}(\cup_{j=1}^{K} A_j) = \mathbb{P}(A_1) + \sum_{k=2}^{K} \mathbb{P}(A_k \cap (\cup_{j=1}^{k-1} A_j^c)), \tag{3.2.3}$$

where A_j^c is the complementary event of A_j. The above formula provides an analytical way to compute the p-value $\mathbb{P}(T^{AW} > -\log(t))$. For example, when $K = 2$,

$$\begin{aligned}
\mathbb{P}(T^{AW} > -\log(t)) &= \mathbb{P}(\cup_{j=1}^{K} T_j < t_j) = 1 - \mathbb{P}(T_1 \ge t_1, T_2 \ge t_2) \\
&= 1 - \int_{t_1}^{1}\int_{t_2/p_1 1}^{1} 2\cdot \mathbb{1}(p_1 \le p_2) dp_2 dp_1 = 1 - \int_{t_1}^{1}\int_{\max\{t_2/p_1 1, p_1\}}^{1} 2 dp_2 dp_1 \\
&= 1 - \int_{t_1}^{1} 2[1 - \max\{t_2/p_1, p_1\}] dp_1.
\end{aligned}$$

In the case $t_1^2 \ge t_2$,

$$\int_{t_1}^{1} 2[1 - \max\{t_2/p_1, p_1\}] dp_1 = \int_{t_1}^{1} 2\{1 - p_1\} dp_1 = (1 - t_1)^2.$$

If $t_1^2 < t_2$,

$$\begin{aligned}
\int_{t_1}^{1} 2[1 - \max\{t_2/p_1, p_1\}] dp_1 &= \int_{t_1}^{1} 2dp_1 - \int_{\sqrt{t_2}}^{1} 2p_1 dp_1 - \int_{t_1}^{\sqrt{t_2}} 2t_2/p_1 dp_1 \\
&= 2(1 - t_1) - (1 - t_2) - t_2\log(t_2/t_1^2) = 1 - t_1 + t_2 - t_2\log(t_2/t_1^2).
\end{aligned}$$

Therefore, for $K = 2$, the p-value $\mathbb{P}(T^{AW} > -\log(t))$ for given observed test statistic $-\log(t)$ can be computed analytically by

$$\mathbb{P}(T^{AW} > -\log(t)) = \begin{cases} 2t_1 - t_1^2 & t_1^2 \ge t_2 \\ t_2\log(t_2/t_1^2) + 2t_1 - t_2, & t_1^2 < t_2 \end{cases}. \tag{3.2.4}$$

Theoretically, $\mathbb{P}(T^{AW} > -\log(t))$ can be computed analytically for any $K \geq 2$. However, the computation of $\mathbb{P}(A_k \cap (\cup_{j=1}^{k-1} A_j^c))$ involves the evaluation of a k-fold integral and the integration domain becomes very complicated for $k \geq 3$, which makes the derivation of the close form of $\mathbb{P}(A_k \cap (\cup_{j=1}^{k-1} A_j^c))$ very tedious and fallible. Hence in this paper, we propose to use importance sampling technique to obtain a numerical approximation of $\mathbb{P}(T^{AW} > -\log(t))$. Importance sampling is a method to reduce variance when using Monte Carlo sampling method. The idea behind importance sampling is to draw samples from a different distribution rather than the distribution of interest and assign a weight to each sample based on the two distributions. This technique is very useful when evaluating an unknown distribution or obtaining the mean of a certain function.

To evaluate $\mathbb{P}(T^{AW} > -\log(t(p_1, \ldots, p_K)))$, we use beta-distribution to draw P_i so that we can "over-sample" those small p-values that result in a large $-\log(t)$.

$$\begin{aligned}
\mathbb{P}(T^{AW} > -\log(t)) &= \mathbb{E}[\mathbb{1}(T^{AW} > -\log(t))] \\
&= \int \mathbb{1}(T^{AW} > -\log(t)) \frac{f(t^{AW})}{f^*(t^{AW})} f^*(t^{AW}) dt^{AW} \\
&= \mathbb{E}^*[\mathbb{1}(T^{AW} > -\log(t)) * W(T^{AW}])
\end{aligned}$$

where $W(\cdot) = f(\cdot)/f^*(\cdot)$ and $\mathbb{E}^*(\cdot)$ is the expectation with respect to $f^*(\cdot)$.

Under the null hypothesis, $P_i \sim U(0,1)$, so the joint distribution of $\{P_i\}_{i=1}^K$ is $f(p_1, \ldots, p_k) = 1$. If we instead using Beta$(1, a)$ distribution to draw P_i, the joint distribution of $P_1, \ldots, P_K$ will be $f^*(p_1, \ldots, p_K) = (a-1)^K (p_1 \cdots p_K)^{a-1}$. Optimal a should be chosen such that $\mathbb{P}(T^{AW} > -\log(t))$ has the smallest variance for a given test statistic $-\log(t)$. For a given K, we predefine test statistics $t_1, \cdots, t_{20}$. For each $t_i, i = 1, \cdots, 20$, we select different a and use enough number of samples n such that $\text{sd}(\hat{p}) < \frac{\hat{p}}{100}$.

In Figure 3.2, the p-values in log-scale for $K = 3$ and $K = 20$ are presented and it can be seen that when $-\log(t)$ is large, the p-values in log-scale are almost on a straight line, which

implies that the p-value of a new test statistic may be obtained by interpolation. Therefore, for a given $K \geq 3$, one can generate a table, where the first column are pre-selected test statistics $-\log(t)$ and the second column are the corresponding p-values of the test statistics computed by importance sampling technique. Therefore, when a new test statistic comes, we provide accurate p-value using monotone cubic spline interpolation in log-scale. We will provide an R-package that can generate p-values for $K = 2, 3, \cdots, 1000$.

3.3 ASYMPTOTICAL PROPERTIES OF THE AW-FISHER'S METHOD

In the last section, a fast algorithm of searching the adaptive weights and an accurate and fast method of computing the p-value $\mathbb{P}(T^{AW} > -\log(t))$ are given, which make the AW-Fisher's method more practical and very appealing in real data analysis. In this section we will investigate some asymptotical properties of the AW-Fisher's method, such as consistency of the adaptive weights and the asymptotical Bahadur optimality (ABO).

3.3.1 Assumptions and notations

Suppose we have K independent studies for testing $H_0 : \theta = 0$ with sample size n_k and p-value p_k for $k = 1, \cdots, K$. Assume that for each $k = 1, \cdots, K$, the statistical test for study k has exact slope $c_k(\theta)$, i.e.,

$$-\frac{2}{n_k}\log(p_k) \to c_k(\theta) \text{ as } n_k \to \infty.$$

Obviously by definition, $c_k(\theta)$ is always non-negative. This assumption states that when $c_k(\theta)$ is positive, the p-value p_k will decay to 0 exponentially as n goes to infinity.
Furthermore, we assume

$$\lim_{n\to\infty} \frac{n_k}{n} = \lambda_k \text{ for } k = 1, \cdots, K,$$

where $n = \frac{1}{K}\sum_{k=1}^{K} n_k$.

Obviously n is the averaged sample size and it holds $\sum_{k=1}^{K}\lambda_k = K$ and $-\frac{2}{n}\log(p_k) \to \lambda_k c_k(\theta)$. This assumption guarantees no study will dominate the others and the sample sizes of all the studies will tend to infinity at the same rate.

3.3.2 Consistency of the estimated weights $\hat{\mathbf{W}}$

In this section we will show that the consistency of the estimated adaptive weights $\hat{\mathbf{W}}$, i.e., $\hat{\mathbf{W}} \to \mathbf{W}^*$ with probability one when the averaged sample size n tends to infinity. To this end, we first give the following lemmas.

Without loss of generality, let's assume the $100(1-\alpha)\%$ quantile of χ^2_m is $\chi^{-2}_m(\alpha)$, i.e., $\mathbb{P}(\chi^2_m \geq \chi^{-2}_m(\alpha)) = \alpha$. Obviously $\chi^{-2}_m(\alpha) \to \infty$ as $\alpha \to 0$. The next lemma states that $\chi^{-2}_{2k+2l}(\alpha) > \chi^{-2}_{2k}(\alpha)$ for $k, l \geq 1$ and small α.

Lemma 3.3.1. *For small enough $\alpha > 0$ and $k, l \geq 1$, it holds*

$$\chi^{-2}_{2k+2l}(\alpha) > \chi^{-2}_{2k}(\alpha). \tag{3.3.1}$$

Proof. Note that

$$\begin{aligned}
&\mathbb{P}(\chi^2_{2k+2l} > \chi^{-2}_{2k}(\alpha)) = \int_{\chi^{-2}_{2k}(\alpha)}^{\infty} \frac{x^{k+l-1}e^{-\frac{x}{2}}}{2^{k+l}\Gamma(k+l)}dx = \int_{\chi^{-2}_{2k}(\alpha)}^{\infty} (\frac{x}{2})^l \frac{\Gamma(k)}{\Gamma(k+l)} \frac{x^{k-1}e^{-\frac{x}{2}}}{2^k\Gamma(k)}dx \\
> \;& (\frac{\chi^{-2}_{2k}(\alpha)}{2})^l \frac{\Gamma(k)}{\Gamma(k+l)} \int_{\chi^{-2}_{2k}(\alpha)}^{\infty} \frac{x^{k-1}e^{-\frac{x}{2}}}{2^k\Gamma(k)}dx = (\frac{\chi^{-2}_{2k}(\alpha)}{2})^l \frac{\Gamma(k)}{\Gamma(k+l)} \mathbb{P}(\chi^2_{2k} > \chi^{-2}_{2k}(\alpha)) \\
= \;& (\frac{\chi^{-2}_{2k}(\alpha)}{2})^l \frac{\Gamma(k)}{\Gamma(k+l)}\alpha.
\end{aligned}$$

Therefore, given $k, l \geq 1$, there always exists $\alpha > 0$ such that $(\frac{\chi^{-2}_{2k}(\alpha)}{2})^l \frac{\Gamma(k)}{\Gamma(k+l)} > 1$ and thus $\mathbb{P}(\chi^2_{2k+2l} > \chi^{-2}_{2k}(\alpha)) > \alpha$, which implies $\chi^{-2}_{2k+2l}(\alpha) > \chi^{-2}_{2k}(\alpha)$.

Lemma 3.3.1 shows that $\chi^{-2}_{2k+2l}(\alpha) > \chi^{-2}_{2k}(\alpha)$ holds for some small $\alpha > 0$. Furthermore, the following lemma shows that in fact $\chi^{-2}_{2k+2l}(\alpha) - \chi^{-2}_{2k}(\alpha)$ will tend to infinity as $\alpha \to 0$.

Lemma 3.3.2. *Given significance level $\alpha > 0$ and integers $k, l \geq 1$, it holds*

$$\triangle_{2l} := \chi^{-2}_{2k+2l}(\alpha) - \chi^{-2}_{2k}(\alpha) \to \infty \text{ as } \alpha \to 0.$$

Proof. Recall that the cumulative distribution function of χ^2_k is $F(x,k) = \frac{\gamma(\frac{k}{2}, \frac{x}{2})}{\Gamma(\frac{k}{2})}$, where $\gamma(\cdot,\cdot)$ is the lower incomplete Gamma function. Therefore, it holds

$$\alpha = \mathbb{P}(\chi^2_{2k} \geq \chi^{-2}_{2k}(\alpha)) = 1 - \frac{\gamma(k, \frac{\chi^{-2}_{2k}(\alpha)}{2})}{\Gamma(k)} = e^{-\frac{\chi^{-2}_{2k}(\alpha)}{2}} \sum_{j=0}^{k-1} \frac{(\frac{\chi^{-2}_{2k}(\alpha)}{2})^j}{j!}.$$

Similarly, it holds

$$\alpha = \mathbb{P}(\chi^2_{2k+2l} \geq \chi^{-2}_{2k+2l}(\alpha)) = 1 - \frac{\gamma(k+l, \frac{\chi^{-2}_{2k+2l}(\alpha)}{2})}{\Gamma(k+l)} = e^{-\frac{\chi^{-2}_{2k+2l}(\alpha)}{2}} \sum_{j=0}^{k+l-1} \frac{(\frac{\chi^{-2}_{2k+2l}(\alpha)}{2})^j}{j!}.$$

By setting $\chi^{-2}_{2k+2l}(\alpha) = \chi^{-2}_{2k}(\alpha) + \triangle_{2l}$, it holds $\triangle_{2l} > 0$ for small enough $\alpha > 0$. Furthermore, it is easy to show that

$$e^{-\frac{\triangle_{2l}}{2}} = \frac{\sum_{j=0}^{k-1} \frac{(\frac{\chi^{-2}_{2k}(\alpha)}{2})^j}{j!}}{\sum_{j=0}^{k+l-1} \frac{(\frac{\chi^{-2}_{2k}(\alpha)+\triangle_{2l}}{2})^j}{j!}} \to 0 \text{ as } \alpha \to 0,$$

which proves the lemma.

Recall that $\triangle_{2l} = \chi^{-2}_{2k+2l}(\alpha) - \chi^{-2}_{2k}(\alpha)$ is the discrepancy of $100(1-\alpha)\%$ quantiles between χ^2_{2k+2l} and χ^2_{2k}. Therefore, **Lemma 3.3.2** states that when α is small, in order for χ^2_{2k+2l} to preserve the same significant level α as χ^2_{2k} does, the l newly included studies should have very small p-values $\{p_j\}_{j=k}^{k+l}$ such that $-2\sum_{j=k}^{k+l} \log(p_j) \geq \chi^{-2}_{2k+2l}(\alpha) - \chi^{-2}_{2k}(\alpha)$. Next theorem demonstrates that the adaptive weights $\hat{W}$ of AW-methods are asymptotically consistent.

Theorem 3.3.3. *$\hat{W} \to W^*$ as $n \to \infty$, i.e. all and only the studies with non-zero effect sizes will contribute to the test statistic, and the convergent rate of $\hat{W}$ is of the order $\mathcal{O}(n^{-1})$.*

Proof. Let $-\frac{2}{n}\sum_{j=1}^{i}\log(p_j) \to \sum_{j=1}^{i}\lambda_j c_j(\theta) = C_i$ for $i = 1, \ldots, K$. Suppose i studies have weight 1.

Firstly, if $\theta_{i+1} \neq 0$,

$$\lim_{n\to\infty}\frac{1-\chi^2_{2i}(-2\sum_{j=1}^{j}\log(p_j))}{1-\chi^2_{2(i+1)}(-2\sum_{j=1}^{i+1}\log(p_j))} \to \lim_{n\to\infty}\frac{1-\chi^2_{2i}(2nC_i)}{1-\chi^2_{2(i+1)}(2nC_{i+1})}$$

$$\to \lim_{n\to\infty}\frac{iC_i^{i-1}}{iC_{i+1}^{i-1}+nC_{i+1}^{i}}\exp\{n(C_{i+1}-C_i)\} \to \infty$$

since $C_{i+1}-C_i = \lambda_{i+1}c_{i+1} > 0$ and $1-\chi^2_{2i}(2nC_i) = \mathbb{P}(\chi^2_{2i} > 2nC_i)$. Therefore study $i+1$ will eventually get a weight 1. Therefore, we have shown that if $\theta_{i+1} \neq 0$, including the $(i+1)$th study into analysis results in a more significant p-value which converges to 0 at accelerated convergent rate $n_{i+1}\exp\{-n_{i+1}c_{i+1}(\theta)\}$.

Secondly, if there exists a study with zero effect size that has a weight 1. Without loss of generality, let $\theta_i = 0$. In order to have weight 1 for study i, one must have

$$1-\chi^2_{2i}(-2\sum_{j=1}^{i}\log p_j) \leq 1-\chi^2_{2(i-1)}(-2\sum_{j=1}^{i-1}\log p_j).$$

Since

$$\chi^2_{2i}(t) = 1-\sum_{j=0}^{i-1}\frac{1}{j!}\left(\frac{t}{2}\right)^j\exp\{-\frac{t}{2}\},$$

we have

$$\sum_{j=0}^{i-1}\frac{1}{j!}\left(-\sum_{j=1}^{i}\log p_j\right)^j\exp\{\sum_{j=1}^{i}\log p_j\}\leq\sum_{j=0}^{i-2}\frac{1}{j!}\left(-\sum_{j=1}^{i-1}\log p_j\right)^j\exp\{\sum_{j=1}^{i-1}\log p_j\}.$$

Thus,

$$p_i=\exp\{\log p_i\}\leq\frac{\sum_{j=0}^{i-2}\frac{1}{j!}\left(-\sum_{j=1}^{i-1}\log p_j\right)^j}{\sum_{j=0}^{i-1}\frac{1}{j!}\left(-\sum_{j=1}^{i}\log p_j\right)^j}\to n^{-1}(i-1)C_{i-1}^{i-2}C_i^{-i+1}\to n^{-1}(i-1)C_i^{-1},$$

as $n\to\infty$, i.e. $p_i\leq n^{-1}(i-1)C_i^{-1}\to 0$. Therefore, eventually, no studies with zero effect size will have weight 1. Now we have proven that the adaptive weights are asymptotically consistent. Since the convergent rate of the adaptive weights depends on how fast one assigns weights 0 to the studies with zero effect size, one concludes from the above discussion that the convergence rate of the adaptive weights is of order $\mathcal{O}(n^{-1})$.

3.3.3 The asymptotic Bahadur optimality (ABO) of AW-Fisher's method

In last subsection we proved that the adaptive weights are asymptotically consistent. In this subsection we will investigate the optimality of the AW-Fisher's method. One way to compare the performance of different tests is the Bahadur relative efficiency (Bahadur 1967), which is defined as the ratio of the exact slopes of different statistical tests and therefore the test with larger exact slope is viewed as superior. In this paper we will use the Bahadur relative efficiency as our primary index of comparing combined p-value procedures. In fact, by assumption $-\frac{2}{n}\log(p_i)\to c_i(\theta)$ as $n\to\infty$, the exact slope measures how quickly the p-value of a test converges to 0 as n tends to infinity when $\theta\neq 0$, and therefore it can be used to compare the performances of different tests. Assuming there are two tests for testing the same hypotheses and have exact slopes $c_1(\theta)$ and $c_2(\theta)$ respectively, then the ratio $c_1(\theta)/c_2(\theta)$

is the exact Bahadur efficiency of test 1 relative to test 2, and $c_1(\theta)/c_2(\theta) > 1$ implies that p_1 converges to 0 faster than p_2 for large enough n. Note that the exact Bahadur efficiency is asymptotic property of two tests, so $c_1(\theta)/c_2(\theta) > 1$ doesn't imply $p_1 < p_2$ for small n.

Lemma 3.3.4. *For $\theta = 0$, it holds $-\frac{2}{n}\log(p) \to 0$ with probability one, i.e., if the effect size is 0, the exact slope $c(\theta)$ of the statistical test is 0.*

Proof. Recall that for $\theta = 0$, the p-value p is distributed uniformly on the interval $(0, 1)$. So it holds

$$\mathbb{E}(-\frac{2}{n}\log(p)) = -\frac{2}{n}\int_0^1 \log(x)dx = \frac{2}{n} \to 0. \tag{3.3.2}$$

Since $-\frac{2}{n}\log(p)$ is always positive for $p \in (0, 1)$, one concludes that

$$-\frac{2}{n}\log(p) \to 0, \tag{3.3.3}$$

i.e., $c(0) = 0$.

Lemma 3.3.4 states that non-significant studies have no impact on the exact slope of combined p-value procedures which fits our intuition.

Littel and Folk had shown that given K independent studies with p-values, sample sizes and exact slopes $\{(p_k, n_k, c_k(\theta))\}_{k=1}^K$ respectively, the exact slope of Fisher's method is $c_{Fisher}(\theta) = \sum_{k=1}^K \lambda_k c_k(\theta)$ (Littel and Folk, 1971) and the Fisher's method is ABO, i.e., $c_{Fisher}(\theta)$ is the largest among all combining p-value procedures (Littel and Folk, 1973), under the assumption $\theta_k \equiv \theta \neq 0$. This assumption is very stringent, since the global null hypothesis $H_0 : \bigcap_{k=1}^K \{\theta_k = 0\}$ will be rejected if at lease one θ_k is nonzero. In this paper we will consider the alternative hypothesis $H_a : \bigcap_{k=1}^r \{\theta_k \equiv \theta \neq 0\}, \bigcap_{j=r+1}^K \{\theta_j = 0\}$. Therefore in this case, when the null hypothesis is false, the exact slope of Fisher's method is $c_{Fisher}(\theta) = \sum_{k=1}^K \lambda_k c_k(\theta) =$

$\sum_{k=1}^{r} \lambda_k c_k(\theta)$, which is still ABO. In the sequel we will investigate if the AW-Fisher's method is ABO under this assumption.

In the sequel we will calculate $c_{AW}(\theta)$. Next theorem shows that $c_{AW}(\theta) = c_{Fisher}(\theta)$, i.e., the AW-Fisher's method is ABO.

Theorem 3.3.5. *It holds $c_{AW}(\theta) = c_{Fisher}(\theta)$, i.e., the AW-Fisher's method is ABO.*

Proof. Let $\{p'_1, \ldots, p'_j\}$ is a subset of $\{p_1, \ldots, p_K\}$ with size j for $j = 1, \ldots, K$, then for a given test statistic $-\log(t)$, it holds

$$\begin{aligned}
\mathbb{P}(T^{AW} \geq -\log(t)) &= \mathbb{P}(\bigcup_{j=1}^{K} \{-2 \sum_{i=1}^{j} \log(P'_i) \geq \chi^{-2}_{2j}(t))\}) \\
&\leq \bigcup_{j=1}^{K} \mathbb{P}(\{-2 \sum_{i=1}^{j} \log(P'_i) \geq \chi^{-2}_{2j}(t)\}) \\
&= (2^K - 1)t
\end{aligned}$$

as n goes to infinity.

Since the adaptive weights are consistent, it holds

$$\lim_{n\to\infty} -\frac{2}{n}\log(t) = \lim_{n\to\infty} -\frac{2}{n}\log(1 - \chi^2_{2r}(-2\sum_{i=1}^{r}\log(p_i))) = \sum_{i=1}^{r} \lambda_i c_i(\theta),$$

we have

$$\lim_{n\to\infty} -\frac{2}{n}\log(\mathbb{P}(T^{AW} \geq -\log(t))) \geq \lim_{n\to\infty} -\frac{2}{n}\{\log(t) + \log(2^K - 1)\} = \sum_{i=1}^{r} \lambda_i c_i(\theta).$$

On the other hand, since

$$\mathbb{P}(T^{AW} \geq -\log(t)) \geq \mathbb{P}(\{-2\sum_{i=1}^{r}\log(P_i) \geq \chi^{-2}_{2r}(t)\}) = t,$$

by utilizing the consistency of the adaptive weights again, we have

$$\lim_{n\to\infty} -\frac{2}{n}\log(\mathbb{P}(T^{AW} \geq -\log(t))) \leq \lim_{n\to\infty} -\frac{2}{n}\log(t) = \sum_{i=1}^{r} \lambda_i c_i(\theta),$$

i.e., $c_{AW}(\theta) = c_{Fisher}(\theta)$ and so AW-Fisher's method is ABO too under the alternative hypothesis $H_a : \bigcap_{k=1}^{r}\{\theta_k \equiv \theta \neq 0\}, \bigcap_{j=r+1}^{K}\{\theta_j = 0\}$.

3.4 SIMULATIONS

3.4.1 ABO of AW-Fisher's method

In this subsection, the following hypotheses $H_0 : \theta_1 = \theta_2 = 0$ vs. $H_A : \theta_1 = 0.3, \theta_2 = 0$ for $K = 2$ and $H_0 : \theta_1 = \theta_2 = \theta_3 = 0$ vs. $H_A : \theta_1 = 0.3, \theta_2 = \theta_3 = 0$ for $K = 3$ were tested. We expect the estimated exact slopes of AW-Fisher's method and the Fisher's method are similar, provided the sample sizes are large enough. The sample sizes were chosen as $\{20, 30, 50, 75, 100, 125, 150, 175, 200, 250, 300, 350,$
$400, 450, 500, 750, 1000, 2000, 3000, 4000, 5000, 7500, 10000, 12500, 15000, 17500, 20000\}$. For each sample size, 5000 simulations were performed and the mean exact slopes were calculated.

Figure 3.3 shows clearly that the AW-Fisher's method is ABO, since the exact slopes of AW-Fisher's method and Fisher's method converge as n tends to infinity.

3.4.2 Comparison of AW-Fisher and Fisher's method

In this subsection, $K = 20$ independent studies are combined, in which 10 studies have zero effect sizes and 10 studies have nonzero effect sizes. Figure 3.4 compares the p-values

of AW-Fisher and Fisher's method and the plot shows that in most cases the AW-Fisher's method will produce more significant p-values.

Table 3.1 summarizes the powers that AW-Fisher and Fisher's method achieve at different significance levels α. The results shows that AW-Fisher is powerful than Fisher's method.

3.4.3 Accuracy of importance sampling algorithm

In this chapter, we proposed an importance sampling algorithm to obtain accurate p-values for $K \geq 2$. Since the close form of p-value calculation is available for $K = 2$, in this section, we will compare the p-values generated by the close form and by the importance sampling algorithm for $K = 2$ for validation. Figure 3.5 shows the scatter plots of log-scaled p-values generated by close form and importance sampling algorithms for three hypotheses setting where 0, 1 or 2 studies have non-zero effect sizes. The plots indicate that the importance sampling algorithm can reach to the accuracy of 10^{-30} which validates our numerical algorithm for p-value approximation.

3.5 DISCUSSION AND CONCLUSION

The AW-Fisher's method proposed in Li & Tseng 2011 possesses good properties such as admissibility and better overall power compared to minP, maxP and Fsiher's method. Furthermore, the adaptive weights give the indication which studies contribute to the statistical significance, which makes it more appealing in genomic data analysis.

In genomic meta-analysis, it is very crucial to have a fast and accurate algorithm to reduce the computing cost and increase the reliability of the results. However, in Li & Tseng 2011, the computational complexity of searching the adaptive weights is exponential and the p- values were evaluated by permutation tests, which is inaccurate and inefficient. Furthermore, some asymptotical properties such as the consistency of the weights and the asymptotic Bahadur optimality (ABO) remain open in Li & Tseng (2011).

In this chapter, we investigated the theoretical questions and concluded that the adaptive weights are consistent and the AW-Fisher is also ABO. Furthermore, based on order statistic technique, we successfully reduce the complexity of searching the best weights from exponential to linear. Now the total cost of the fast searching algorithm involves sorting the vector of K p-values and searching k non-zero vector of weights, which results in an overall complexity of at most $\mathcal{O}(K^2)$ depending on the sorting algorithm used. The close form of the p-value computation is derived for $K = 2$, and for $K \geq 3$, an importance sampling algorithm was proposed to evaluate the p-values. Since the p-values are almost linear to large test statistics in log-scale, for each $K \geq 3$, a table containing a column of test statistics and their corresponding p-values can be generated in a very accurate manner and then saved for later use. When a new statistic is coming, one can compute the corresponding log-scaled p-value by cubic interpolation, which is very efficient and accurate.

Since the AW-Fisher has good theoretical properties and the computation of the p-values is fast and accurate, we expect it will find more applications in the future.

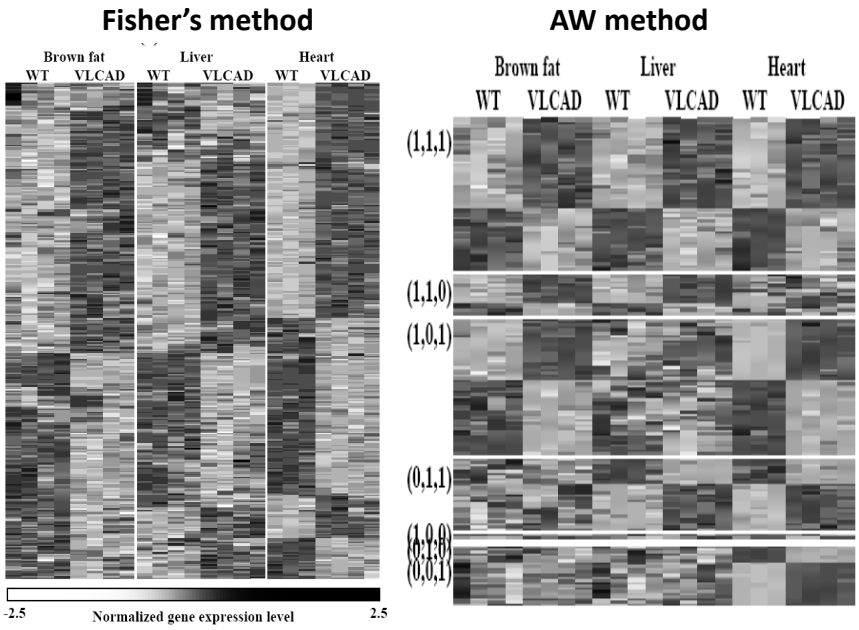

Figure 7: Heatmaps of gene expressions for DE genes identified by Fisher's and AW-Fisher's methods in the mouse energy metabolism datasets.

Table 9: Toy example of finding the adaptive weights

Weights	Weighted statistic	Null distribution	p-value
(1,1,1)	13.82	χ^2_6	0.032
(1,1,0)	0	χ^2_4	1
(1,0,1)	13.82	χ^2_4	0.008
(0,1,1)	13.82	χ^2_4	0.008
(1,0,0)	0	χ^2_2	1
(0,1,0)	0	χ^2_2	1
(0,0,1)	13.82	χ^2_2	0.001

Table 10: Comparison of complexities $2^K - 1$ vs. K. Total cost: sorting (at most $O(K^2)$) and linear searching ($O(K)$)

K/Methods	exponential searching	Linear searching
2	3	2
3	7	3
4	15	4
5	31	5
6	63	6
7	127	7
8	255	8
9	511	9
10	1023	10
11	2047	11
12	4095	12
13	8191	13
14	16383	14
15	32767	15

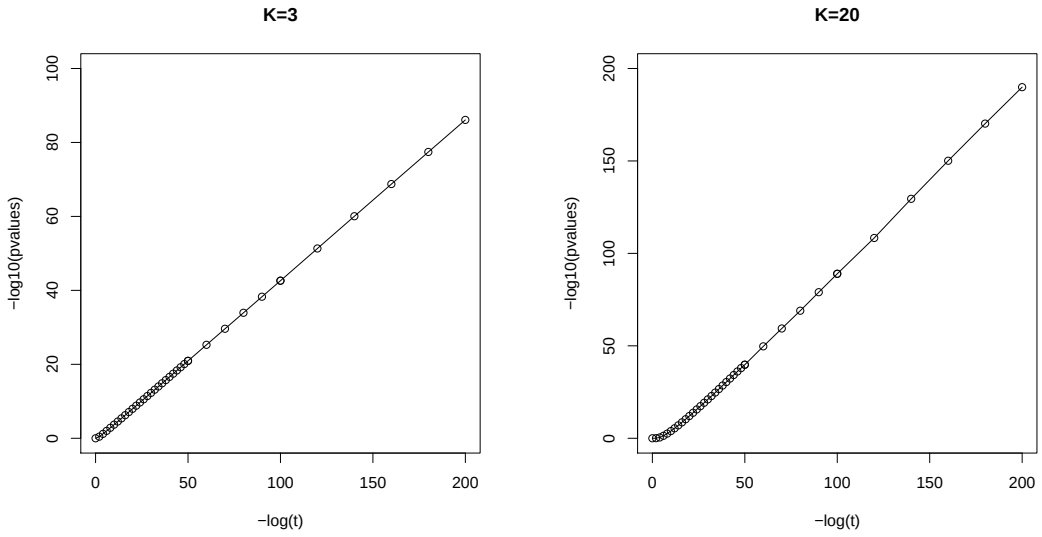

Figure 8: p-values of AW-Fisher's method in log scale for $K = 3$ and $K = 20$

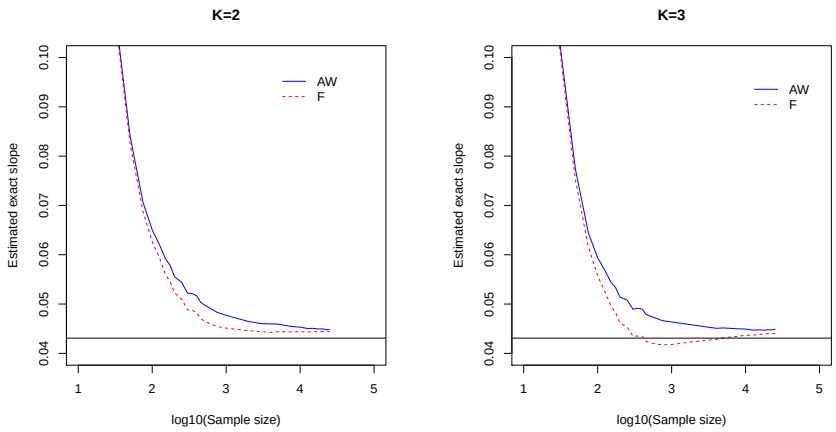

Figure 9: Comparison of the approximated exact slopes for AW-Fisher and Fisher's method for $K = 2$ and $K = 3$. Only the first study has non-zero effect size 0.3.

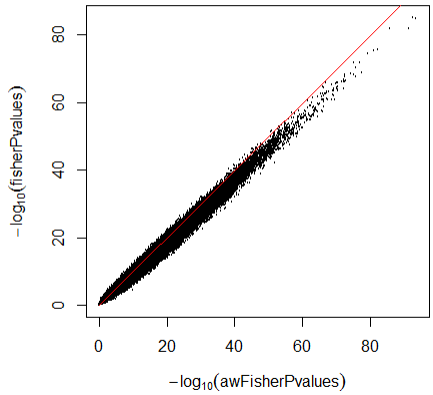

Figure 10: Comparison of the p-values of AW-Fisher and Fisher's method

Table 11: Powers of AW-Fisher and Fisher's method at different significance levels α

Method/α	10^{-3}	10^{-5}	10^{-7}	10^{-9}
AW	0.98	0.93	0.86	0.77
Fisher	0.97	0.91	0.82	0.72

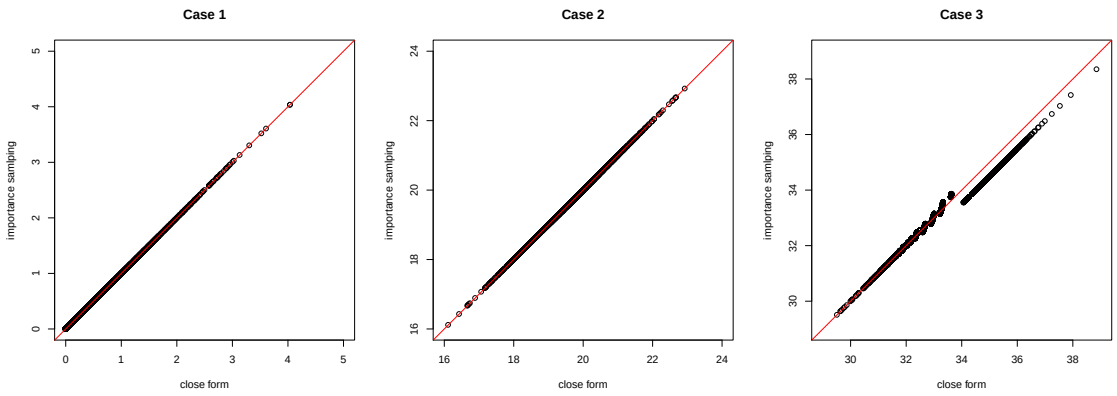

Figure 11: Comparison of the p-values in log-scale. Case 1: $P_1, P_2 \sim \text{Uniform}(0,1)$; Case 2: $P_1 \sim \text{Uniform}(0,1), P_2 \sim \text{Beta}(1, 10^{20})$; Case 3: $P_1 \sim \text{Beta}(1, 10^{15}), P_2 \sim \text{Beta}(1, 10^{20})$.

4.0 CONCLUSION AND FUTURE WORKS

4.1 CONCLUSION

The rapid development of high-throughput experimental technology in the past decade has made the generation of genomic data more and more affordable. However, the sample size in each individual study is usually small and therefore using meta-analysis to combine multiple genomic studies can increase the statistical power and make it possible to draw a stronger scientific statement. In this dissertation, I proposed a series of tools for univariate meta-analysis methods which can be easily applied to genomic data.

In chapter 2, I proposed three imputation methods for evidence aggregation p-value combination methods. In order to apply the conventional p-value combination methods such as the Fisher's method or the Stouffer's method, all the p-values should be known in advance. However, in many medical publications, usually only a partial DE gene list was reported with respect to a given p-value threshold, which results in the artificially "censored" p-values and the conventional meta-analysis methods can not be applied directly to those "censored" p-values. Based on the null hypothesis that the true effect size is zero and thus the p-value follows uniform distribution on $[0, 1]$, I proposed three imputation methods (the mean imputation, the single random imputation and the multiple imputation) and imputed the "censored" p-values accordingly. The conventional evidence aggregation meta-analysis methods now can be applied to the imputed p-values. In order to compute the p-values accurately, the analytical null distributions of the test statistic were derived for

the mean imputation and the single random imputation methods and the asymptotical null distribution was derived for the multiple imputation. The proposed imputation methods were compared with the available case method for simulated expression profiles and two real prostate cancer and MDD data. The results indicate that the proposed imputation methods have higher DE gene detection capacity than the available method and among the imputation methods, the mean imputation and the multiple imputation methods are more powerful than the single random imputation method. The proposed imputation methods were also applied to two real examples in colorectal cancer and pain data and the results indicate that both mean imputation and multiple imputation performed among the best in terms of detection capacity and biological validation by pathway analysis. The mean imputation method is recommended in practice, since it possesses similar detection capacity as the multiple imputation method does and it requires lower computational burden.

In chapter 3, I revisited the AW-Fisher's method. The AW-Fisher's method aims to assigning weight 1 to the studies with non-zero effect size and assigning 0 to the studies with zero effect size. Therefore the estimated adaptive weights provide insights on which studies contribute to the statistical significance. In chapter 3, I investigated some asymptotical properties of the AW-Fisher's method such as the consistency of the adaptive weights and asymptotical Bahadur optimality (ABO). Furthermore, two computing issues were discussed. Firstly, I provided a fast algorithm basing on ordered p-values to search for the adaptive weights and the computational complexity of total cost is reduced from $\mathcal{O}(2^K)$ to at most $\mathcal{O}(K^2)$ (at most $\mathcal{O}(K^2)$ for sorting the p-values and $\mathcal{O}(K)$ for searching the best weights). Secondly, I provided a closed form to compute the p-values for $K = 2$ and an importance sampling algorithm was proposed to compute the p-values for $K \geq 3$. Since for each given $K \geq 3$, the p-values in log-scale are almost linear for large test statistic, we provided a fast algorithm to compute the p-value for a given test statistic. We firstly generate a table of some preselected test statistics and their corresponding p-values by importance sampling algorithm, then for any new test statistic we approximate the p-value by monotone cubic interpolation. The fast algorithm was compared to the closed form for $K = 2$ and the results indicate that the numerical approximation can reach the precision of 10^{-30} which is accurate enough in

practice. The fast algorithms make the AW-Fisher's method very useful and appealing in practice.

4.2 FUTURE WORKS

In the future, I intend to further extend my current work in several ways, including methodology and software.

Firstly, I will polish my R codes and make it easy to use. An R package will be developed for the methods discussed in this dissertation. And the package will be submitted to the comprehensive R archive network (CRAN, http://cran.r-project.org).

Secondly, I will extend my methodologies into the vertical integration of multiple data types. So far, I mainly focus on the horizontal genomic meta-analysis, i.e., combining multiple relevant studies (e.g. microarray or GWAS) to increase statistical power. It should be noted that recent technology has made it possible to simultaneously perform multi-platform genomic profiling of biological samples and produce multi-dimensional genomic data. Such data describe the biological properties of the same cohort from different angles and it provides great opportunity to study the coordination between regulatory mechanisms on multiple levels. With the rapid decline of sequencing costs, such data will soon accumulate rapidly and the need for integrating the information contained in the multi-dimensional genomic data will increase. Since existing tools are designed for one-dimensional or at most two-dimensional genomic data, novel methods, for example, vertical genomic integrative analysis, for analyzing multi-dimensional datasets for effective information extraction and hypothesis testing should be developed.

Unlike horizontal genomic meta-analysis, the vertical genomic integrative analysis, which integrates multiple studies that measure multiple dimension of genetic information of the

same cohort (e.g. transcription, genotyp- ing, copy number variation, methylation, miRNA etc), is more challenging. This is a relatively new research filed and it contains rich topics and open questions waiting for researchers to solve. In future research I will devote myself to methodology research of the vertical genomic integrative analysis and their applications.

BIBLIOGRAPHY

[1] 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*, **57**: 289-300.

[2] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**: 1165-1188.

[3] Berger R.L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, **24**: 295-300.

[4] Bertucci F., Salas S., Eysteries S., et al. (2004). Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters. *Oncogene*, **23**: 1377-91.

[5] Bianchini M., Levy E., Zucchini C., et al. (2006). Comparative study of gene expression by cDNA microarray in human colorectal cancer tissues and normal mucosa. *Int J Oncol*, **29**: 83-94.

[6] Birnbaum A. (1954). Combining independent tests of significance. *Journal of the American Statistical Association*, **49**: 559-574.

[7] Birnbaum A. (1955). Characterizations of complete classes of tests of some multiparametric hypothesis, with applications to likelihood ratio tests. *ANN. Math. Statist.*, **26**: 21-36.

[8] Bellot G.L., Tan W.H., Tay L.L., Koh D. et al. (2012). Reliability of tumor primary cultures as a model for drug response prediction: expression profiles comparison of tissues versus primary cultures from colorectal cancer patients. *J Cancer Res Clin Oncol*, **138(3)**: 463-482.

[9] Borovecki F. et al (2005). Genome-wide expression profiling og human blood reveals biomarkers for huntingtons disease. *Proceedings of the National Academy of Sciences*, **102**: 11023-11028.

[10] Cardoso J. et al. (2007). Expression and genomic profiling of colorectal cancer. *Biochimica et Biophysica Acta-Reviews on Cancer*, **1775**: 103-137.

[11] Chan S.K., Griffith O.L., Tai I.T. and Jones S.J.M. (2008). Meta-analysis of colorectal cancer gene expression profiling studies identifies consistently reported candidate biomarkers. *Cancer Epidemiol Biomarkers Prev*, **17(3)**: 543-552.

[12] Chang L.C., Lin H.M., Sibille E. and Tseng G.C. (2013). Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. *BMC Bioinformatics*, **14**: 368.

[13] Choi J.K. et al. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19**: 84-90.

[14] Choi H. et al. (2007). A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. *BMC Bioinformatics*, **8**: 364-383.

[15] Cooper H.M. and Hedges L.V. (Eds.) (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

[16] Croner RS, Foertsch T, Brueckl WM, et al. (2005). Common denominator genes that distinguish colorectal carcinoma from normal mucosa. *Int J Colorectal Dis*, **20**: 353-362.

[17] Duval S. and Tweedie R.L. (2000a). Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, **56(2)**: 455-463.

[18] Duval S. and Tweedie R.L. (2000b). A nonparametric "Trim and Fill" method of accounting for publication bias in meta-analysis. *JASA*, **95(1)**: 89-98.

[19] Fisher R.A. (1932). *Statistical methods for research workers.* Edinburgh, Oliver & Boyd, 4th edition.

[20] Fleiss J.L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.),*The handbook of research synthesis* (pp. 245-260), New York: Russell Sage Foundation.

[21] Folks J.L. (1984). Combination of independent tests. In *Handbook of statistics 4. Nonparametric methods*, P.R. Krishnaiah and P.K. Sen (eds):New York, North-Holland

[22] Gorlov I.P. et al. (2009). Candidate pathways and genes for prostate cancer: a meta-analysis of gene expression data. *BMC Medical Genomics*, **2**:48.

[23] Grade M., Hormann P., Becker S., et al. (2007). Gene expression profiling reveals a massive, aneuploidy-dependent transcriptional deregulation and distinct differences between lymph node-negative and lymph node-positive colon carcinomas. *Cancer Res*, **67**:41-56.

[24] Griffith O.L., Jones S.J.M. and Wiseman S.M. (2006). Meta-analysis and meta-review of thyroid cancer gene expression profiling studies identifies important diagonstic biomarkers. *J. Clin. Oncol.*, **24**:5043-51.

[25] Hedges L.V. and Olkin I. (1980). *Vote-counting methods in research synthesis. Psychological Bulletin*, **88**:359.

[26] Hedges L.V. and Olkin I. (1985). *Statistical methods for meta-analysis.* Academic Press Inc.: Orlando, Florida.

[27] Hedges L.V. and Vevea J.L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, **3**:486-504.

[28] Hedges L.V. (2007). Meta-analysis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp.919-953), New York: Russell Sage Foundation.

[29] Hunter J.E. and Schmidt F.L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment*, **8**:275-292.

[30] Ioannidis J.P.A., Allison D.B., Ball C.A. and et. al. (2009). Repeatability of published microarray gene expression analysis. *Nature Genetics*, **41**:149-155.

[31] Jiang X., Tan J., Li J., Kivimäe S. et al. (2008). DACT3 is an epigenetic regulator of Wnt/beta-catenin signaling in colorectal cancer and is a therapeutic target of histone modifications. *Cancer Cell*, **13(6)**:529-41.

[32] Kim H., Nam S.W., Rhee H., et al. (2004). Different gene expression profiles between microsatellite instability-high and microsatellite stable colorectal carcinomas. *Oncogene*, **23**: 6219-25.

[33] Kwon H.C., Kim S.H., Roh M.S., et al. (2004). Gene expression profiling in lymph node-positive and lymph node-negative colorectal cancer. *Dis Colon Rectum*, **47**: 141-152.

[34] LaCroix-Fralish M.L. et. al.. (2011). Patterns of pain: Meta-analysis of microarray studies of pain. *Pain*, **152**: 1888-1898.

[35] Lancaster H. (1961). The combination of probabilities: an application of orthonormal functions. *Australian Journal of Statistics*, **3**: 20-33.

[36] Littell C.L. and Floks J.L.. (1971). Asymptotic Optimality of Fisher's Method of Combining Independent Tests. *Journal of the American Statistical Association*, **66(336)**: 802-805.

[37] Littell C.L. and Floks J.L.. (1973). Asymptotic Optimality of Fisher's Method of Combining Independent Tests II. *Journal of the American Statistical Association*, **68(341)**: 193-194.

[38] Lau J., Antman E.M. and Jimenez-Silva J. et. al. (1992). Cumulative meta-analysis of therapeutic trials for Myocardial infarction. *The new England Journal of Medicine*, **327**:248-254.

[39] McCarley R.W., Wible C.G., Frumin M., Hirayasu Y., Levitt J.J., Shenton M.E. (2001). Why vote-count reviews don't count [letter to the editor]. *Biological Psychiatry*, **49**: 161-163.

[40] Li J. and Tseng G.C. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *Annals of Applied Statistics*, **5(2A)**:994-1019.

[41] Littell R.C. and Folks J.L.(1971). Asymptotic optimality of Fisher's method of combining independent tests. *Journal of the American Statistical Association*, **66**: 802-806.

[42] Littell R.C. and Folks J.L. (1973). Asymptotic optimality of Fisher's method of combining independent tests ii. *Journal of the American Statistical Association*, **68**: 193-194.

[43] Moreau Y. et al. (2003). Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends in Genetics*, **19**: 570-577.

[44] Olkin I. and Saner H. (2001). Approximations for trimmed Fisher procedures in research synthesis. *Statist. Methods Med. Res.*, **10**: 267-276.

[45] Owen A.B. (2009). Karl pearson's meta-analysis revisited. *Annals of Statistics*, **37**: 3867-3892.

[46] Pirooznia M., Nagarajan V. and Deng Y. (2007). Gene venn - a web application for comparing gene lists using venn diagram. *Bioinformation*, **1**: 420-422.

[47] Rhodes D. et al. (2002). Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer research*, **62**: 4427-4433.

[48] Rosenthal R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.),*The handbook of research synthesis* (pp. 231-244), New York: Russell Sage Foundation.

[49] Roy S.N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics*, **24(2)**: 220-238.

[50] Little R. and Rubin D. (2002). *Statistical analysis with missing data, second edition*. John Wiley & Sons, Inc.: Hoboken, New Jersey.

[51] Segal E. et al. (2004). A module map showing conditional activity of expression modules in cancer. *Nature Genetics*, **3**: 1090-1098.

[52] Song C. and Tseng G.C. (2014). Order statistic for robust genomic meta-analysis. *Annals of Applied Statistics*. Accepted.

[53] Stouffer S. et al. (1949). *The American solider, volumn I: adjustement during army life*. Princeton University press.

[54] Sterne J. (editor) (2009). *Meta-analysis in Stata: an updated collection from the Stata Journal*. Stata press.

[55] Tippett L.H.C. (1931). *The methods in statistics*. Williams and Norgate, LTD., 1st edition.

[56] Tseng G.C. et al. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research*, **40(9)**:3785-99.

[57] Wang X., Lin Y., Song C., Sibille E. and Tseng G.C. (2012). Detecting disease-associated genes with confounding variable adjustment and the impact on genomic meta-analysis: with application to major depressive disorder. *BMC Bioinformatics*, **13**: 52.

[58] Wilkinson B. (1951). A statistical consideration in psychological research. *Psychological Bulletin*, **48**: 156-157.