Statistical methods for gene set based significance analysis

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Sang Mee Lee

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

Doctor of Philosophy

July, 2011

Acknowledgements

First and foremost, praises and thanks to God, the Almighty, for having made everything possible by giving me wisdom, strength and courage to complete this work. I love Him deeply and more than words can express.

I would like to express my deep and sincere gratitude to my thesis advisor, Dr. Baolin Wu, for providing invaluable guidance throughout this research. It was a great privilege and honor to work and study under his guidance.

I am deeply indebted to my co-advisor Dr. Wei Pan whose help and encouragement motivated me at all times of my graduate study.

I would also like to thank the other members of my committee, Dr. Cavan Reilly and Dr. Xiaotong Shen for their encouragement, insightful comments, and hard questions.

Special thanks are also extended to all other faculty members, staff, and fellow students in the Division of Biostatistics, University of Minnesota for their help and friendship, which allowed me to devote to and to enjoy my study.

My heartfelt thanks to my fellows of the Korean Presbyterian Church of Minnesota, and to all members of the Minnesota United Prayer Meeting for their prayer, love and support.

I owe my loving thanks to my family who gave me their unconditional support and encouragement throughout. My special gratitude is due to my nephew, Chang-Sub. They all let me own a happy family.

Abstract

Gene set enrichment analysis (GSEA) is a method to identify groups of genes, which are statistically more differentially expressed than all other genes across different treatments within a microarray study. Most of the existing approaches have largely relied on nonparametric methods and require repeated computation of permutation and resampling data to assess the significance of a gene set. In this dissertation, we study parametric approaches for GSEA by formulating the enrichment analysis into a simple model comparison problem. The methods not only gain the flexibility in statistical modeling corresponding to biological problems but also achieve computational efficiency.

First, we propose a likelihood based approach assuming a finite mixture model for a two-class comparison problem and the implementation of the analysis is achieved by a likelihood ratio based testing approach. In addition we extend the parametric methods to flexible two-component mixture models for one-sided enrichment analysis which aims to test for enrichment of up (or down) regulation only. Also, we develop chi-square mixture models which incorporate the idea of two-class comparison studies into multiple category microarray experiments. Applications to gene expression data, along with simulations, demonstrate the computational efficiency and the competitive performance of the proposed methods.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Microarray gene expression data are becoming routinely used in biomedical research, bearing the hope of molecularly diagnosing and developing more effective therapeutic treatments for various diseases. The common feature of these data is typically that the number of observed samples is much smaller than the number of genes. Detecting individual genes that are differentially expressed across several states of interest has been the most commonly used gene expression analysis. One of the important statistical issues associated with differential expression detection for large scale microarray data lies in the extreme multiple testing: typically tens of thousands of genes are tested simultaneously and many false positives are likely to be identified just by chance. To address the issue, a variety of statistical methods has been proposed in the literature including the frequentist [55, 13, 54, e.g.], Bayesian [27, 53, 28, 39, 26, e.g.], and empirical Bayes approaches [23, 14, 30, e.g.] etc. Many have proposed shrunken estimation [55, 62, 10, e.g.] or empirical Bayes modeling [34, 48, e.g.] approaches to stabilize the variance estimate and improve the gene selection power. Most existing methods have calculated a (often univariate) summary statistic for each gene, which are then modeled for inference. For example, Dudoit *et al.* [13] proposed to use the t-statistic with a permutation test to rank genes. Tusher *et al.* [55] proposed the regularized t-statistic for gene ranking motivated by the commonly observed large variation associated with gene expression. Efron [14] has proposed to model the distribution of ordinary or regularized t-statistics over all genes non-parametrically and make inference using the empirical Bayes approach.

Among existing methods, the empirical Bayes approach (EB) has proven very useful for studying the simultaneous significance testing problems commonly encountered in analyzing current large-scale microarray gene expression data and has been studied extensively. For example, Kendziorski *et al.* [30] and Smyth [48] have proposed parametric EB approaches with different gene expression distribution assumptions. Efron

[23, 14, 17] has proposed and studied in detail the nonparametric EB approach. The nonparametric EB method models summary statistics across genes, which are often some univariate statistics (e.g., ordinary/moderated t/F-statistics), to borrow information for improving the individual gene inference and overall detection power.

The finite mixture model is also a popular approach for differential gene expression detection in analyzing microarray data, and has been widely used due to its ease of implementation and interpretability. For example, Pan *et al.* [42] and McLachlan *et al.* [36] proposed the two-component normal mixture with unknown variance for modeling null and differentially expressed genes. Most normal mixture models essentially try to approximate some transformed Z-scores (e.g., the two-sample t-statistics) with a normal distribution. Jiao and Zhang [29] proposed to use the t-mixture with unknown degrees of freedom and variance for more robust inference of differential expression, and showed its competitive performance compared to the normal mixture model. Very often these finite mixture modeling approaches have ignored the effect size distribution of gene expressions and instead directly modeled the Z-scores with some convenient distribution assumptions. Some parametric and nonparametric modeling approaches have been proposed to explicitly model the effect size distribution in detecting differential expressions. For example, Wu *et al.* [63] proposed a multi-component normal mixture with constant variance equal to one for analyzing the two-sample t-statistics that explicitly models the effect size distribution and captures the heterogeneity of differentially expressed genes. It is well known that the normal distribution might be a very poor approximation for analyzing Z-scores with relatively small sample size microarray data. Ruppert *et al.* [45] proposed a very novel semi-parametric approach with nonparametric B-splines modeling of the gene expression effect size distribution for differential expression detection. A quadratic programming procedure is proposed to solve the semi-parametric model that might require intensive computations.

The classical method for significance testing of microarray data is to test one gene at a time, to compute a p-value for each gene, and then to adjust for multiple comparisons through controlling the family-wise error rate (FWER) or false discovery rate (FDR, [4]). Although this single gene oriented analysis gives many important insights, it has a few limitations [49]. A number of genes which contributes to subtle changes in expression may not be detected because the cut-off is determined after a correction for multiple testing. On the other hand, statistical analysis results in a long list of significant genes. It is a burden on biologists to interpret and figure out any genetic patterns. Also, the gene-specific analysis method may ignore a critical function of biological processes such as metabolic pathways and transcriptional programs that are distributed across an entire network of genes, yet subtle at the level of individual genes. Often a set of genes together influence a biological process. Recently many researchers have proposed methods to address these analytical challenges of a gene-specific manner. These approaches are based on gene sets which have already been annotated by functional categories, and accordingly they yield more biologically interpretable results. Now gene expression microarray analysis is assessed at the level of gene sets.

Gene sets often contain a collection of genes that work together to affect the system function. Gene annotation information (e.g., Gene Ontology, [2]) often divides genes into sets with similar functions. One of the main research questions for gene set inference is called gene set enrichment analysis (GSEA): we want to evaluate whether the gene set is enriched in terms of certain characteristics of interest (e.g., differential expression) relative to the other (random) gene sets.

A widely used approach starts from the list of significant differentially expressed genes derived from single gene analysis, and then evaluates over-representation of a gene set within a list of genes using Fisher's exact test, hypergeometric test, or other independent tests in a 2×2 contingency table. This approach has been modified by

many authors (see, e.g., [31] for a review), but the results of significance could be highly dependent on the selected cutoff value. An alternative approach is based on distribution comparisons. Typically a gene score, known as the local statistic for each gene that measures the difference of that gene's expression across different experimental conditions, is computed. Then a gene set score (global statistic) associated with local statistics within a gene set is compared to those of its complement. Several different variations of testing methods have been developed (see, e.g., [37], [43], [60] and [49]). Among the existing methods, the random set based methods proposed by Efron and Tibshirani [21] and Newton *et al.* [40] have standardized test statistics, which are then compared to random gene sets with significance assessed by permutation and random sampling. These random set based methods are state-of-the-art currently in the field.

In this thesis, we propose parametric approaches for GSEA, which could offer very competitive performance by combining information across all genes. In Chapter 2, we develop a general likelihood based approach for GSEA focused on two-sided enrichment analysis: testing overall differential expression (either down- or up-regulation). We extend the likelihood based approach for overall enrichment analysis to one-sided methods that test for enrichment of up (or down) regulation only in Chapter 3. Chapter 4 presents another extension of the proposed method to analyze microarray data with more than two experimental conditions. A short discussion on future work is provided in 5. In Chapter 2-4, the EM algorithms for implementing the proposed methods are also detailed, together with simulation studies and applications to real microarray data to illustrate the performance of the new methods over a current method.

Chapter 2

Likelihood based approach to gene set enrichment analysis with a finite mixture model

2.1 Introduction

One of the main research questions for gene set enrichment analysis (GSEA) is to evaluate whether the gene set is enriched in terms of certain characteristics of interest (e.g., differential expression) relative to the other (random) gene sets. The sets of genes are externally defined and derived from a variety of sources, such as Gene ontology [2], KEGG [41] and some public database of gene pathways, e.g., BioCarta (`http://www.biocarta.com`). A variety of approaches have been proposed for gene set enrichment analysis with respect to differential expression. Here we give a general description of commonly used gene set enrichment analysis methods. Typically the procedures begin by computing a measurement of differential expression for each gene (which we refer to as a "gene score") such as a two-sample t-statistic. Next, a gene set score is constructed as a function of gene scores within the set. The statistical significance for each gene set can be evaluated by comparing the gene set score to random set scores [40]. Alternatively, we can derive the null distribution of the gene set score statistic based on sample permutations to compute significance of enrichment [37, 49, 22].

Of the gene set enrichment analysis methods, GSEA proposed by Mootha *et al.* [37] and improved by Subramanian *et al.* [49] is one of the most popular approaches. They employed the Kolmogorov-Smirnov statistic between gene ranks of *a priori* set of genes and those of the rest of genes, and adopted permutation techniques to compute its significance. Barry *et al.* [3] proposed a more general framework, Significance Analysis of Function and Expression (SAFE), and offered more options for the individual gene scores. Tian *et al.* [52] and Goeman and Bühlmann [25] addressed the differences between subject permutation methods and gene sampling methods for generating the null distribution. Newton *et al.* [40] introduced a random-set statistical framework for significance assessment. Efron and Tibshirani [22] proposed a very novel "maxmean" statistic as gene set score and assessed its significance based on gene restandardization

and subject permutation. They show that the modified method has superior power and makes more accurate inferences possible. The random set based methods are state-of-the-art currently in the field. In this paper, we will approach the GSEA under a likelihood based testing framework and compare the performance to the approach of Efron and Tibshirani [21].

The rest of the chapter is organized as following. Statistical methods are introduced in Section 2.2, and we develop efficient numerical algorithms for model estimation in Section 2.3. Section 2.4 is devoted to simulation studies and Section 2.5 to an application to leukemia gene expression data. We end the chapter with a discussion in Section 2.6. All technical details are delegated to the Appendix (Section A.1).

2.2 Statistical methods

A Gene pathway often contains a collection of genes that work together to affect the system function. Gene annotation information such as Gene Ontology (GO) database [2] often divides genes into sets with similar functions. For the following discussion, we will summarize them as providing gene set information.

Consider a two-class microarray dataset, and denote the (modified) two-sample t-statistics as z_i for gene $i = 1, \cdots, m$. In the following, we discuss a finite normal mixture model for analyzing z_i. Consider the following finite mixture model with K+1 components

$$\sum_{k=0}^{K} \theta_k f_k(z), \quad f_0 = N(\mu_0, \sigma_0^2), \quad f_k = N(\mu_k, 1), \quad \theta_k > 0, \quad \sum_{k=0}^{K} \theta_k = 1. \tag{2.1}$$

Here the first component, $N(\mu_0, \sigma_0^2)$, empirically models null genes, which is different from theoretical null (standard normal distribution) and could take into account the potential dependence among genes [15]. We can interpret θ_0 as the proportion of null genes, and θ_k the proportion of genes with μ_k magnitude of differential expression. The

collection of all μ_k captures the heterogeneity of differential expressions across genes. Typically we use some moderated t-statistic combined with a normal transformation to achieve a better fit with a finite normal mixture model (see Section 2.5 for more details). In the finite mixture model, we fix the variance to one to make the model identifiable and relatively easy to estimate. Another motivation is that the t-statistic is known to approximately follow a normal distribution with variance one for relatively large sample size. We choose K based on BIC [47].

In enrichment analysis, we try to test whether a given gene set A is significantly different from any random gene set. Note that a random gene set can be treated as a random sampling from all genes. Thus comparing the given set A to a random set is equivalent to comparing A to all genes, which is again equivalent to comparing A to other genes (since A is a subset of all genes). Conceptually the modified two-sample t-statistics of genes in a given set can be modeled by a similar finite mixture model with different proportions for each component,

$$\sum_{k=0}^{K} \nu_k f_k(z), \quad \sum_{k=0}^{K} \nu_k = 1. \tag{2.2}$$

Under no enrichment, the gene set A and any random gene set have the same proportion of differentially expressed genes. Therefore gene set A and all the other genes (denoted as A^c) can be modeled respectively with

$$\nu_0 f_0(z) + \sum_{k=1}^{K} \nu_{jk} f_k(z), \quad \nu_0 + \sum_{k=1}^{K} \nu_{jk} = 1, \quad j = 1, 2. \tag{2.3}$$

Under enrichment, the gene set A and A^c have different proportions of differentially expressed genes, and hence can be modeled separately with

$$\sum_{k=0}^{K} \eta_{jk} f_k(z), \quad \sum_{k=0}^{K} \eta_{jk} = 1, \quad j = 1, 2. \tag{2.4}$$

Enrichment analysis corresponds to evaluating $\eta_{10} = \eta_{20}$, which can be tested by a likelihood ratio statistic, e_A, comparing models (2.3) and (2.4). The significance of e_A

can be approximately assessed using chi-square distribution with 1 degree of freedom. Enrichment analysis is a one-sided test assessing whether gene set A is enriched with more differentially expressed genes compared to a random set. Therefore we adjust p-value calculation as follows

$$\begin{cases} 0.5 + F(e_A; 1)/2 & \text{if } \hat{\eta}_{10} \geq \hat{\eta}_{20}, \\ 0.5 - F(e_A; 1)/2 & \text{otherwise,} \end{cases}$$

where $F(\cdot; df)$ is the χ^2_{df} distribution function.

In the following we discuss estimation of the empirical null distribution, and EM algorithms [11] for solving the proposed models (2.1), (2.2), and (2.3).

2.3 Model estimation

2.3.1 Empirical null distribution estimation and finite mixture model fitting

Efron [16] proposed two methods for estimating $(\theta_0, \mu_0, \sigma_0)$. One is "Central Matching", which approximates the marginal log density with a quadratic curve near zero assuming the central peak of the t-value histogram consists mainly of null genes. "Mle Fitting" is another method based on a truncated normal model by assuming the non null distribution has zero support in a pre-chosen small interval. Efron [15] shows in simulation studies that Central matching yields almost unbiased estimates if θ_0 exceeds 0.9, but it has large variation for estimating σ_0. Mle Fitting generally gives more stable estimates while it depends on the pre-chosen interval. Both methods have been implemented in the R package, *locfdr*. In our simulation studies, we have observed that the Mle Fitting method gives satisfactory results.

Given K and the estimated empirical null distribution parameters $(\hat{\theta}_0, \hat{\mu}_0, \hat{\sigma}_0^2)$, we can estimate (μ_k, θ_k) for model (2.1) iteratively based on the EM algorithm as follows

(see Appendix for technical details)

$$\begin{aligned}\theta_k^{(t+1)} &= (1-\hat{\theta}_0)\frac{\sum_{i=1}^m T_{k,i}^{(t)}}{\sum_{j=1}^K \sum_{i=1}^m T_{j,i}^{(t)}},\\ \mu_k^{(t+1)} &= \frac{\sum_{i=1}^m T_{k,i}^{(t)} z_i}{\sum_{i=1}^m T_{k,i}^{(t)}}, \quad k=1,\cdots,K,\end{aligned}$$

where

$$T_{k,i}^{(t)} = \frac{\theta_k^{(t)}\phi(z_i-\mu_k^{(t)})}{\hat{\theta}_0 f_0(z_i;\hat{\mu}_0,\hat{\sigma}_0^2)+\sum_{j=1}^K \theta_j^{(t)}\phi(z_i-\mu_j^{(t)})}, \quad k>0, \quad f_0(z_i;\hat{\mu}_0,\hat{\sigma}_0^2) = \frac{1}{\hat{\sigma}_0}\phi\Big(\frac{z_i-\hat{\mu}_0}{\hat{\sigma}_0}\Big).$$

Here $\phi(\cdot)$ is the standard normal density function.

Occasionally, the Mle Fitting method implemented in *locfdr* gives abnormal estimate of $\hat{\theta}_0$ which is larger than 1. We then estimate θ_0 in the EM algorithm together with other parameters as follows

$$\begin{aligned}\theta_k^{(t+1)} &= \frac{1}{m}\sum_{i=1}^m T_{k,i}^{(t)}, \quad k\geq 0,\\ \mu_k^{(t+1)} &= \frac{\sum_{i=1}^m T_{k,i}^{(t)} z_i}{\sum_{i=1}^m T_{k,i}^{(t)}}, \quad k>0,\end{aligned}$$

where

$$T_{0,i}^{(t)} = \frac{\theta_0^{(t)} f_0(z_i;\hat{\mu}_0,\hat{\sigma}_0^2)}{\theta_0^{(t)} f_0(z_i;\hat{\mu}_0,\hat{\sigma}_0^2)+\sum_{k=1}^K \theta_k^{(t)}\phi(z_i-\mu_k^{(t)})}.$$

2.3.2 Gene set model fitting

Given an estimated $(\hat{\mu}_0,\hat{\sigma}_0^2,\hat{\mu}_k)$ based on all genes, we can estimate individual model (2.2) for given set A as follows (see Appendix for technical details)

$$\nu_k^{(t+1)} = \frac{\sum_{i\in A} T_{k,i}^{(t)}}{m_A},$$

where m_A is the size of set A and for gene i in set A

$$T_{0,i}^{(t)} = \frac{\nu_0^{(t)} f_0(z_i;\hat{\mu}_0,\hat{\sigma}_0^2)}{\nu_0^{(t)} f_0(z_i;\hat{\mu}_0,\hat{\sigma}_0^2)+\sum_{k=1}^K \nu_k^{(t)}\phi(z_i-\hat{\mu}_k)},$$

$$T_{k,i}^{(t)} = \frac{\nu_k^{(t)}\phi(z_i - \hat{\mu}_k)}{\nu_0^{(t)} f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{j=1}^K \nu_j^{(t)}\phi(z_i - \hat{\mu}_j)}, \quad k = 1, \cdots, K.$$

2.3.3 Model fitting for a gene set and all the other genes under no enrichment

Under no enrichment, we can similarly estimate the mixture model (2.3) using the EM algorithm. Denote the complement of set A as A^c. Let

$$\begin{aligned}
T_{0,i}^{(t)} &= \frac{\nu_0^{(t)} f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2)}{\nu_0^{(t)} f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{l=1}^K \nu_{1l}^{(t)}\phi(z_i - \hat{\mu}_l)}, \quad i \in A, \\
T_{k,i}^{(t)} &= \frac{\nu_{1k}^{(t)}\phi(z_i - \hat{\mu}_g)}{\nu_0^{(t)} f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{l=1}^K \nu_{1l}^{(t)}\phi(z_i - \hat{\mu}_l)}, \quad i \in A, \quad k > 0, \\
T_{0,j}^{(t)} &= \frac{\nu_0^{(t)} f_0(z_j; \hat{\mu}_0, \hat{\sigma}_0^2)}{\nu_0^{(t)} f_0(z_j; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{l=1}^K \nu_{2l}^{(t)}\phi(z_j - \hat{\mu}_l)}, \quad j \in A^c, \\
T_{k,j}^{(t)} &= \frac{\nu_{2k}^{(t)}\phi(z_j - \hat{\mu}_k)}{\nu_0^{(t)} f_0(t_j; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{l=1}^K \nu_{2l}^{(t)}\phi(z_j - \hat{\mu}_l)}, \quad j \in A^c, \quad k > 0.
\end{aligned}$$

We then iteratively solve parameters as follows (see Appendix for technical details)

$$\begin{aligned}
\nu_0^{(t+1)} &= \frac{1}{m}(\sum_{i \in A} T_{0,i}^{(t)} + \sum_{j \in A^c} T_{0,j}^{(t)}), \\
\nu_{1k}^{(t+1)} &= (1 - \nu_0^{(t)}) \frac{\sum_{i \in A} T_{k,i}^{(t)}}{\sum_{l=1}^K \sum_{i \in A} T_{l,i}^{(t)}}, \\
\nu_{2k}^{(t+1)} &= (1 - \nu_0^{(t)}) \frac{\sum_{j \in A^c} T_{k,j}^{(t)}}{\sum_{l=1}^K \sum_{j \in A^c} T_{l,j}^{(t)}}.
\end{aligned}$$

Next we conduct a simulation study to compare the proposed likelihood based method to the state-of-the-art GSA approach (using the maxmean test statistic) studied by Efron and Tibshirani [21].

2.4 Simulation study

For $m = 10^4$ genes from two groups each with n samples, we simulate their expressions from normal distributions. Expression variance σ^2 is simulated individually for each gene from χ^2 distribution with 30 degrees of freedom. This mimics the commonly observed large variation of gene variances in microarray data. We simulate dependence by dividing genes into blocks each with 100 genes and within-block pairwise gene correlation being ρ. We randomly set $m\theta_0$ genes as null. The standardized differences of non-null genes, $(\mu_{1i} - \mu_{2i})/\sigma_i$, are simulated from a mixture of $(-1, -0.5, 0.5, 1)$ with proportions $2 : 3 : 3 : 2$. We compare the proposed likelihood based method (denoted as Lrt) to the approach of Efron and Tibshirani [21] (denoted as GSA). For size evaluation, we randomly sample m_e genes from all genes and compute the enrichment p-values based on Lrt and GSA in each simulation. Due to the random sampling, we know the set is not enriched. The random sampling is repeated 1000 times to obtain 1000 p-values, which is then used to assess Type I error rate. For power comparison, we consider gene set with m_e genes. We randomly select $m_e\theta_e$ genes from the null set, and $m_e(1 - \theta_e)$ from non-null set. We repeat the random sampling 1000 times to compute p-values and calculate power for any given type I error α. We have done simulations for $n = 25, 50$, $\rho = 0.2, 0.7$, $\theta_0 = 0.9, 0.95$, and $m_e = 50, 100$. Very similar patterns have been observed across different settings. Here we report the simulation results for $n = 25, \rho = 0.7, \theta_0 = 0.9, m_e = 100$. The complete results are available at the Appendix A.1.

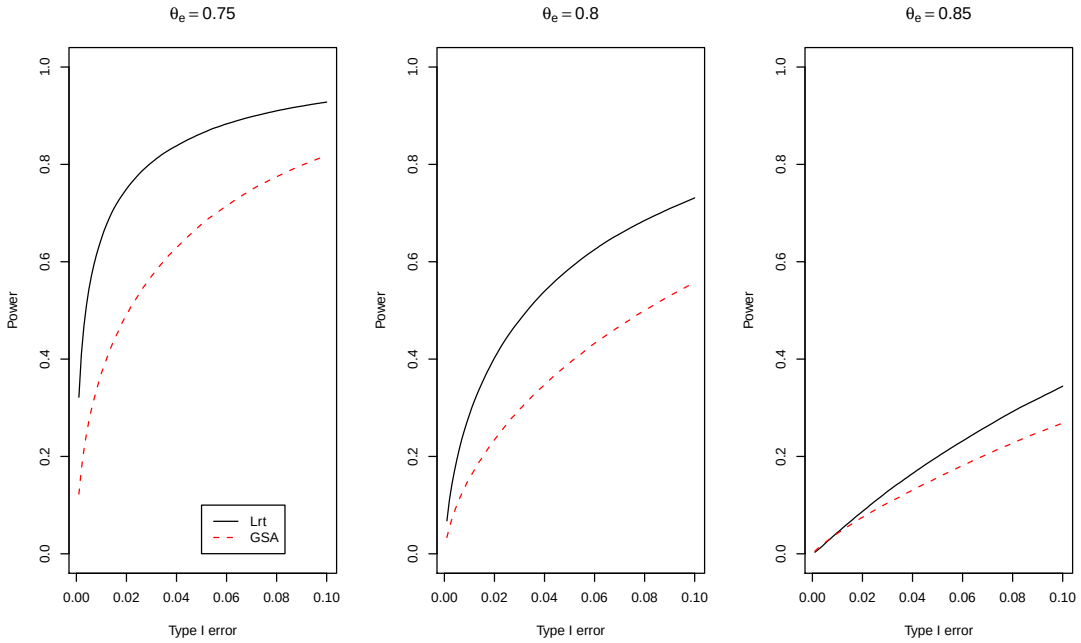
Table 2.1 summarizes the results for true Type-I error $\alpha = (0.005, 0.01, 0.05, 0.10)$ over 100 simulations. We can see that both methods have approximately the right size.

Figure 2.1 summarizes the power for $\theta_e = (0.75, 0.8, 0.85)$ over 100 simulations. Overall we can see that the proposed Lrt has very competitive performance compared to GSA.

Table 2.1: Estimated type I error of Lrt and GSA over 100 simulations. Listed within parenthesis are the standard errors.

	$\hat{\alpha}$			
α	0.005	0.01	0.05	0.1
Lrt	0.004 (0.0002)	0.009 (0.0003)	0.045 (0.0007)	0.090 (0.0009)
GSA	0.006 (0.0003)	0.012 (0.0004)	0.054 (0.0008)	0.104 (0.0011)

Figure 2.1: Power of Lrt and GSA averaged over 100 simulations for three different values of θ_e. The horizontal axis corresponds to type I error.



Next we analyze a leukemia gene expression microarray data to illustrate the relative performance of the proposed likelihood based method and GSA.

2.5 Application to leukemia gene expression data

The leukemia gene expression data reported at Kumar *et al.* [33] measured the expressions of 45,101 genes from 5 paired controls and Meis1-knockdown cases. We identified 522 gene pathways from the C2 functional collection in the *Molecular Signature Database* [49]. Pathway sizes range from 2 to 365 genes. We analyze in total 357 pathways that have more than 10 genes.

To improve the accuracy of the normal distribution approximation, we apply the empirical Bayes modeling approach [48], which computed a moderated t-statistic, t_i, for gene i by pooling information across all genes for an improved sample variance estimate (implemented in R package, *limma*). We then apply the normal distribution transformation to the moderated t-statistic

$$z_i = \Phi^{-1}(T_d(t_i))$$

where $\Phi(\cdot)$ is the standard normal distribution function and $T_d(\cdot)$ is the t-distribution function with d degrees of freedom. Here, the degree of freedom d is estimated from all genes using the empirical Bayes modeling approach.

When applied to the leukemia microarray data, controlling FDR at 0.05/0.1, the proposed likelihood based method (Lrt) detected 29/51 significant gene sets, while no gene pathway is identified as significant with GSA. Figure 2.2 shows the number of significant pathways versus the estimated FDR for Lrt and GSA.

Table 2.2 lists the top 29 significant pathways identified by the proposed method. Many of them are closely related to cancer development. For example, several identified pathways are related to the cell cycle, which is known to play an important role in cancer

development: cell cycle machinery controls cell proliferation, and cancer is a disease of inappropriate cell proliferation [8]. The atrbrcaPathway is also closely related to cell cycle and cancer. Specifically the ATR gene serves as a checkpoint kinase that halts cell cycle progression and induces DNA repair when DNA is damaged. Loss of ATR results in a loss of checkpoint control in response to DNA damage, leading to cell death (see `http://www.biocarta.com/pathfiles/h_ATRBRCAPATHWAY.asp`). The DNA damage signaling pathway is linked to DNA repair, cell-cycle control, growth arrest, and plays an important role in cancer development.

Figure 2.2: The number of significant pathways versus FDR

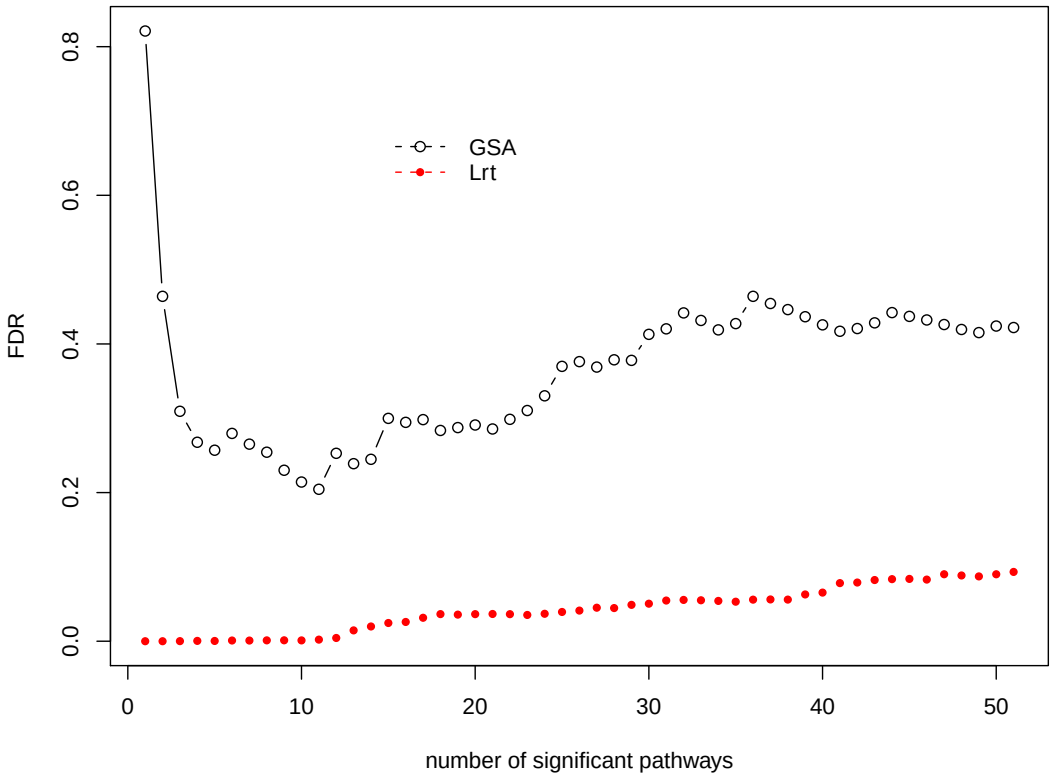


Table 2.2: Top 29 most significant pathways identified with the proposed likelihood based method.

Pathway	number of genes	p-value
Cell_Cycle	73	2E-13
CR_CELL_CYCLE	74	5E-11
atrbrcaPathway	18	7E-07
CR_REPAIR	35	4E-06
GLUT_DOWN	230	5E-06
cell_cycle_checkpoint	22	1E-05
DNA_DAMAGE_SIGNALLING	85	1E-05
HTERT_UP	94	2E-05
CR_DNA_MET_AND_MOD	20	3E-05
LEU_DOWN	130	3E-05
cell_cycle_regulator	20	6E-05
rbPathway	11	0.0001
cell_cycle_arrest	27	0.0005
hdacPathway	28	0.0008
RAP_DOWN	169	0.0010
SA_REG_CASCADE_OF_CYCLIN_EXPR	12	0.0011
il7Pathway	16	0.0015
mRNA_processing	40	0.0018
shh_lisa	15	0.0019
GLUCOSE_DOWN	122	0.0020
MAP00020_Citrate_cycle_TCA_cycle	16	0.0022
cellcyclePathway	22	0.0022
mRNA_splicing	45	0.0023
SIG_IL4RECEPTOR_IN_B_LYPHOCYTES	26	0.0025
caspasePathway	21	0.0028
crebPathway	25	0.0030
eif4Pathway	24	0.0034
MAP00240_Pyrimidine_metabolism	38	0.0035
nfatPathway	49	0.0040

2.6 Discussion

The GSEA approach first proposed and studied by Mootha *et al.* [37] and Subramanian *et al.* [49] provides a very novel way to interpret the large-scale gene expression data. Compared to individual gene oriented analysis, gene set based inference can often produce results that are meaningful and easy to interpret, and provide additional insights into the underlying biological processes. Many simple and ad hoc statistical methods based on categorization are becoming routinely used in practice (e.g., the widely used hypergeometric testing approach) for gene set significance assessment. Nonparametric methods based on permutation and random sampling have been proposed and proven to be more powerful but might be quite computationally intensive. In this chapter, we approach the GSEA from a likelihood framework and transform it into a model comparison problem, which can be addressed using the powerful likelihood ratio test approach. Through applications and simulation studies we have demonstrated the competitive performance of the proposed method. An interesting extension is to develop a similar method for one-sided enrichment analysis testing for enrichment of up (or down) regulation only. We will describe it in the next chapter.

Chapter 3

Extension to one-sided gene set enrichment analysis

3.1 Introduction

In Chapter 2, we propose the likelihood based method for analyzing overall enrichment. We formulate enrichment analysis into a model comparison problem and adopt the asymptotic chi-square approximation for efficient computation of significance instead of relying on the computing intensive random sampling or permutation approaches. In this chapter, we extend the likelihood based approach to enrichment analysis of one-sided differential expressions, i.e., assessing whether a gene set has more significantly up- or down-regulated genes. This approach is motivated from biological research where investigators want to distinguish between up- and down-regulations. We will consider a two-component mixture model based on the finite normal mixture model assuming each distribution is separately generated from either genes of interest or the rest of genes. The interesting genes are defined as "Non-null" while the non-interesting genes as "Null".

The rest of the chapter is organized as following. Section 3.2 discusses in detail the proposed method. We analyze a leukemia gene expression data to illustrate the performance of the proposed method in Section 3.3. Section 3.4 presents simulation results based on the leukemia gene expression data. Concluding remarks are provided in Section 3.5.

3.2 Statistical Methods

In the following discussion, we describe statistical methods for detecting gene sets enriched with up-regulated genes. It is understood that the methods could be readily extended to down-regulation enrichment analysis by taking the negative of the gene scores.

3.2.1 Finite mixture modeling of up-regulation

We begin with a two-sample t-statistic z_j for gene $j = 1, \cdots, m$. Typically z_j is normally transformed. We can divide all genes into two groups, "Non-null" and "Null". Non-null consists of up-regulated genes of our interest and Null includes down-regulated and not differentially expressed genes. Overall, we analyze z_j with a two-component mixture model [19]

$$f(z) = p_0 f_0(z) + (1 - p_0) f_1(z), \tag{3.1}$$

where $f_1(z)$ is density of Non-null genes and $f_0(z)$ is density of Null genes. p_0 indicates the proportion of Null genes. In the following, we discuss a finite mixture model for estimating f_0 and f_1.

For those not differentially expressed genes, we empirically model them with a normal distribution $N(\mu_0, \sigma_0^2)$ to incorporate the commonly observed gene interactions [15, 16]. For differentially expressed genes (including both up- and down-regulated), we model them with a finite normal mixture with each component having variance fixed at 1. Therefore, for all genes we have (Chapter 2)

$$\theta_0 N(\mu_0, \sigma_0^2) + \sum_{k=1}^{K} \theta_k N(\mu_k, 1), \quad \sum_{k=0}^{K} \theta_k = 1, \tag{3.2}$$

where K is typically selected based on BIC.

Under the previous finite mixture model, intuitively Non-null genes of interest (up-regulated genes) can be modeled by those components with positive mean, $\mu_k > 0$, and Null genes (including down-regulated and not differentially expressed) are modeled with those components with $\mu_k < 0$ and $N(\mu_0, \sigma_0^2)$. In our limited experience from data analysis, we have found that the empirical distribution component, $N(\mu_0, \sigma_0^2)$, can not completely take into account the gene dependence, and sometimes we observe μ_k with very small magnitude. Here we adopt the idea of non-interesting genes proposed by Ruppert *et al.* [45] by putting a threshold c_0 (e.g., $c_0 = 0.5$) on μ_k, which yields the

following finite mixture model for Non-null genes of our interest (up-regulated genes)

$$f_1(z) = \frac{1}{1-\theta_c}\sum_{k=1}^{K}\theta_k N(\mu_k, 1)I(\mu_k > c_0),$$

where

$$\theta_c = \theta_0 + \sum_{k}^{K}\theta_k I(\mu_k \leq c_0).$$

And Null genes are modeled with

$$f_0(z) = \frac{1}{\theta_c}\theta_0 N(\mu_0, \sigma_0^2) + \frac{1}{\theta_c}\sum_{k=1}^{K}\theta_k N(\mu_k, 1)I(\mu_k \leq c_0).$$

We first estimate the empirical null component $N(\mu_0, \sigma_0^2)$ following Efron [18]. All the other components of model (3.2) can then be readily estimated using EM algorithm (see e.g. Chapter 2)

3.2.2 Enrichment analysis

In enrichment analysis, we try to test whether a given gene set, denoted as A, is significantly enriched with up-regulated genes compared to any random gene set. Intuitively comparing the given gene set A to a random set is equivalent to comparing A to all the other genes, denoted A^c. Conceptually, gene set A and A^c can be separately modeled by the two-component mixture model in (3.1) with different Null proportions. Under no enrichment of up-regulation, A and A^c have the same proportion of Null genes, p_0. Under enrichment, A and A^c can be modeled respectively with

$$p_{k0}f_0(z) + (1-p_{k0})f_1(z), \quad k = 1, 2. \tag{3.3}$$

The maximum likelihood estimate of p_{10} can be obtained based on the following recursion rule (see Section A.2 for details)

$$p_{10}^{(t+1)} = \frac{1}{m_A}\sum_{j\in A}\frac{p_{10}^{(t)}f_0(z_j)}{p_{10}^{(t)}f_0(z_j) + (1-p_{10}^{(t)})f_1(z_j)}$$

where m_A is the size of set A. We can similarly estimate p_{20}.

Enrichment analysis corresponds to evaluating $p_{10} = p_{20}$ versus $p_{10} < p_{20}$, which can be tested using a likelihood ratio statistic (denoted as e_A) comparing models (3.1) and (3.3). We note that model (3.1) is the null model for all gene sets and needs to be estimated just once computationally. The significance p-value of enrichment can be approximately computed based on χ^2 distribution as $0.5 + F(e_A; 1)/2$ when $\hat{p}_{10} \geq \hat{p}_{20}$ and $0.5 - F(e_A; 1)/2$ otherwise. Here $F(\cdot; 1)$ is the χ^2 distribution function with 1 degree of freedom.

We first analyze a leukemia gene expression data to empirically compare the performance of the proposed method to a representative method, GSA, proposed by Efron and Tibshirani [22].

3.3 Application to leukemia gene expression data

The leukemia microarray data studied at Chen *et al.* [6] compares the expression differences between *Mll-AF9* knockin and wild type mice in four cell types: 10 hematopoietic stem cells, 15 common lymphoid progenitors, 9 common myeloid progenitors, and 9 granulocyte-monocyte progenitors. Among the measured 45,101 probes from Affymetrix murine 430 2.0 genechip, we select 35,848 probes that have unigene annotations for analysis. We study the enrichment of differentially expressed genes for C2 functional pathways from the Molecular Signature Database (`http://www.broadinstitute.org/gsea/msigdb/index.jsp`). There are in total 1892 gene pathways, and we studied 1736 sets that have at least 15 genes.

We fit an additive effects model to analyze the gene expression differences between *Mll-AF9* and wild type samples. Specifically, for the expression values of gene j,

$\{x_{1j}, \cdots, x_{n_j}\}$, we assume

$$\log(x_{ij}) = \alpha_j + \beta_j y_i + \sum_{k=1}^{3} \theta_{kj} z_{ki} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma_j^2) \quad i = 1, \cdots, n,$$

where y_i is the indicator for the *Mll-AF9* versus the wild type and (z_{1i}, z_{2i}, z_{3i}) are three indicators for modeling cell type differences. We adopt the empirical Bayes approach of Smyth [48] to obtain a moderated estimate of σ_j^2 that has been shown in simulation studies to be more robust than the sample variance and very effective for analyzing large-scale gene expression data. For gene j, the moderated t-statistic for β_j is used to summarize its differential expression, and we further carry out a normal transformation for follow-up analysis. For the GSA method, we use 10,000 permutations to compute enrichment p-values for each gene set.

Figure 3.1 shows the estimated FDR for analyzing enrichment of up- and down-regulation. In general, we can see that the proposed method (Lrt) identified more significant pathways than GSA at the same FDR level. Controlling FDR at 0.05, Lrt identified 32/255 pathways significantly enriched with up/down-regulated genes, while 5/7 pathways are detected by GSA, respectively. Table 3.1 and 3.2 list the selected significant pathways (only top 30 pathways enriched with down-regulated genes are listed for Lrt. The complete list of 255 pathways are provided in Appendix A.2). Nearly all pathways identified by GSA are also selected by Lrt.

Most detected pathways are directly associated with leukemia induction. For example, pathways including an "AML" or "ALL" term in the name consist of genes related to acute myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL), respectively. In addition, many selected pathways are associated with the "MLL" gene which is involved in chromosomal translocations associated with AML leukemia [51]. Specifically, those pathways such as VERHAAK_AML_NPM1_MUT_VS_WT_UP, VERHAAK_AML_NPM1_MUT_VS_WT_DN, ALCALAY_AML_NPMC_UP and ALCALAY_

AML_NPMC_DN are related to mutations of gene NPM1. Interestingly, the gene NPM1 has been used for a diagnostic factor for acute myeloid leukemia [58, 1]. The two pathways, CHIARETTI_T_ALL_DIFF and CHIARETTI_T_ALL, consist of genes expressed in T-cell acute lymphocytic leukemia [7]. ROSS_MLL_FUSION and ROSS_PML_RAR pathways are collection of genes which distinguish pediatric AML subtypes [44]. Activation of HOXA9 genes has been reported to be one of the required targets for leukemia development by MLL fusion genes [32]. Indeed, a fusion between this HOXA9 and the NUP98 gene has been associated with myeloid leukemia induction [50, 38]. The proposed Lrt method detected 7 pathways from up- and 10 from down- regulation related to HOXA9 or NUP98 genes. In addition, some pathways detected by Lrt are related to Hematopoietic stem cells, which play a critical role in cellular blood components. It is known that abnormalities in this cellular developmental program lead to blood cell diseases including leukemia [46].

In the next section, we conduct a simulation study based on the leukemia gene expression data to compare the performance of the proposed method and GSA.

Figure 3.1: Estimated FDR for top 30 ranked pathways enriched with up-regulated (left panel) and down-regulated (right panel) genes identified by GSA and Lrt.

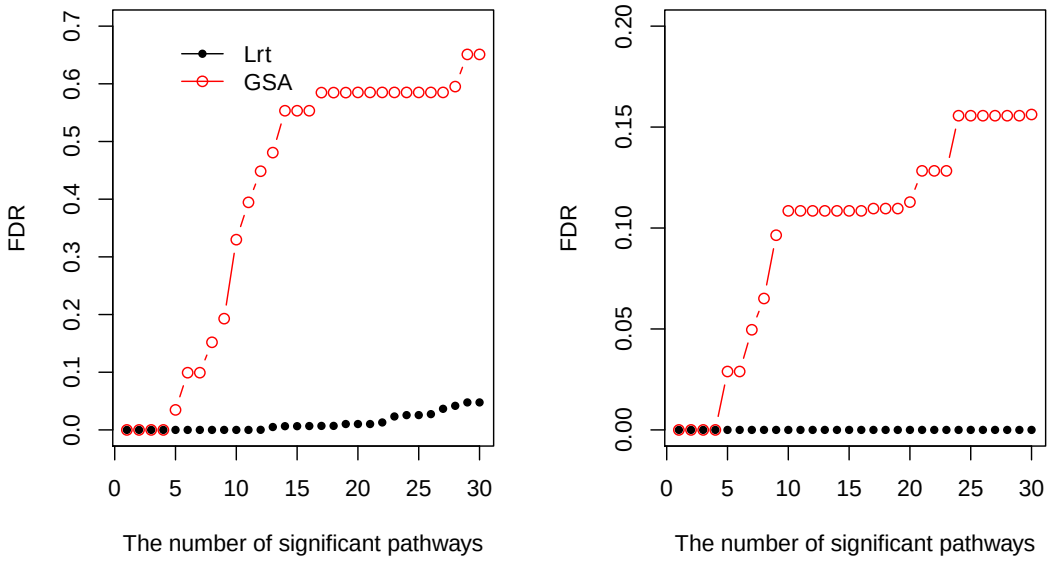


Table 3.1: The identified pathways enriched with up-regulation by Lrt and GSA (FDR≤ 0.05)

Lrt

Pathway	# genes	p-value
VERHAAK_AML_NPM1_MUT_VS_WT_UP	285	6E-15
HOX_GENES	83	4E-14
ALCALAY_AML_BY_NPM1_LOCALIZATION_UP	254	6E-12
KUMAR_HOXA_DIFF	789	7E-12
WANG_MLL_CBP_VS_GMP_UP	116	8E-12
ROSS_MLL_FUSION	129	1E-10
TAKEDA_NUP8_HOXA9_3D_UP	296	2E-10
TAKEDA_NUP8_HOXA9_16D_UP	244	2E-10
TAKEDA_NUP8_HOXA9_10D_UP	278	3E-10
HEMATOPOESIS_RELATED_TRANSCRIPTION_FACTORS	170	2E-09
TAKEDA_NUP8_HOXA9_8D_UP	242	3E-09
CHIARETTI_T_ALL_DIFF	488	1E-06
CHIARETTI_T_ALL	449	4E-05
EPHA4PATHWAY	25	5E-05
HSA04514_CELL_ADHESION_MOLECULES	249	6E-05
MONOCYTEPATHWAY	27	6E-05
RIBAVIRIN_RSV_DN	90	7E-05
LAIRPATHWAY	28	7E-05
GH_AUTOCRINE_DN	235	0.0001
PASSERINI_SIGNAL	645	0.0001
CELL_ADHESION_MOLECULE_ACTIVITY	207	0.0001
TAKEDA_NUP8_HOXA9_6H_UP	134	0.0002
PARK_MSCS_DIFF	83	0.0003
GOLUB_ALL_VS_AML_DN	30	0.0004
JISON_SICKLECELL_DIFF	610	0.0004
JISON_SICKLE_CELL	58	0.0004
LEE_DENA_DN	112	0.0006
HSA04670_LEUKOCYTE_TRANSENDOTHELIAL_MIGRATION	228	0.0007
CMV-UV_HCMV_6HRS_DN	190	0.0008
IRITANI_ADPROX_DN	108	0.0008
HIPPOCAMPUS_DEVELOPMENT_NEONATAL	53	0.0009
ZHAN_MMPC_SIMAL	85	0.0009

GSA

Pathway	# genes	p-value	rank by Lrt
TAKEDA_NUP8_HOXA9_16D_UP	244	0	8
ROSS_MLL_FUSION	129	0	6
HEMATOPOESIS_RELATED_TRANSCRIPTION_FACTORS	170	0	10
WANG_MLL_CBP_VS_GMP_UP	116	0	5
HOX_GENES	83	0.0001	2

Table 3.2: Top 30 pathways enriched with down-regulation identified by Lrt and GSA (FDR$\leq$ 0.05). The full list is available in Appendix A.2.

Lrt

Pathway	# genes	p-value
LEE_TCELLS3_UP	183	0
LEE_TCELLS2_UP	2008	0
SERUM_FIBROBLAST_CELLCYCLE	243	0
LE_MYELIN_UP	198	7E-16
LI_FETAL_VS_WT_KIDNEY_DN	349	5E-15
STEMCELL_HEMATOPOIETIC_UP	2796	2E-14
HSC_HSCANDPROGENITORS_FETAL	1112	3E-13
HSC_HSCANDPROGENITORS_ADULT	1134	6E-13
POD1_KO_UP	786	1E-12
STEMCELL_NEURAL_UP	3845	1E-12
TAKEDA_NUP8_HOXA9_8D_DN	326	1E-12
MATSUDA_VALPHAINKT_DIFF	1058	3E-12
HOFFMANN_BIVSBII_BI_TABLE2	437	4E-12
KUMAR_HOXA_DIFF	789	5E-12
STEMCELL_EMBRYONIC_UP	2773	8E-12
GAY_YY1_DN	632	8E-12
CHIARETTI_T_ALL_DIFF	488	1E-11
BRCA_ER_NEG	1671	5E-11
HADDAD_HSC_CD7_UP	97	6E-11
UVB_NHEK3_ALL	848	6E-11
IDX_TSA_UP_CLUSTER3	208	1E-10
TAKEDA_NUP8_HOXA9_10D_DN	207	1E-10
DOX_RESIST_GASTRIC_UP	79	2E-10
HADDAD_HPCLYMPHO_ENRICHED	501	9E-10
MIDDLEAGE_DN	33	2E-09
HADDAD_HSC_CD10_UP	470	2E-09
CROONQUIST_IL6_STARVE_UP	75	3E-09
P21_P53_ANY_DN	90	2E-09
PRMT5_KD_UP	381	5E-09
CANCER_UNDIFFERENTIATED_META_UP	120	8E-09

GSA

Pathway	# genes	p-value	rank by Lrt
CMV_HCMV_TIMECOURSE_8HRS_DN	35	0	60
YE_INTRAMETASTATIC_HCC_UP	41	0.0002	118
IDX_TSA_DN_CLUSTER1	83	0	33
3AB_GAMMA_DN	24	0	87
MATSUDA_VALPHAINKT_DIFF	1058	0.0001	12
HDACI_COLON_SUL2HRS_DN	33	0.0001	63
CORDERO_KRAS_KD_VS_CONTROL_DN	100	0	67

3.4 Simulation study

We consider $m = 40,000$ genes from two groups each with $n = 20$ samples and simulate their expression levels from normal distribution $N(\mu_{kj}, \sigma_j^2)$ for group $k = 1, 2$ and gene j. The individual gene variance σ_j^2 is simulated from chi-square distribution with 30 degrees of freedom. The dependence is generated by dividing genes into non-overlapping blocks each with 200 genes and within-block pairwise gene correlation being $\rho = 0.5$. We set null gene proportion as $\theta_0 = 0.97$. The standardized differences of non-null genes, $(\mu_{1j} - \mu_{2j})/\sigma_j$, are simulated from a mixture of $\sqrt{\frac{2}{n}}(-2.2, -5.2, 4.46)$ with proportions $(0.026, 0.0012, 0.0028)$. We fix the threshold $c_0 = \pm 0.5$ to define up/down-regulated genes of interest. We randomly sample m_e genes from all genes to empirically compute type I errors. To simulate enriched gene sets, $m_e\theta_e$ genes are randomly sampled from the null set and $m_e(1 - \theta_e)$ genes from non-null set. We repeat the random sampling 1000 times to compute size and power for any given Type I error α. For GSA we used 10,000 permutations to compute enrichment p-values. Here we report the simulation results for $m_e = (200, 400)$, $\theta_e = (0.9, 0.91, 0.92)$ for down-regulation and $\theta_e = (0.96, 0.97, 0.98)$ for up-regulation. We have done simulations under various other settings and observed very similar patterns (see Appendix A.2 information for complete results) .

Table 3.3 presents the nominal Type I error for $\alpha = 0.005, 0.01, 0.05, 0.10$. Both approaches have approximately the right sizes. GSA slightly over-estimates the Type I error and Lrt is slightly conservative. It is mainly because the null/non-null densities f_0/f_1 in model (3.1) are estimated using all genes, which reduces the essential degrees of freedom difference between enrichment models for each gene set. As a partial remedy, we model the null distribution of enrichment likelihood ratio statistics with a scaled chi-square, $e_A/\lambda_0 \sim \chi_1^2$. It is not hard to verify that the maximum likelihood estimate of λ_0 is the sample average. Table 3.3 shows the results for this scaled chi-square modeling approach (denoted as aLrt), which is less conservative than Lrt.

Table 3.3: Average type I error over 100 simulations (listed within parenthesis are the standard errors).

α	0.005	0.01	0.05	0.1
		$m_e = 200$ (down-regulation)		
GSA	0.0064 (0.0002)	0.0115 (0.0004)	0.0516 (0.0007)	0.1010 (0.0010)
Lrt	0.0042 (0.0002)	0.0088 (0.0003)	0.0447 (0.0007)	0.0899 (0.0010)
aLrt	0.0049 (0.0002)	0.0100 (0.0003)	0.0480 (0.0005)	0.0944 (0.0007)
		$m_e = 200$ (up-regulation)		
GSA	0.0060 (0.0002)	0.0110 (0.0003)	0.0509 (0.0007)	0.1005 (0.0009)
Lrt	0.0037 (0.0002)	0.0080 (0.0003)	0.0398 (0.0007)	0.0840 (0.0012)
aLrt	0.0041 (0.0002)	0.0086 (0.0003)	0.0419 (0.0006)	0.0866 (0.0012)
		$m_e = 400$ (down-regulation)		
GSA	0.0062 (0.0002)	0.0112 (0.0003)	0.0511 (0.0007)	0.1008 (0.0010)
Lrt	0.0043 (0.0002)	0.0089 (0.0003)	0.0449 (0.0007)	0.0903 (0.0009)
aLrt	0.0051 (0.0002)	0.0104 (0.0003)	0.0484 (0.0006)	0.0954 (0.0006)
		$m_e = 400$ (up-regulation)		
GSA	0.0064 (0.0003)	0.0114 (0.0003)	0.0510 (0.0007)	0.1001 (0.0010)
Lrt	0.0043 (0.0002)	0.0082 (0.0003)	0.0428 (0.0007)	0.0852 (0.0011)
aLrt	0.0053 (0.0002)	0.0097 (0.0003)	0.0473 (0.0007)	0.0910 (0.0008)

Figure 3.2 through 3.5 summarize the estimated power and Type I error averaged over 100 simulations. Overall, the proposed aLrt performs slightly better than Lrt, and both can detect significantly more truly enriched gene sets than GSA.

Figure 3.2: Power curve for Lrt and GSA: $m_e = 200$ and up-regulation.

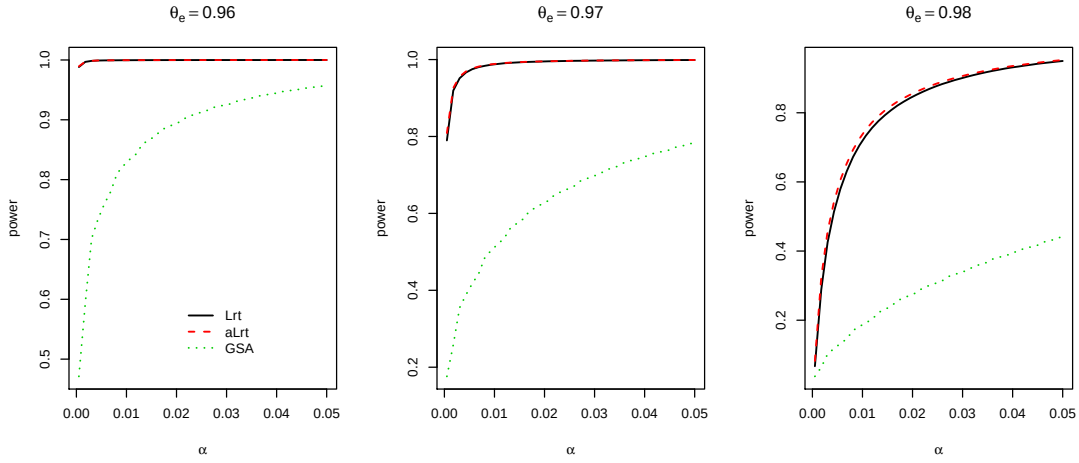


Figure 3.3: Power curve for Lrt and GSA: $m_e = 400$ and up-regulation.

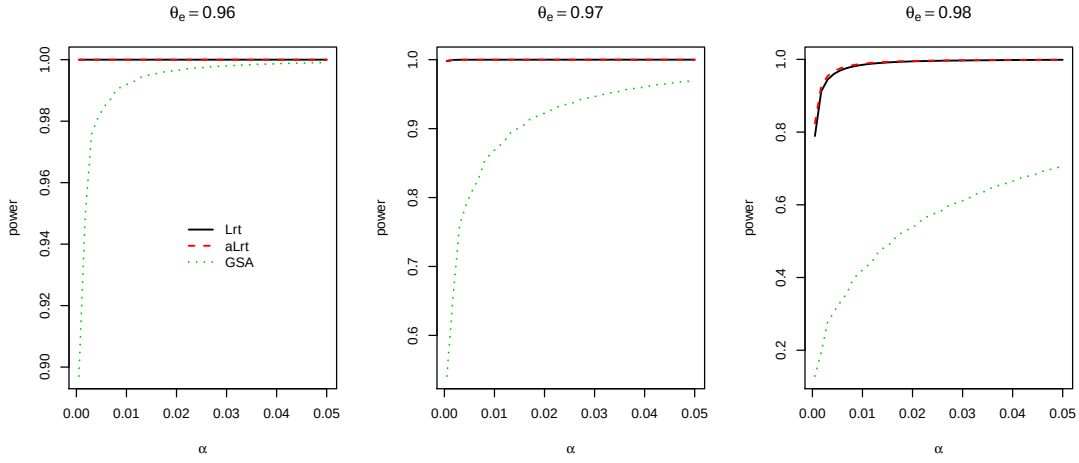


Figure 3.4: Power curve for Lrt and GSA: $m_e = 200$ and down-regulation.

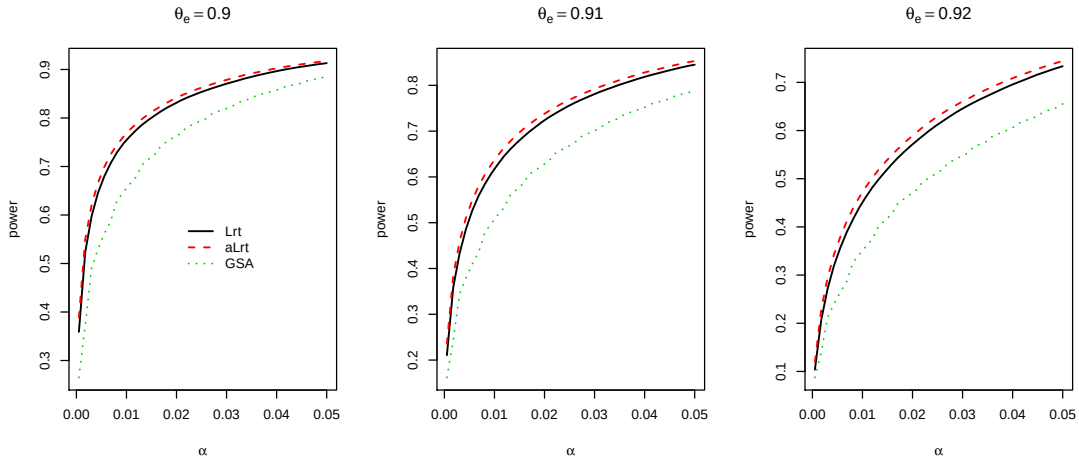
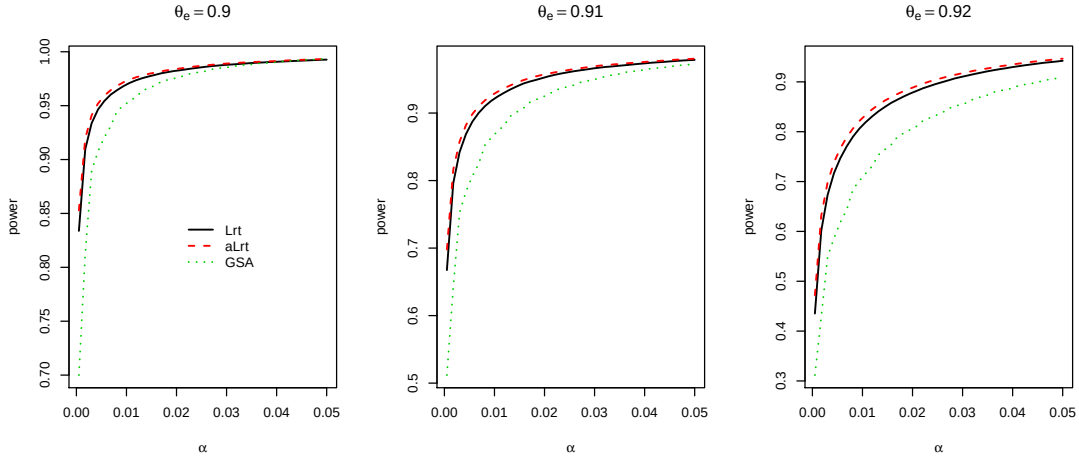


Figure 3.5: Power curve for Lrt and GSA: $m_e = 400$ and down-regulation.



3.5 Discussion

In this chapter, we studied statistical methods for one-sided gene set enrichment analysis, which is often motivated from biological research where only a specific type of differential expression is of interest. The proposed method extended the likelihood based approach for the overall enrichment analysis proposed at Chapter 2. Specifically we have adopted a finite mixture model to approximate the distribution of differentially expressed genes of interest (either up or down-regulated genes). In our limited experience, we found that the fitted finite mixture model selected based on BIC could potentially miss those genes of interest. For example, in detecting gene sets enriched with up-regulated genes, we could select a model without $\mu_k > 0$. Currently, we increased the number of finite mixture components to always include at least one component of interest in the selected mixture model. It will be interesting to explore an alternative approach that uses a continuous distribution to approximate the underlying effective size distribution of differentially expressed genes as proposed by Ruppert *et al.* [45].

Chapter 4

Enrichment analysis for multi-class microarray data

4.1 Introduction

Most enrichment analysis methods mainly focus on two-class gene expression data although many microarray studies are often conducted in more than two experimental conditions. Besides, existing methods for multi-class data have largely relied on non-parametric approaches that often require computing intensive sample permutation or gene random sampling to assess the enrichment significance. In this chapter, we attend to multi-class differential expression and derive a novel tool to extend the main idea of likelihood based approach for GSEA in Chapter 2. We first develop an empirical null distribution based on a finite mixture model for multi-class differential expressions which takes into account potential gene dependence. We then propose an efficient and powerful mixture model based likelihood testing approach to multi-class enrichment analysis. Simulation studies and application to public microarray data are used to illustrate the competitive performance of the proposed method.

The rest of the chapter is organized as follows. In Section 4.2 we introduce the proposed method for the multi-class gene set enrichment analysis, and develop a computationally efficient algorithm for the proposed approach. Section 4.3 is devoted to a simulation study to investigate the performance of the proposed method. We analyze a breast cancer microarray data for illustration in Section 4.4. Concluding remarks are provided in Section 4.5.

4.2 Statistical methods

Consider a microarray data measuring differential expressions of m genes under G experiment conditions. For testing equal expressions across G levels, we compute the F-statistic, which is then transformed to a chi-square statistic z_i for gene $i = 1, \cdots, m$ following χ^2 distribution with $\nu = G - 1$ degrees of freedom.

4.2.1 Finite mixture model for multi-class differential expression and enrichment analysis

We analyze z_i with a finite chi-square mixture model

$$\frac{\theta_0}{\sigma_0^2} f_\nu(\frac{z}{\sigma_0^2}; \lambda_0) + \sum_{k=1}^{K} \theta_k f_\nu(z; \lambda_k), \quad \theta_k > 0, \quad \sum_{k=0}^{K} \theta_k = 1, \tag{4.1}$$

where $f_\nu(\cdot; \lambda)$ denotes the χ^2 distribution function with non-centrality parameter λ and degrees of freedom ν. The coefficients θ_k's are the mixing proportions and K is the number of components, which can be chosen based on BIC.

In the chi-square mixture model (4.1), we empirically model null genes with a non-central χ^2 distribution with scale σ_0^2 and non-centrality λ_0 instead of the theoretical central χ^2 distribution with ν degrees of freedom. The empirical modeling approach could partially account for gene dependence and help to improve the accuracy in estimating the null distribution, which has been extensively studied for two-class microarray data [15, 16]. To model the expression heterogeneity of differentially expressed genes, we use a mixture of K chi-square distributions each with a non-centrality parameter λ_k.

In our experiences of simulation and data analysis, we have found that the empirical null distribution can not completely take into account the gene dependence. Sometimes we observed some non-null component with very small λ_k. We adopt the approach of Ruppert *et al.*[45] by putting a threshold t_0 (e.g., $t_0 = 1$) on λ_k to identify those truly differentially expressed genes of our interest.

Based on model (4.1), we propose to model null genes by

$$g_0(z) = \frac{\theta_0}{\theta_e \sigma_0^2} f_\nu(\frac{z}{\sigma_0^2}; \lambda_0) + \frac{1}{\theta_e} \sum_{k=1}^{K} \theta_k f_\nu(z; \lambda_k) I(\lambda_k \le t_0), \tag{4.2}$$

and non-null genes by

$$g_1(z) = \frac{1}{1 - \theta_e} \sum_{k}^{K} \theta_k f_\nu(z; \lambda_k) I(\lambda_k > t_0) \tag{4.3}$$

where $\theta_e = \theta_0 + \sum_{k=1}^{K} \theta_k I(\lambda_k \leq t_0)$. Overall we have a two-component mixture model for analyzing multi-class differential expressions

$$\gamma_0 g_0(z) + (1 - \gamma_0) g_1(z) \tag{4.4}$$

where γ_0 is the proportion of empirically defined null genes.

In enrichment analysis, we try to test whether a given gene set, denoted A_1, is significantly enriched compared to any random gene set, which intuitively can be approached by comparing gene set A_1 to all the other genes, denoted A_2. Under no enrichment, the gene set A_1 and A_2 have the same proportion of null genes, γ_0. Under enrichment, they can be modeled separately with

$$\gamma_{j0} g_0(z) + (1 - \gamma_{j0}) g_1(z) \quad j = 1, 2 \tag{4.5}$$

Enrichment analysis of gene set A_1 is evaluating $H_0 : \gamma_{10} = \gamma_{20}$, which can be approached by comparing model (4.4) and (4.5) and tested using a likelihood ratio statistic, denoted e_A. The significance of e_A can be approximately assessed based on the asymptotic χ^2 distribution. Enrichment analysis is a one-sided test: we are testing whether gene set A_1 is enriched with more differentially expressed genes compared to A_2. Therefore we compute enrichment significance p-value as $0.5 + F_1(e_A)/2$ if $\hat{\gamma}_{10} \geq \hat{\gamma}_{20}$ and $0.5 - F_1(e_A)/2$ otherwise. Here $F_1(\cdot)$ is the χ^2 distribution function with 1 degrees of freedom.

We discuss computational methods for estimating the proposed models in the next section.

4.2.2 Model estimation

For empirical null estimation, we assume all $\{z_i \leq c_0\}$ come from null genes, where c_0 is pre-chosen, e.g., the 75% percentile of all z_i. We have a constrained χ^2 density for

modeling genes with $z_i \le c_0$

$$h(z; c_0) = \frac{f_\nu(\frac{z}{\sigma_0^2}; \lambda_0)/\sigma_0^2}{F_\nu(\frac{c_0}{\sigma_0^2}; \lambda_0)} I(z \le c_0)$$

where $F_\nu(\cdot; \lambda_0)$ is the χ^2 distribution function with degrees of freedom ν and non-centrality λ_0. Define the null probability

$$p_0 = \Pr(z \le c_0) = \theta_0 F_\nu(\frac{c_0}{\sigma_0^2}; \lambda_0).$$

We then obtain the following log likelihood

$$m_0 \log(p_0) + (m - m_0) \log(1 - p_0) + \sum_{z_i \le c_0} \log h(z_i; c_0), \quad m_0 = \sum_{i=1}^{m} I(z_i \le c_0),$$

which can be numerically maximized to estimate $(\theta_0, \sigma_0^2, \lambda_0)$.

Given $(\hat{\theta}_0, \hat{\sigma}_0^2, \hat{\lambda}_0)$ and fixed K, we can estimate (θ_k, λ_k) based on EM algorithm as follows. First we compute the conditional probabilities in the E-step

$$\tau_{i0}^{(t)} = \frac{\frac{\hat{\theta}_0}{\hat{\sigma}_0^2} f_\nu(\frac{z_i}{\hat{\sigma}_0^2}; \hat{\lambda}_0)}{\frac{\hat{\theta}_0}{\hat{\sigma}_0^2} f_\nu(\frac{z_i}{\hat{\sigma}_0^2}; \hat{\lambda}_0) + \sum_{k=1}^{K} \theta_k^{(t)} f_\nu(z_i; \lambda_k^{(t)})},$$

$$\tau_{ik}^{(t)} = \frac{\theta_k^{(t)} f_\nu(z_i; \lambda_k^{(t)})}{\frac{\hat{\theta}_0}{\hat{\sigma}_0^2} f_\nu(\frac{z_i}{\hat{\sigma}_0^2}; \hat{\lambda}_0) + \sum_{k=1}^{K} \theta_k^{(t)} f_\nu(z_i; \lambda_k^{(t)})}.$$

In the M-step, we need to maximize the conditional expected log likelihood

$$\sum_{i=1}^{m} \left\{ \tau_{i0}^{(t)} \left(\log \hat{\theta}_0 - \log \hat{\sigma}_0^2 + \log f_\nu\Big(\frac{z_i}{\hat{\sigma}_0^2}; \hat{\lambda}_0\Big) \right) + \sum_{k=1}^{K} \tau_{ik}^{(t)} \Big(\log \theta_k^{(t)} + \log f_\nu(z_i; \lambda_k^{(t)}) \Big) \right\}.$$

We can easily check that for $k \ge 1$

$$\theta_k^{(t+1)} = (1 - \hat{\theta}_0) \frac{\sum_{i=1}^{m} \tau_{ik}^{(t)}}{\sum_{j=1}^{K} \sum_{i=1}^{m} \tau_{ij}^{(t)}},$$

and

$$\lambda_k^{(t+1)} = \arg\max_{\lambda} \sum_{i=1}^{m} \tau_{ik}^{(t)} \log f_\nu(z_i; \lambda), \tag{4.6}$$

which can be numerically solved, e.g., using *optimize* function in R. An alternative and computationally convenient approach could be based on moment matching. Note that we can easily compute the first moment of a non-central χ^2 distribution. For $X \sim f_\nu(\cdot, \lambda)$, we can check that $\mathbb{E}(X) = \nu + \lambda$. The estimation of λ_k in equation (4.6) can be treated as a weighted log likelihood of z_i following distribution $f_\nu(\cdot, \lambda_k)$ and with weight τ_{ik}. Therefore we can empirically estimate its mean using the weighted sample average

$$\bar{\lambda}_k = \frac{\sum_{i=1}^m \tau_{ik} z_i}{\sum_{i=1}^m \tau_{ik}},$$

which can be matched to the theoretical expectation for estimation. Therefore we propose to update $\hat{\lambda}_k = \bar{\lambda}_k - \nu$ when $\bar{\lambda}_k - \nu > 0$, and keep the original λ_k otherwise.

Theoretically the computationally intensive numerical optimization approach will estimate λ_k more accurately than the computationally efficient moment matching approach. But for enrichment analysis, what matters is really the distribution functions of the gene differential expression levels, i.e., (4.2) and (A.34). In our extensive simulation studies, we have found that the moment matching and numerical optimization approaches provide comparable density estimation and yield similar enrichment analysis results.

Given an estimated finite mixture model for all genes, we can develop an EM algorithm to efficiently estimate the gene set model (4.5) based on the following recursion rule (see section A.3 for details)

$$\gamma_{j0}^{(t+1)} = \frac{1}{m_j} \sum_{i \in A_j} \frac{\gamma_{j0}^{(t)} g_0(z_i)}{\gamma_{j0}^{(t)} g_0(z_i) + (1 - \gamma_{j0}^{(t)}) g_1(z_i)}, \quad j = 1, 2.$$

where m_j is the number of genes in set A_j.

Next we conduct simulation studies to compare the performance of the proposed method to GSA [22].

4.3 Simulation study

We consider 3 groups each with n samples. For each sample we generate the expression levels of $m = 10^4$ genes from a normal distribution with expression variance σ_i^2 simulated from a χ^2 distribution with 30 degrees of freedom. We simulate gene dependence by dividing genes into 50 blocks each with 200 genes. We assume independence between blocks and within-block pairwise gene correlation equal to ρ. The first group samples always have zero mean expression for all genes. We randomly select $m\theta_0$ non-null genes with the effective size of the second and third group samples simulated from $\{\delta_j = -1.25 + 2.5 * (j-1)/9 : j = 1, \ldots, 10\}$ with relative probabilities proportional to a scaled beta distribution with parameters $a = b = 3$, Beta$((\delta_j + 1.5)/3; 3, 3)$. We fix c_0 as the 75% percentile of z_i's in empirical null estimation, and set $t_0 = 1$ in defining null gene distribution. For size evaluation, we randomly sample 1000 gene sets with each consisting of randomly selected m_e genes from all genes. The computed 1000 p-values are used to compute Type I error rate. To compute power, $m_e\theta_e$ genes are randomly sampled from the null set and $m_e(1-\theta_e)$ genes from non-null set. We repeat 1000 times to compute a p-value and power for any given Type I error α. We conducted simulations for $\rho = 0.25$, $\theta_0 = (0.9, 0.95)$, $n = (15, 25)$, and $m_e = 50, 100$. We have observed very similar patterns across different simulation settings. Here we report the results for $\theta_0 = 0.9, n = 25, m_e = 100$ and the complete results are available at Appendix A.3. We used 10000 permutations to compute significance for GSA.

Table A.12 summarizes the estimated Type I error over 100 simulations. The proposed method is denoted as Lrt-wt/Lrt-mle corresponding to moment matching and numerical optimization approaches respectively. Both methods have approximately the right size, although proposed Lrt is slightly conservative while GSA is slightly optimistic especially for small significance values.

For power comparison, we randomly sample 2000 gene sets each with m_e genes.

Table 4.1: The estimated type I error over 100 simulations (listed within parenthesis are the standard errors).

α	Lrt-wt	Lrt-mle	GSA
0.005	0.0048 (0.0002)	0.0046 (0.0002)	0.0090 (0.0003)
0.01	0.0091 (0.0003)	0.0092 (0.0003)	0.0155 (0.0004)
0.05	0.0470 (0.0007)	0.0463 (0.0007)	0.0565 (0.0007)
0.1	0.0939 (0.0009)	0.0939 (0.0009)	0.1032 (0.0009)

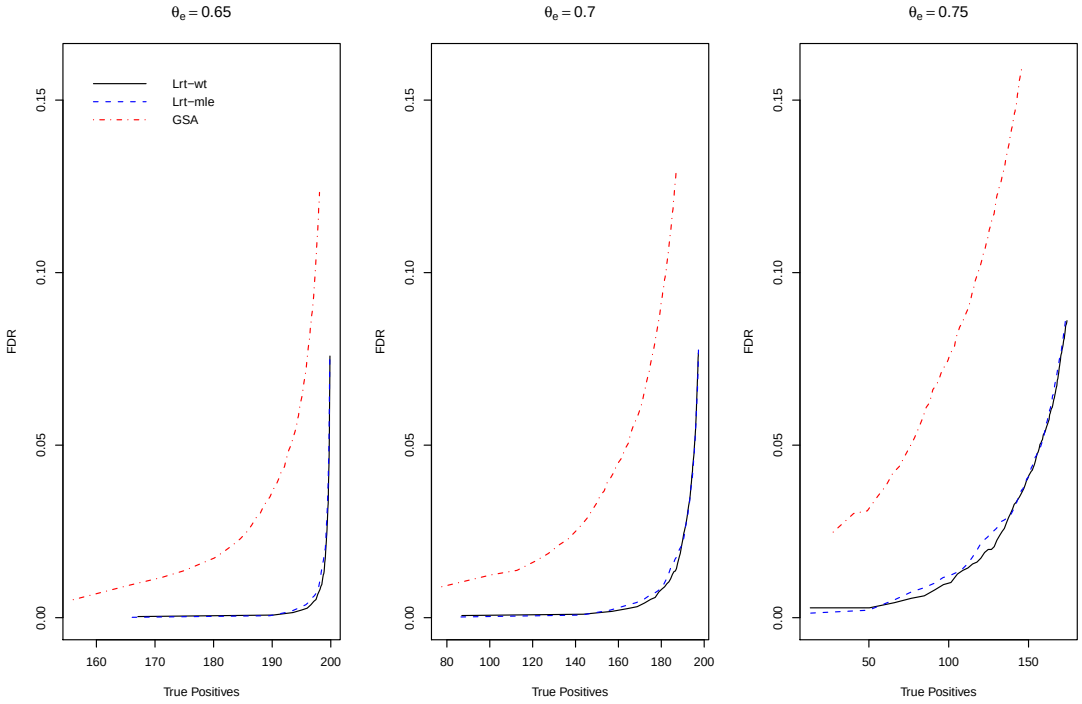
Assume 10% of the gene sets are truly enriched with each containing $m_e\theta_e$ genes randomly sampled from the null set and $m_e(1-\theta_e)$ genes from non-null set. We compute true positives and FDR based on estimated Type I error and power. Figure 4.1 shows the FDR and true positives averaged over 100 simulations. The proposed Lrt-wt and Lrt-mle yield very similar results. And both have smaller FDR and can detect more truly enriched gene sets than GSA.

In the next section, we analyze a breast cancer microarray data to illustrate the proposed method.

4.4 Application to Breast Cancer Microarray Data

The breast cancer microarray data studied in [12] were used to identify molecular signatures that correlated with response to neoadjuvant chemotherapy. It consists of 34 patients split into three groups according to clinical response to chemotherapy: complete, partial and none, with sample sizes of 10, 14 and 10, respectively. The data measured the expression profiles of 54,613 probes. We retrieved 1892 gene pathways from C2 functional collection in the Molecular Signature Database [49]. In total we analyze 1775 gene pathways that contain at least 15 genes.

Figure 4.1: FDR versus true positives averaged over 100 simulations



We first compute the moderated F-statistics summarizing 3-group differential expressions for all genes [48]. The moderated F-statistic is then transformed into chi-square statistic for enrichment analysis with the proposed method. The moment matching and numerical optimization approaches provide nearly identical results. We use 10000 permutations to compute enrichment significance for GSA.

Figure 4.2 shows the estimated FDR for the top ranked 50 gene pathways identified by both methods. Overall the proposed method (Lrt) detected much more significantly enriched pathways than GSA.

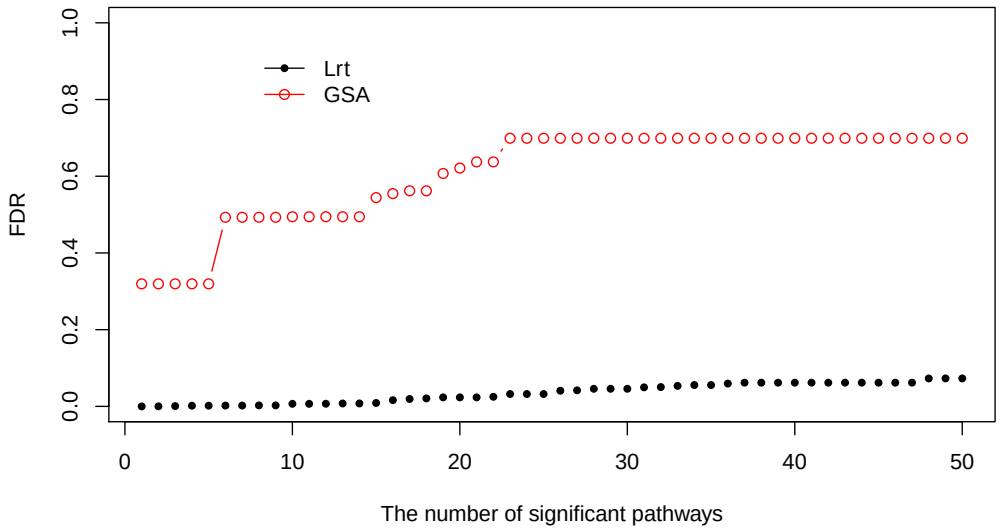
Controlling FDR at 0.1, Lrt detected 78 significant pathways, while GSA did not identify any significantly enriched gene pathway. Table 4.2 and 4.3 listed the top 78

significant pathways identified by Lrt. Many of them are directly associated with breast cancer development or involved in other human cancers or diseases. For example, the top ranked pathway, BRCA_ER_POS, consists of a set of genes positively correlated with estrogen receptor status in breast cancer [56]. BRCA_BRCA1_NEG and BRCA_ER_NEG consist of breast cancer susceptibility genes, BRCA1 and BRCA2. In normal cells, both genes are involved in the cellular response to DNA damage, including blocking cell cycle progression and helping DNA repair to ensure the stability of the genome during cell division. Certain harmful mutations of these genes lead to breast cancer [57]. Another group of pathways are related to estrogen which is directly related to breast cancer risk. For example, TFF2_KO_UP [35], IGF1RPATHWAY [5], EDG1PATHWAY [61], and FRASOR_ER_DN [24] are estrogen related pathways. BREAST_CANCER_ESTROGEN_SIGNALING contains breast cancer related genes involved in estrogen-dependent signaling [9]. PTENPATHWAY and SA_PTEN_PATHWAY are both related to PTEN, which is a tumor suppressor gene which regulates cell cycle, prevents cells from growing and dividing too rapidly (`http://www.biocarta.com/pathfiles/h_PTENPATHWAY.asp`). Frequent inactivation of the PTEN gene has been observed in breast cancer development [59].

4.5 Discussion

In this chapter we focused on detecting gene sets enriched with differentially expressed genes for multi-class microarray data. We proposed a finite chi-square mixture model for detecting multi-class differential expressions and incorporated an empirical null concept into the model to take account of gene correlation [15, 16]. We also developed a computationally very efficient moment matching based EM algorithm for model estimation. Its performance was satisfactory in our simulation studies and applications compared to the numerical optimization based EM algorithm. Overall we treated the underlying

Figure 4.2: Estimated FDR for top 50 ranked gene pathways detected by Lrt and GSA.



magnitude of differential expressions (effect size) as nuisance parameters and instead focused on estimating distributions of differential expression statistics using a discrete approximation of effect size. It will be interesting to explore an alternative approach based on a continuous distribution modeling of differential expression effect size, which could adapt to detecting more flexible enrichment patterns.

Table 4.2: Top 78 most significant pathways identified with the proposed likelihood based method

Pathways	# genes	# probes	p-value
BRCA_ER_POS	527	919	1E-15
AGUIRRE_PANCREAS_CHR9	24	59	1E-07
TARTE_PC	81	157	1E-06
PENG_RAPAMYCIN_UP	190	477	4E-06
NING_COPD_UP	180	462	5E-06
PENG_GLUTAMINE_UP	296	780	8E-06
ROSS_CBF_MYH	54	118	8E-06
CAMPTOTHECIN_PROBCELL_DN	31	56	1E-05
BREAST_CANCER_ESTROGEN_SIGNALING	101	299	1E-05
HEPARAN_SULFATE_BIOSYNTHESIS	8	17	4E-05
CHONDROITIN	8	17	4E-05
PTENPATHWAY	18	62	5E-05
TFF2_KO_UP	25	49	0.0001
SIG_PIP3_SIGNALING_IN_B_LYMPHOCYTES	36	102	0.0001
MYC_ONCOGENIC_SIGNATURE	212	378	0.0001
STEMCELL_HEMATOPOIETIC_UP	1452	2690	0.0001
ADIP_DIFF_UP	69	138	0.0002
CROMER_HYPOPHARYNGEAL_MET_VS_NON_DN	83	163	0.0002
ELECTRON_TRANSPORTER_ACTIVITY	128	279	0.0003
HYPOXIA_NORMAL_UP	219	379	0.0003
CARIES_PULP_HIGH_UP	96	186	0.0003
SIG_INSULIN_RECEPTOR_PATHWAY_IN_CARDIAC_MYOCYTES	51	160	0.0003
IGF1RPATHWAY	15	41	0.0004
ARFPATHWAY	16	38	0.0004
TNFALPHA_ADIP_DN	59	105	0.0005
BCNU_GLIOMA_NOMGMT_24HRS_UP	14	32	0.0006
GLUTATHIONE_METABOLISM	31	55	0.0006
EIF4PATHWAY	24	78	0.0007
HSA05214_GLIOMA	64	182	0.0008
HSA05223_NON_SMALL_CELL_LUNG_CANCER	54	145	0.0008
HSA05220_CHRONIC_MYELOID_LEUKEMIA	76	193	0.0009
YANG_OSTECLASTS_SIG	48	61	0.0009
HSA00030_PENTOSE_PHOSPHATE_PATHWAY	26	45	0.001
HCMVPATHWAY	16	45	0.0011
SA_PTEN_PATHWAY	17	53	0.0011
REOVIRUS_HEK293_DN	231	516	0.0012
HOFMANN_MANTEL_LYMPHOMA_VS_LYMPH_NODES_UP	50	117	0.0013
CHAUVIN_ANDROGEN_REGULATED_GENES	46	109	0.0014
EDG1PATHWAY	26	87	0.0014

Table 4.3: Top 78 most significant pathways identified with the proposed likelihood based method(Conti.)

Pathways	# genes	# probes	p-value
HSA00051_FRUCTOSE_AND_MANNOSE_METABOLISM	42	83	0.0014
MARCINIAK_CHOP_DIFF	26	34	0.0015
ICHIBA_GVHD	335	442	0.0015
LONGEVITYPATHWAY	14	35	0.0016
HSA05212_PANCREATIC_CANCER	73	178	0.0016
TRKAPATHWAY	14	37	0.0016
O_GLYCAN_BIOSYNTHESIS	16	32	0.0016
HPV31_DN	48	77	0.0016
PLCPATHWAY	8	23	0.002
FRASOR_ER_DN	74	163	0.0021
CARIES_PULP_UP	212	424	0.0021
METHOTREXATE_PROBCELL_DN	14	23	0.0021
BRCA_BRCA1_NEG	154	256	0.0022
CREB_BRAIN_2WKS_UP	24	46	0.0024
PENTOSE_PHOSPHATE_PATHWAY	25	47	0.0024
ARGININE_AND_PROLINE_METABOLISM	46	84	0.0026
SA_REG_CASCADE_OF_CYCLIN_EXPR	13	27	0.0026
KANNAN_P53_UP	40	82	0.0026
ZHAN_PCS_MULTIPLE_MYELOMA_SPKD	25	34	0.0026
CELL_ADHESION_MOLECULE_ACTIVITY	120	275	0.003
GLEEVECPATHWAY	22	58	0.003
RACCYCDPATHWAY	23	57	0.003
HESS_HOXAANMEIS1_DN	70	116	0.0031
HESS_HOXAANMEIS1_UP	70	116	0.0031
HSA05218_MELANOMA	71	178	0.0031
HSA00480_GLUTATHIONE_METABOLISM	39	62	0.0031
CANCERDRUGS_PROBCELL_DN	15	27	0.0032
BRENTANI_CELL_CYCLE	86	219	0.0033
SIG_PIP3_SIGNALING_IN_CARDIAC_MYOCTES	67	199	0.0033
HSA04115_P53_SIGNALING_PATHWAY	68	175	0.0035
HSA05219_BLADDER_CANCER	42	98	0.0037
ALCALAY_AML_NPMC_UP	139	280	0.0039
BCNU_GLIOMA_NOMGMT_48HRS_UP	18	35	0.0041
AGEING_KIDNEY_SPECIFIC_UP	189	357	0.0042
CELL_ADHESION	201	446	0.0042
INSULIN_ADIP_SENS_UP	13	23	0.0042
HSA04620_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	102	199	0.0042
BRCA_ER_NEG	929	1644	0.0043
BCNU_GLIOMA_MGMT_48HRS_UP	18	37	0.0044

Chapter 5

Conclusion and Discussion

We have proposed a parametric modeling approach to gene set enrichment analysis, which is intended to meet the biological research question for analyzing and interpreting large-scale gene expression data in a microarray experiment. Chapter 2 develops a likelihood based method for analyzing overall enrichment and transforms enrichment analysis into a model comparison problem, which can be addressed using the likelihood ratio test approach. In Chapter 3, we extend the likelihood based approach to enrichment analysis of one-sided differential expressions. Furthermore, Chapter 4 propose a finite chi-square mixture model for detecting multi-class differential expressions and a computationally efficient moment matching based EM algorithm for model estimation. Application to real gene expression data, along with simulation studies, demonstrate the competitive performance of the proposed methods over GSA which is the-state-of-the-art of this field.

The proposed likelihood based methods based on a finite mixture model use a discrete approximation of the underlying magnitude of differential expressions (effective size) instead of estimating distributions of differential expression statistics. For many microarray data, the effective size is more likely to be on a continuous scale over hundreds of thousands of genes. Currently, we are exploring an alternative approach based on a continuous distribution modeling of differential expression effect size and we expect the method could adapt to detect more flexible enrichment patterns. In addition the distribution of the effective size is often of inferential interest itself [20], for example, we want to assess more flexible patterns of differential expression levels, e.g., only identifying those genes with a (pre-defined) large magnitude of differential expression. However, it has been largely ignored or treated as nuisance parameters in most existing approaches. Recently [45] proposed a novel semi-parametric modeling approach that calculates p-values parametrically assuming t-distribution and approximates the effect

size distribution non-parametrically with B-splines. It is interesting to study an alternative semi-parametric modeling approach useing t-distribution and B-splines to directly approximate the test statistic distribution.

References

[1] Alcalay, M., Tiacci, E., Bergomas, R., Bigerna, B., Venturini, E., Minardi, S. P., Meani, N., Diverio, D., Bernard, L., Tizzoni, L., Volorio, S., Luzi, L., Colombo, E., Lo Coco, F., Mecucci, C., Falini, B., Pelicci, P. G., and for the Gruppo Italiano Malattie Ematologiche Maligne dell'Adulto (GIMEMA) Acute Leukemia Working Party. Acute myeloid leukemia bearing cytoplasmic nucleophosmin (NPMc+ AML) shows a distinct gene expression profile characterized by up-regulation of genes involved in stem-cell maintenance. *Blood 106*, 3 (2005), 899–902.

[2] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics 25*, 1 (May 2000), 25–29.

[3] Barry, W. T., Nobel, A. B., and Wright, F. A. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics 21*, 9 (May 2005), 1943–1949.

[4] Benjamini, Y., and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical*

Society. Series B (Methodological) 57, 1 (1995), pp. 289–300.

[5] Camirand, A., Zakikhani, M., Young, F., and Pollak, M. Inhibition of insulin-like growth factor-1 receptor signaling enhances growth-inhibitory and proapoptotic effects of gefitinib (iressa) in human breast cancer cells. *Breast Cancer Res 7*, 4 (2005).

[6] Chen, W., Kumar, A. R., Hudson, W. A., Li, Q., Wu, B., Staggs, R. A., Lund, E. A., Sam, T. N., and Kersey, J. H. Malignant transformation initiated by mll-af9: Gene dosage and critical target cells. *Cancer Cell 13*, 5 (May 2008), 432–440.

[7] Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J., and Foa, R. Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood 103*, 7 (2004), 2771–2778.

[8] Collins, K., Jacks, T., and Pavletich, N. P. The cell cycle and cancer. *Proceedings of the National Academy of Sciences of the United States of America 94*, 7 (1997), 2776–2778.

[9] Cos, S., González, A., Martínez-Campa, C., Mediavilla, M. D. D., Alonso-González, C., and Sánchez-Barceló, E. J. Estrogen-signaling pathway: a link between breast cancer and melatonin oncostatic actions. *Cancer detection and prevention 30*, 2 (2006), 118–128.

[10] Cui, X., Hwang, J. T. G., Qiu, J., Blades, N. J., and Churchill, G. A. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics 6*, 1 (2005), 59–75.

[11] Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*, 1 (1977), 1–38.

[12] Dressman, H. K., Hans, C., Bild, A., Olson, J. A., Rosen, E., Marcom, P. K., Liotcheva, V. B., Jones, E. L., Vujaskovic, Z., Marks, J., Dewhirst, M. W., West, M., Nevins, J. R., and Blackwell, K. Gene expression profiles of multiple breast cancer phenotypes and response to neoadjuvant chemotherapy. *Clin Cancer Res 12*, 3 Pt 1 (February 2006), 819–826.

[13] Dudoit, R., Yang, Y. H., Callow, M. J., and Speed, T. P. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica Sinica 12* (2002), 111–139.

[14] Efron, B. Robbins, empirical bayes and microarrays. *The Annals of Statistics 31*, 2 (2003), pp. 366–378.

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

[16] Efron, B. Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association 102*, 477 (March 2007), 93–103.

[17] Efron, B. Size, power and false discovery rates. *The Annals of Statistics 35* (2007), 1351–1377.

[18] Efron, B. Size, power and false discovery rates. *The Annals of Statistics 35*, 4 (August 2007), 1351–1377.

[19] Efron, B. Microarrays, empirical bayes and the two-groups model.

[20] Efron, B. Empirical bayes estimates for large-scale prediction problems. *Journal of the American Statistical Association 104*, 487 (2009), 1015–1028.

[21] Efron, B., and Tibshirani, R. On testing the significance of sets of genes. *Annals of Applied Statistics 2007* (2006).

[22] Efron, B., and Tibshirani, R. On testing the significance of sets of genes. *Ann. Appl. Stat. 1*, 1 (2007), 107–129.

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

[24] Frasor, J., Stossi, F., Danes, J. M., Komm, B., Lyttle, C. R., and Katzenellenbogen, B. S. Selective Estrogen Receptor Modulators. *Cancer Research 64*, 4 (2004), 1522–1533.

[25] Goeman, J. J., and Bühlmann, P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics (Oxford, England) 23*, 8 (April 2007), 980–987.

[26] Gottardo, R., Raftery, A. E., Yee Yeung, K., and Bumgarner, R. E. Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biometrics 62*, 1 (2006), 10–18.

[27] Ibrahim, J. G., Chen, M.-H., and Gray, R. J. Bayesian models for gene expression with dna microarray data. *Journal of the American Statistical Association 97*, 457 (2002), 88–99.

[28] Ishwaran, H., and Rao, J. S. Detecting differentially expressed genes in microarrays using bayesian model selection. *Journal of the American Statistical Association 98*, 462 (2003), 438–455.

[29] Jiao, S., and Zhang, S. The t-mixture model approach for detecting differentially expressed genes in microarrays. *Functional & Integrative Genomics 8* (2008), 181–186.

[30] Kendziorski, C. M., Newton, M. A., Lan, H., and Gould, M. N. On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine 22*, 24 (2003), 3899–3914.

[31] Khatri, P., and Draghici, S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics 21*, 18 (September 2005), 3587–3595.

[32] Kumar, A. R., Hudson, W. A., Chen, W., Nishiuchi, R., Yao, Q., and Kersey, J. H. Hoxa9 influences the phenotype but not the incidence of Mll-AF9 fusion gene leukemia. *Blood 103*, 5 (2004), 1823–1828.

[33] Kumar, A. R., Li, Q., Hudson, W. A., Chen, W., Sam, T., Yao, Q., Lund, E. A., Wu, B., Kowal, B. J., and Kersey, J. H. A role for meis1 in mll-fusion gene leukemia. *Blood 113*, 8 (2009), 1756–1758.

[34] Lnnstedt, I., and Speed, T. Replicated microarray data. *Statistica Sinica 12* (2001), 31–46.

[35] May, F. E. B., Semple, J. I., Prest, S. J., and Westley, B. R. Expression and motogenic activity of tff2 in human breast cancer cells. *Peptides 25*, 5 (2004), 865 – 872. Molecular Medicine of TFF Peptides: Mucosal Protection and Repair, and More.

[36] McLachlan, G., Bean, R., and Jones, L. B.-T. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics 22*, 13 (2006), 1608–1615.

[37] Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. PGC-1-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics 34*, 3 (June 2003), 267–273.

[38] Nakamura, T., Largaespada, D. A., Lee, M. P., Johnson, L. A., Ohyashiki, K., Toyama, K., Chen, S. J., Willman, C. L., Chen, I.-M., Feinberg, A. P., Jenkins, N. A., Copeland, N. G., and Shaughnessy, J. D. Fusion of the nucleoporin gene nup98 to hoxa9 by the chromosome translocation t(7;11)(p15;p15) in human myeloid leukaemia. *Nature Genetics 12* (1996), 154–158.

[39] Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics 5*, 2 (2004), 155–176.

[40] Newton, M. A., Quintana, F. A., Den Boon, J. A., Sengupta, S., and Ahlquist, P. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Ann. Appl. Stat. 1*, 1 (2007), 85–106.

[41] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl. Acids Res. 27*, 1

(1999), 29–34.

[42] Pan, W., Lin, J., and Le, C. T. A mixture model approach to detecting differentially expressed genes with microarray data. *Functional & Integrative Genomics 3* (2003), 117–124.

[43] Pavlidis, P., Qin, J., Arango, V., Mann, J. J., and Sibille, E. Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. *Neurochemical Research 29* (2004), 1213–1222.

[44] Ross, M. E., Mahfouz, R., Onciu, M., Liu, H.-C., Zhou, X., Song, G., Shurtleff, S. A., Pounds, S., Cheng, C., Ma, J., Ribeiro, R. C., Rubnitz, J. E., Girtman, K., Williams, W. K., Raimondi, S. C., Liang, D.-C., Shih, L.-Y., Pui, C.-H., and Downing, J. R. Gene expression profiling of pediatric acute myelogenous leukemia. *Blood 104*, 12 (2004), 3679–3687.

[45] Ruppert, D., Nettleton, D., and Hwang, J. T. Exploring the information in p-values for the analysis and planning of multiple-test experiments. *Biometrics 63*, 2 (June 2007), 483–495.

[46] Sachs, L. The control of hematopoiesis and leukemia: from basic biology to the clinic. *Proceedings of the National Academy of Sciences of the United States of America 93*, 10 (1996), 4742–4749.

[47] Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics 6*, 2 (1978), 461–464.

[48] Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology 3*, 1 (2004).

[49] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America 102*, 43 (2005), 15545–15550.

[50] Takeda, A., Goolsby, C., and Yaseen, N. R. NUP98-HOXA9 Induces Long-term Proliferation and Blocks Differentiation of Primary Human CD34+ Hematopoietic Cells. *Cancer Research 66*, 13 (2006), 6628–6637.

[51] Thirman, M. J., Gill, H. J., Burnett, Robert C.and Mbangkollo, D., McCabe, N. R., Kobayashi, H., Ziemin-van der Poel, S., Kaneko, Y., Morgan, R., Sandberg, A. A., Chaganti, R., Larson, R. A., Le Beau, M. M., Diaz, M. O., and Rowley, J. D. Rearrangement of the MLL Gene in Acute Lymphoblastic and Acute Myeloid Leukemias with 11q23 Chromosomal Translocations. *N Engl J Med 329* (1993), 909–914.

[52] Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America 102*, 38 (September 2005), 13544–13549.

[53] Townsend, J. P., and Hartl, D. L. Bayesian analysis of gene expression levels: statistical quantification of relative mrna level across multiple strains or treatments.

[54] Troyanskaya, O. G., Garber, M. E., Brown, P. O., Botstein, D., and Altman, R. B. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics 18*, 11 (2002), 1454–1461.

[55] Tusher, V. G., Tibshirani, R., and Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences 98*, 9 (2001), 5116–5121.

[56] van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. Gene expression profiling predicts clinical outcome of breast cancer. *Nature 415*, 6871 (January 2002), 530–536.

[57] Venkitaraman, A. R. Cancer susceptibility and the functions of brca1 and brca2. *Cell 108* (2002), 171–182.

[58] Verhaak, R. G. W., Goudswaard, C. S., van Putten, W., Bijl, M. A., Sanders, M. A., Hugens, W., Uitterlinden, A. G., Erpelinck, C. A. J., Delwel, R., Lowenberg, B., and Valk, P. J. M. Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance. *Blood 106*, 12 (2005), 3747–3754.

[59] Weng, L.-P., Smith, W. M., Dahia, P. L. M., Ziebold, U., Gil, E., Lees, J. A., and Eng, C. PTEN Suppresses Breast Cancer Cell Growth by Phosphatase Activity-dependent G1 Arrest followed by Cell Death. *Cancer Research 59*, 22 (1999), 5808–5814.

[60] William T. Barry, A. B. N., and Wright, F. A. A statistical framework for testing functional categories in microarray data. *The Annals of Applied Statistics 2* (2008), 286–315.

[61] Wittmann, B. M., Wang, N., and Montano, M. M. Identification of a Novel Inhibitor of Breast Cell Growth That Is Down-Regulated by Estrogens and Decreased in Breast Tumors. *Cancer Research 63*, 16 (2003), 5151–5158.

[62] Wu, B. Differential gene expression detection using penalized linear regression models: the improved sam statistics. *Bioinformatics 21*, 8, 1565–1571.

[63] Wu, B., Guan, Z., and Zhao, H. Parametric and nonparametric fdr estimation revisited. *Biometrics 62*, 3 (2006), 735–744.

Appendix A

This appendix contains EM algorithms for model estimaion and full results of simulation studies under different settings to complement the simulation studies of Chapter 2through 4. In addition it provides the complete lists of identified gene pathways as a result of leukemia data analysis in Chapther 3.

A.1 GSEA with a finite mixture model

EM algorithm for finite mixture model fitting

We begin with the finite mixture model in (2.1) given $(\hat{\theta}_0, \hat{\mu}_0, \hat{\sigma}_0^2)$ and K. Define indicators $w_{ik} \in \{0, 1\}$ following a multinomial distribution

$$\Pr(w_{ik} = 1) = \theta_k, \quad \sum_{k=0}^{K} w_{ik} = 1,$$

and conditionally we assume

$$z_i | w_{ik} = 1 \sim f_k.$$

The complete data likelihood function for (z_i, w_{ik}) can be written as

$$\prod_{i=1}^{m} \{\hat{\theta}_0 f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2)\}^{w_{i0}} \prod_{k=1}^{K} \{\theta_k \phi(z_i - \mu_k)\}^{w_{ik}}.$$

In E-step, the conditional probabilities given current estimates of the parameter $(\theta_k^{(t)}, \mu_k^{(t)})$ can be checked to be

$$T_{0,i}^{(t)} = \frac{\hat{\theta}_0 f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2)}{\hat{\theta}_0 f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{k=1}^{K} \theta_k^{(t)} \phi(z_i - \mu_k^{(t)})}, \quad T_{k,i}^{(t)} = \frac{\theta_k^{(t)} \phi(z_i - \mu_k)}{\hat{\theta}_0 f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{j=1}^{K} \theta_j^{(t)} \phi(z_i - \mu_j^{(t)})}.$$

In M-step, the conditional expected log likelihood can be checked to be proportional to

$$\sum_{i=1}^{m} \left\{ T_{0,i}^{(t)} \left(\log \hat{\theta}_0 - \log \hat{\sigma}_0 - \frac{(z_i - \hat{\mu}_0)^2}{2\hat{\sigma}_0^2} \right) + \sum_{k=1}^{K} T_{k,i}^{(t)} \left(\log \theta_k^{(t)} - \frac{(z_i - \mu_k^{(t)})^2}{2} \right) \right\},$$

which can be easily verified to be maximized by

$$\theta_k^{(t+1)} = (1 - \hat{\theta}_0) \frac{\sum_{i=1}^{m} T_{k,i}^{(t)}}{\sum_{j=1}^{K} \sum_{i=1}^{m} T_{j,i}^{(t)}}, \quad \mu_k^{(t+1)} = \frac{\sum_{i=1}^{m} T_{k,i}^{(t)} z_i}{\sum_{i=1}^{m} T_{k,i}^{(t)}}, \quad k \geq 1.$$

Given only $(\hat{\mu}_0, \hat{\sigma}_0^2)$ with θ_0 also being a parameter, we have

$$(\theta_0^{(t+1)}, \cdots, \theta_K^{(t+1)}) = \arg\max_{\theta_k} \sum_{k=0}^{K} \sum_{i=1}^{m} T_{k,i}^{(t)} \log \theta_k, \quad \mu_k^{(t+1)} = \arg\min_{\mu_k} \sum_{i=1}^{m} \sum_{k=1}^{K} T_{k,i}^{(t)} (z_i - \mu_k)^2, \quad k > 0.$$

We can easily check that

$$\theta_k^{(t+1)} = \frac{1}{m} \sum_{i=1}^{m} T_{k,i}^{(t)}, \quad k \geq 0, \quad \mu_k^{(t+1)} = \frac{\sum_{i=1}^{m} T_{k,i}^{(t)} z_i}{\sum_{i=1}^{m} T_{k,i}^{(t)}}, \quad k > 0.$$

EM algorithm for gene set model estimation

The complete data likelihood function for gene set A given $(\hat{\theta}_0, \hat{\mu}_0, \hat{\sigma}_0^2, \hat{\mu}_k)$ is

$$\prod_{i \in A} \{\nu_0 f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2)\}^{w_{i0}} \prod_{k=1}^{K} \{\nu_k \phi(z_i - \hat{\mu}_k)\}^{w_{ik}}.$$

The conditional expected log likelihood can be easily checked to be

$$\sum_{i \in A} \{T_{0,i}^{(t)} \log \nu_0 + \sum_{k=1}^{K} T_{k,i}^{(t)} \log \nu_k\},$$

where

$$T_{0,i}^{(t)} = \frac{\nu_0^{(t)} f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2)}{\nu_0^{(t)} f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{k=1}^{K} \nu_k^{(t)} \phi(z_i - \hat{\mu}_k)}, T_{k,i}^{(t)} = \frac{\nu_k^{(t)} \phi(z_i - \hat{\mu}_k)}{\nu_0^{(t)} f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{j=1}^{K} \nu_j^{(t)} \phi(z_i - \hat{\mu}_j)}, \quad k \geq 1.$$

We can easily verify that

$$\nu_k^{(t+1)} = \frac{\sum_{i \in A} T_{k,i}^{(t)}}{m_A}, \quad k \geq 0.$$

EM algorithm for model fitting under no enrichment

The complete data likelihood can be written as

$$\prod_{i \in A} \{\nu_0 f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2)\}^{w_{i0}} \prod_{k=1}^{K} \{\nu_{1k} \phi(z_i - \hat{\mu}_k)\}^{w_{ik}} \prod_{j \in A^c} \{\nu_0 f_0(z_j; \hat{\mu}_0, \hat{\sigma}_0^2)\}^{w_{j0}} \prod_{k=1}^{K} \{\nu_{2k} \phi(z_j - \hat{\mu}_k)\}^{w_{jk}},$$

where $\nu_0 + \sum_{k=1}^{K} \nu_{lk} = 1$, $l = 1, 2$. The conditional expected log likelihood can be easily checked to be

$$\sum_{i \in A} \{T_{0,i}^{(t)} \log \nu_0 + \sum_{k=1}^{K} T_{k,i}^{(t)} \log \nu_{1k}\} + \sum_{j \in A^c} \{T_{0,j}^{(t)} \log \nu_0 + \sum_{k=1}^{K} T_{k,j}^{(t)} \log \nu_{2k}\},$$

where

$$\begin{aligned}
T_{0,i}^{(t)} &= \frac{\nu_0^{(t)} f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2)}{\nu_0^{(t)} f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{l=1}^{K} \nu_{1l}^{(t)} \phi(z_i - \hat{\mu}_l)}, \quad i \in A, \\
T_{k,i}^{(t)} &= \frac{\nu_{1k}^{(t)} \phi(z_i - \hat{\mu}_g)}{\nu_0^{(t)} f_0(z_i; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{l=1}^{K} \nu_{1l}^{(t)} \phi(z_i - \hat{\mu}_l)}, \quad i \in A, \quad k > 0, \\
T_{0,j}^{(t)} &= \frac{\nu_0^{(t)} f_0(z_j; \hat{\mu}_0, \hat{\sigma}_0^2)}{\nu_0^{(t)} f_0(z_j; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{l=1}^{K} \nu_{2l}^{(t)} \phi(z_j - \hat{\mu}_l)}, \quad j \in A^c, \\
T_{k,j}^{(t)} &= \frac{\nu_{2k}^{(t)} \phi(z_j - \hat{\mu}_k)}{\nu_0^{(t)} f_0(t_j; \hat{\mu}_0, \hat{\sigma}_0^2) + \sum_{l=1}^{K} \nu_{2l}^{(t)} \phi(z_j - \hat{\mu}_l)}, \quad j \in A^c, \quad k > 0.
\end{aligned}$$

To maximize the conditional log likelihood, we use the Lagrange multiplier method

$$\begin{aligned}
Q &= \sum_{i \in A} \{T_{0,i}^{(t)} \log \nu_0^{(t)} + \sum_{k=1}^{K} T_{k,i}^{(t)} \log \nu_{1k}^{(t)}\} + \sum_{j \in A^c} \{T_{0,j}^{(t)} \log \nu_0^{(t)} + \sum_{k=1}^{K} T_{k,j}^{(t)} \log \nu_{2k}^{(t)}\} \\
&- \lambda_1 (\nu_0^{(t)} + \sum_{k=1}^{K} \nu_{1k}^{(t)} - 1) - \lambda_2 (\nu_0^{(t)} + \sum_{k=1}^{K} \nu_{2k}^{(t)} - 1).
\end{aligned}$$

Setting the gradient vector $\nabla Q = 0$ yields the following equations

$$\begin{aligned}
\frac{\partial Q}{\partial \nu_0^{(t)}} &= \frac{\sum_{i \in A} T_{0,i}^{(t)}}{\nu_0^{(t)}} + \frac{\sum_{j \in A^c} T_{0,h}^{(t)}}{\nu_0^{(t)}} - \lambda_1 - \lambda 2 = 0, \\
\frac{\partial Q}{\partial \nu_{1k}^{(t)}} &= \frac{\sum_{i \in A} T_{k,i}^{(t)}}{\nu_{1k}^{(t)}} - \lambda_1 = 0, \qquad k > 0, \\
\frac{\partial Q}{\partial \nu_{2k}^{(t)}} &= \frac{\sum_{j \in A^c} T_{k,j}^{(t)}}{\nu_{2k}^{(t)}} - \lambda_2 = 0, \qquad k > 0, \\
\frac{\partial Q}{\partial \lambda_1} &= \nu_0^{(t)} + \sum_{k=1}^{K} \nu_{1k}^{(t)} - 1 = 0, \\
\frac{\partial Q}{\partial \lambda_2} &= \nu_0^{(t)} + \sum_{k=1}^{K} \nu_{2k}^{(t)} - 1 = 0.
\end{aligned}$$

From the first three equations we can obtain

$$\nu_0^{(t)} = \frac{\sum_{i\in A} T_{0,i}^{(t)} + \sum_{j\in A^c} T_{0,j}^{(t)}}{\lambda_1 + \lambda_2}, \quad \nu_{1k}^{(t)} = \frac{\sum_{i\in A} T_{0,i}^{(t)}}{\lambda_1}, \quad \nu_{2k}^{(t)} = \frac{\sum_{j\in A^c} T_{0,j}^{(t)}}{\lambda 2}, \quad k > 0.$$

When plugging these into the last two equations, we can easily verify that

$$\begin{aligned}
\nu_0^{(t+1)} &= \frac{1}{m}(\sum_{i\in A} T_{0,i}^{(t)} + \sum_{j\in A^c} T_{0,j}^{(t)}), \\
\nu_{1k}^{(t+1)} &= (1-\nu_0^{(t)}) \frac{\sum_{i\in A} T_{k,i}^{(t)}}{\sum_{l=1}^{K}\sum_{i\in A} T_{l,i}^{(t)}}, \\
\nu_{2k}^{(t+1)} &= (1-\nu_0^{(t)}) \frac{\sum_{j\in A^c} T_{k,j}^{(t)}}{\sum_{l=1}^{K}\sum_{j\in A^c} T_{l,j}^{(t)}}.
\end{aligned}$$

Simulation study

We report additional results of simulation study in Chapter 2. We consider all 16 different simulation settings: $\theta_0 = 0.9, 0.95$, $m_e = 50, 100$, $\rho = 0.2, 0.7$ and $n = 25, 50$ as listed inTable A.1.

Table A.2 summarizes the estimated Type I error over 100 simulations. Both Lrt and GSA methods have approximately the right size.

Figure A.1 to A.16 show the estimated FDR and true positives averaged over 100 simulations. Overall, the proposed Lrt can identify more enriched gene sets for any given FDR than GSA.

Table A.1: The 16 different simulation settings

θ_0	m_e	ρ	n	scenarios
0.9	50	0.2	25	scenario 1
			50	scenario 2
		0.7	25	scenario 3
			50	scenario 4
	100	0.2	25	scenario 5
			50	scenario 6
		0.7	25	scenario 7
			50	scenario 8
0.95	50	0.2	25	scenario 9
			50	scenario 10
		0.7	25	scenario 11
			50	scenario 12
	100	0.2	25	scenario 13
			50	scenario 14
		0.7	25	scenario 15
			50	scenario 16

Table A.2: Estimated type I error of Lrt and GSA over 100 simulations (listed within parenthesis are the standard errors)

	Lrt				GSA			
α	0.005	0.01	0.05	0.1	0.005	0.01	0.05	0.1
Scenario 1	0.004	0.009	0.045	0.089	0.006	0.012	0.054	0.104
	(0.0002)	(0.0003)	(0.0007)	(0.001)	(0.0003)	(0.0004)	(0.0007)	(0.001)
Scenario 2	0.005	0.009	0.047	0.096	0.008	0.013	0.057	0.106
	(0.0002)	(0.0003)	(0.0007)	(0.001)	(0.0003)	(0.0004)	(0.0007)	(0.001)
Scenario 3	0.004	0.009	0.045	0.09	0.007	0.013	0.055	0.105
	(0.0002)	(0.0003)	(0.0008)	(0.0009)	(0.0002)	(0.0004)	(0.0008)	(0.001)
Scenario 4	0.005	0.009	0.047	0.095	0.008	0.015	0.058	0.108
	(0.0002)	(0.0003)	(0.0008)	(0.001)	(0.0003)	(0.0004)	(0.0007)	(0.0009)
Scenario 5	0.005	0.009	0.045	0.09	0.006	0.011	0.054	0.104
	(0.0002)	(0.0003)	(0.0006)	(0.0009)	(0.0002)	(0.0003)	(0.0007)	(0.0011)
Scenario 6	0.004	0.009	0.046	0.094	0.007	0.013	0.055	0.106
	(0.0002)	(0.0003)	(0.0006)	(0.001)	(0.0002)	(0.0003)	(0.0006)	(0.0009)
Scenario 7	0.004	0.009	0.045	0.09	0.006	0.012	0.054	0.104
	(0.0002)	(0.0003)	(0.0007)	(0.0009)	(0.0003)	(0.0003)	(0.0008)	(0.0011)
Scenario 8	0.005	0.009	0.047	0.094	0.007	0.013	0.056	0.106
	(0.0002)	(0.0003)	(0.0007)	(0.001)	(0.0003)	(0.0003)	(0.0007)	(0.001)
Scenario 9	0.004	0.009	0.045	0.09	0.006	0.012	0.053	0.104
	(0.0002)	(0.0003)	(0.0006)	(0.0009)	(0.0002)	(0.0003)	(0.0007)	(0.001)
Scenario 10	0.005	0.009	0.044	0.089	0.008	0.014	0.059	0.109
	(0.0002)	(0.0003)	(0.0007)	(0.0011)	(0.0003)	(0.0003)	(0.0007)	(0.0008)
Scenario 11	0.004	0.009	0.044	0.089	0.007	0.013	0.054	0.104
	(0.0002)	(0.0003)	(0.0008)	(0.0012)	(0.0003)	(0.0004)	(0.0007)	(0.0009)
Scenario 12	0.005	0.009	0.045	0.089	0.009	0.015	0.059	0.109
	(0.0002)	(0.0003)	(0.0008)	(0.0011)	(0.0003)	(0.0004)	(0.0008)	(0.001)
Scenario 13	0.004	0.008	0.045	0.09	0.006	0.011	0.053	0.104
	(0.0002)	(0.0003)	(0.0006)	(0.0009)	(0.0002)	(0.0003)	(0.0006)	(0.0009)
Scenario 14	0.005	0.009	0.047	0.094	0.007	0.012	0.056	0.106
	(0.0002)	(0.0003)	(0.0007)	(0.001)	(0.0002)	(0.0004)	(0.0006)	(0.0009)
Scenario 15	0.004	0.009	0.044	0.088	0.006	0.012	0.053	0.103
	(0.0002)	(0.0003)	(0.0007)	(0.001)	(0.0002)	(0.0003)	(0.0007)	(0.001)
Scenario 16	0.005	0.01	0.046	0.093	0.007	0.014	0.056	0.106
	(0.0002)	(0.0003)	(0.0007)	(0.001)	(0.0003)	(0.0004)	(0.0007)	(0.0011)

Figure A.1: (scenario 1) $\theta_0 = 0.9, m_e = 50, \rho = 0.2$, n=25

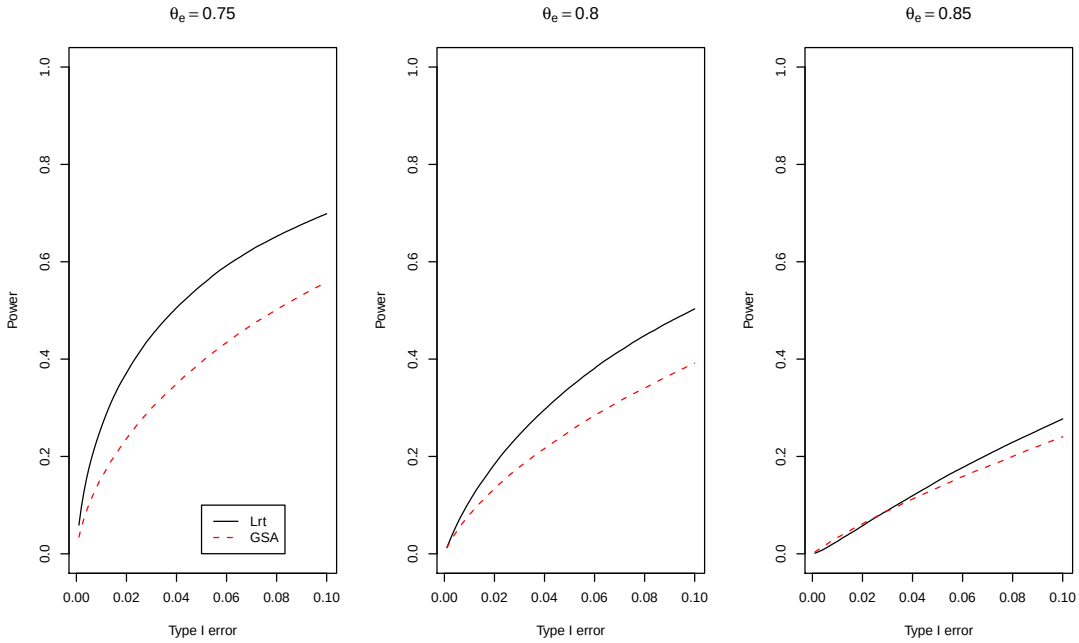


Figure A.2: (scenario 2) $\theta_0 = 0.9, m_e = 50, \rho = 0.2$, n=50

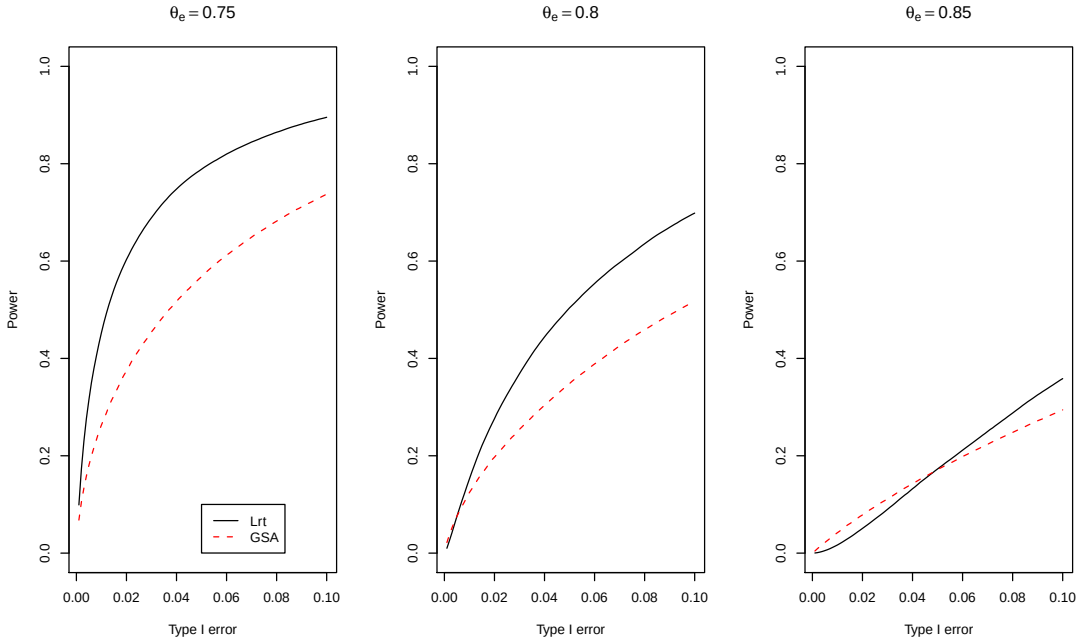


Figure A.3: (scenario 3) $\theta_0 = 0.9, m_e = 50, \rho = 0.7$, n=25

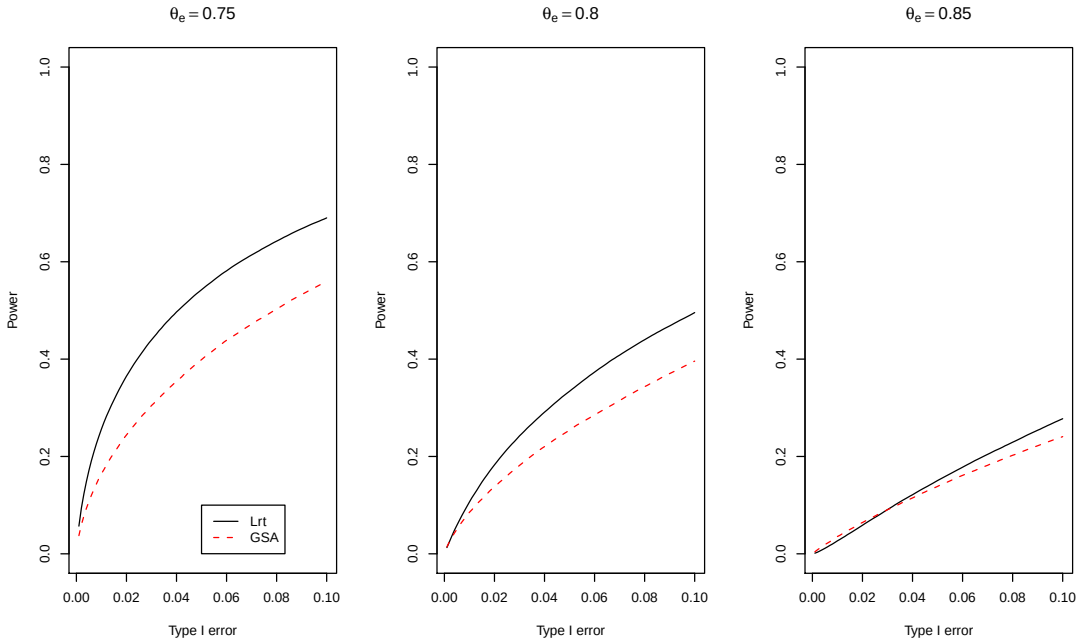


Figure A.4: (scenario 4) $\theta_0 = 0.9, m_e = 50, \rho = 0.7$, n=50

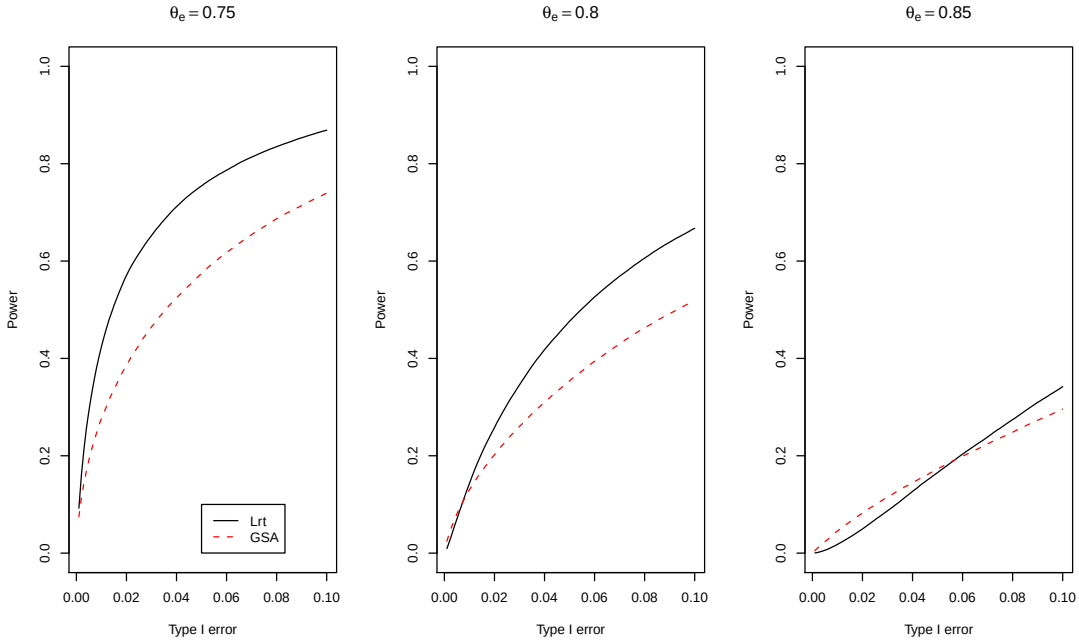


Figure A.5: (scenario 5) $\theta_0 = 0.9, m_e = 100, \rho = 0.2$, n=25

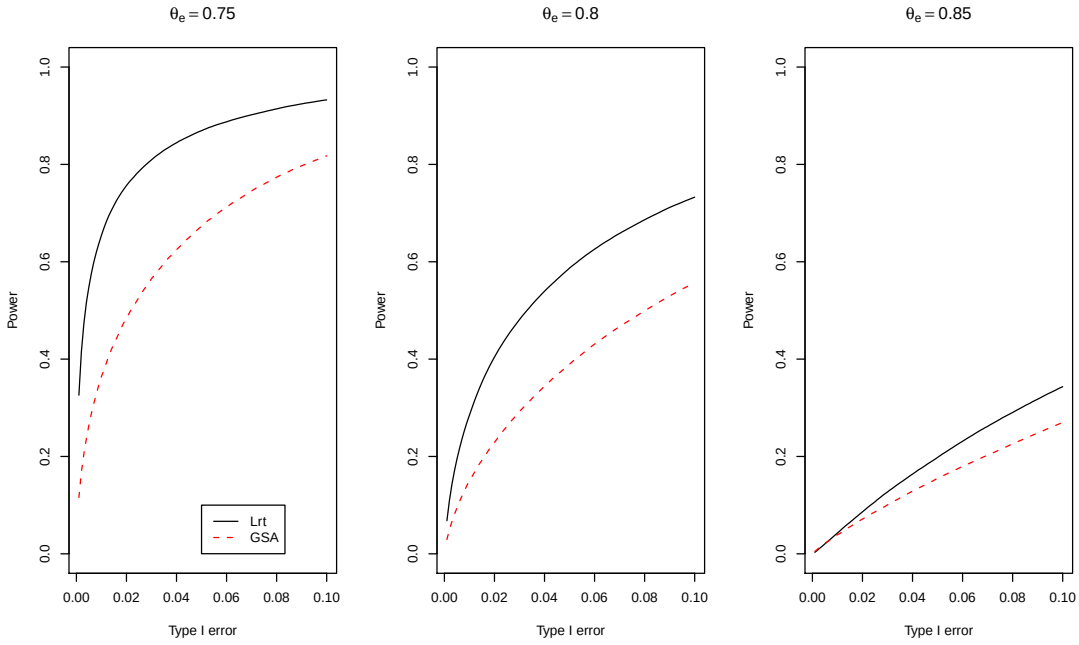


Figure A.6: (scenario 6) $\theta_0 = 0.9, m_e = 100, \rho = 0.2$, n=50

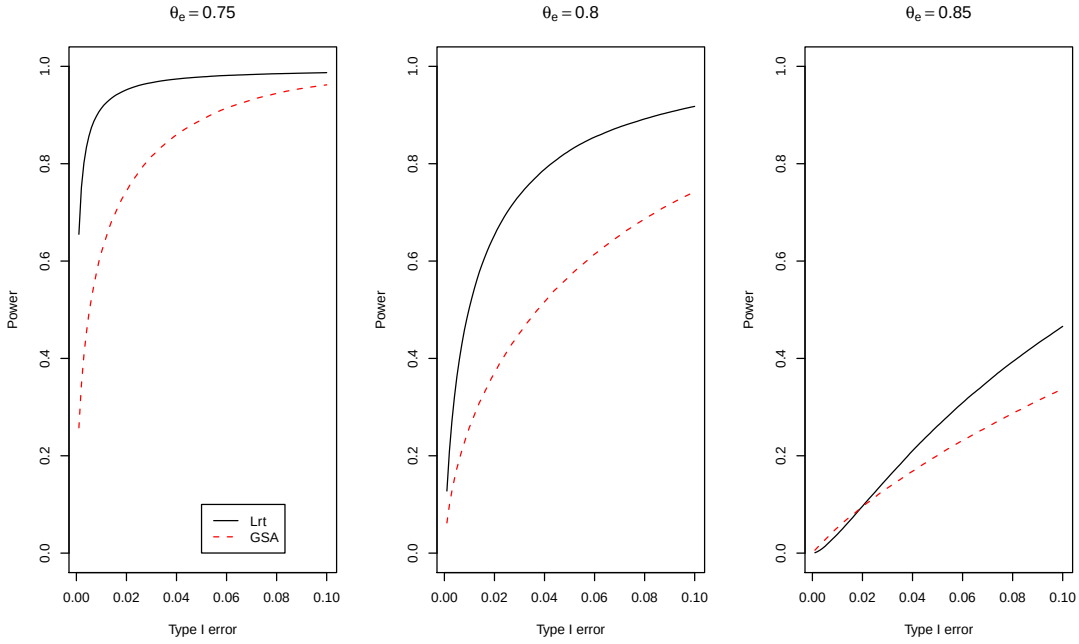


Figure A.7: (scenario 7) $\theta_0 = 0.9, m_e = 100, \rho = 0.7$, n=25

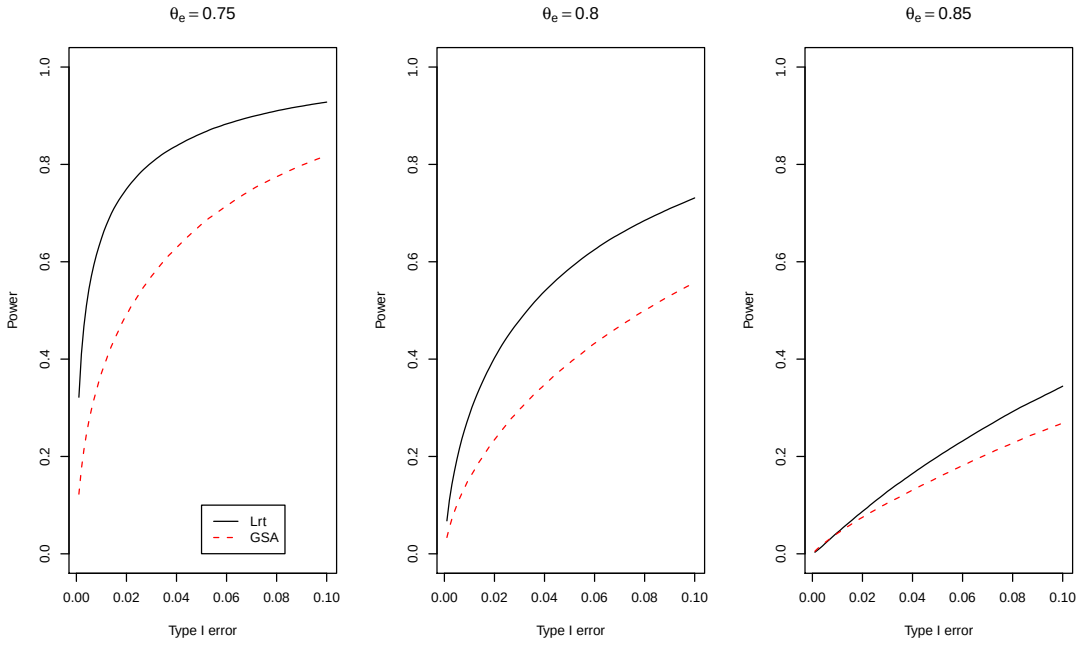


Figure A.8: (scenario 8) $\theta_0 = 0.9, m_e = 100, \rho = 0.7$, n=50

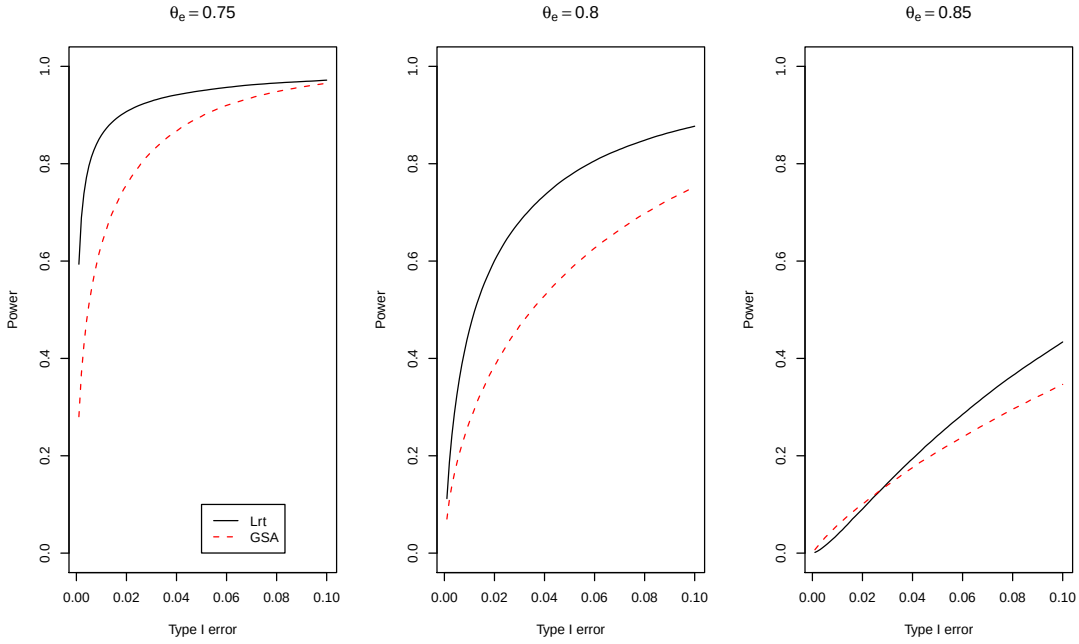


Figure A.9: (scenario 9) $\theta_0 = 0.95, m_e = 50, \rho = 0.2$, n=25

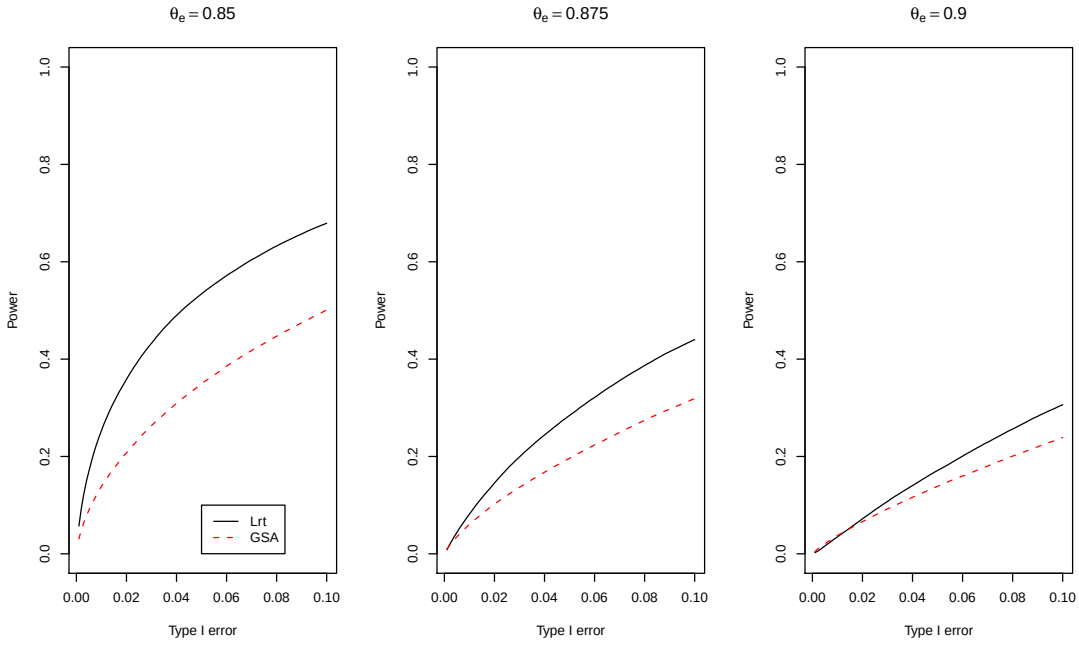


Figure A.10: (scenario 10)$\theta_0 = 0.95, m_e = 50, \rho = 0.2$, n=50

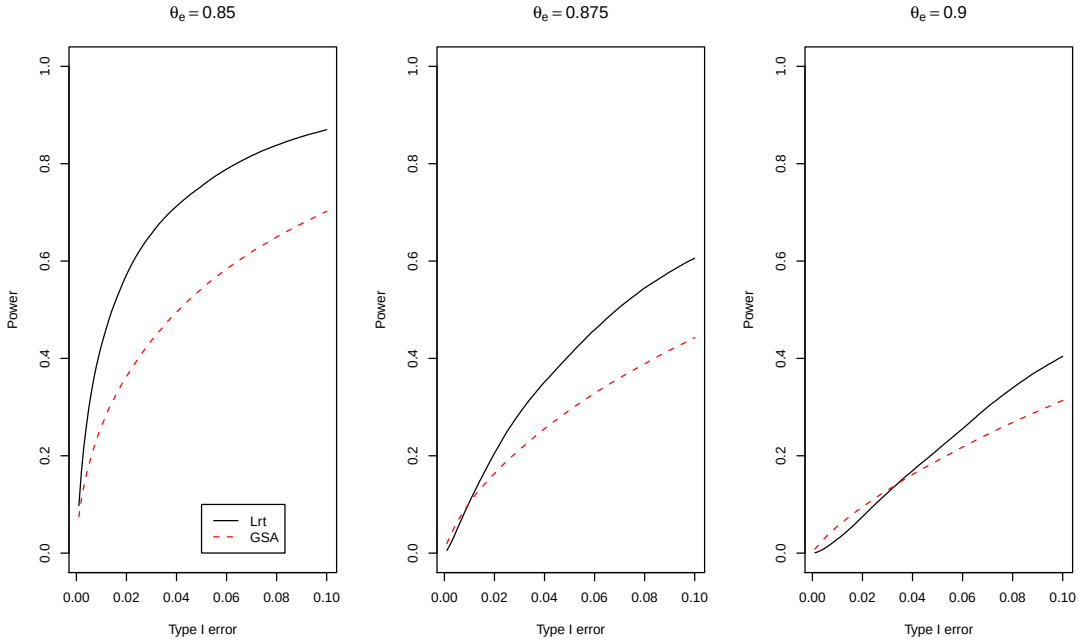


Figure A.11: (scenario 11) $\theta_0 = 0.95, m_e = 50, \rho = 0.7$, n=25

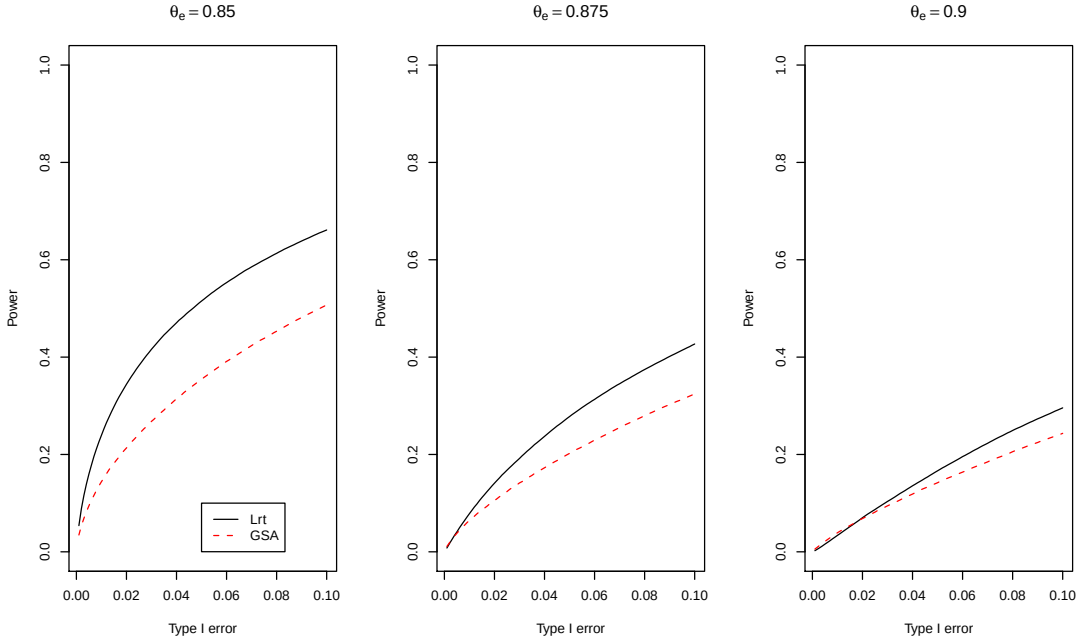


Figure A.12: (scenario 12) $\theta_0 = 0.95, m_e = 50, \rho = 0.7$, n=50

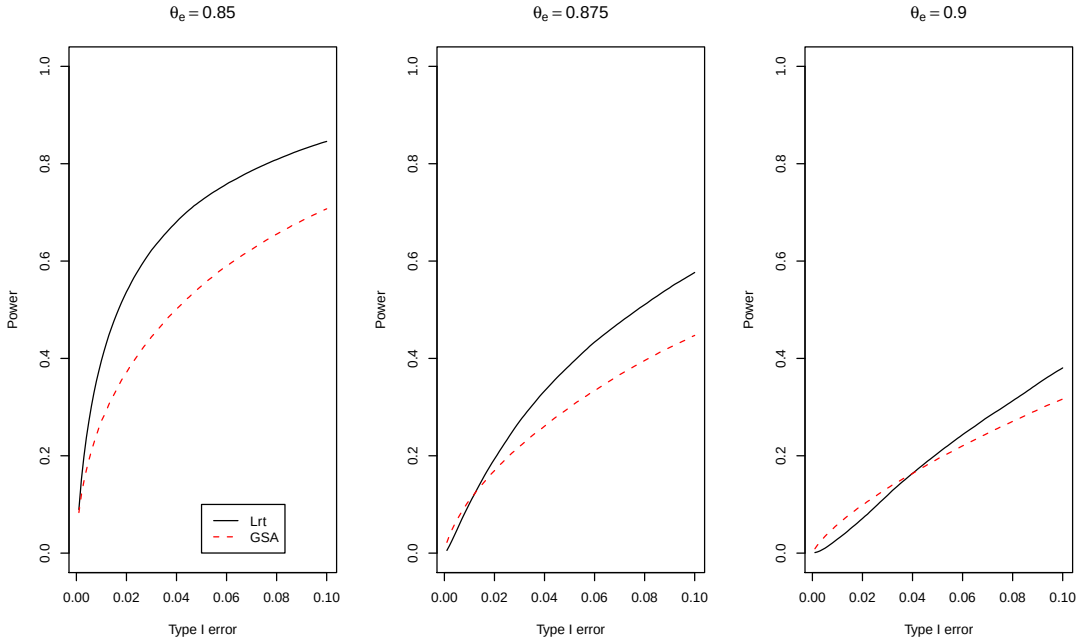


Figure A.13: (scenario 13) $\theta_0 = 0.95, m_e = 100, \rho = 0.2$, n=25

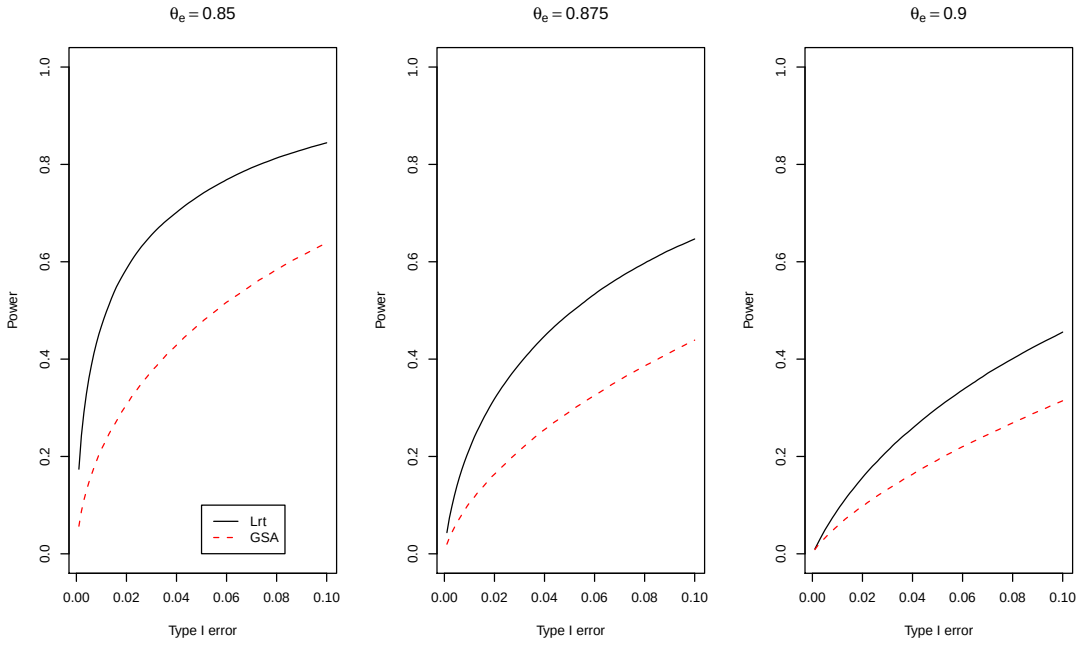


Figure A.14: (scenario 14) $\theta_0 = 0.95, m_e = 100, \rho = 0.2$, n=50

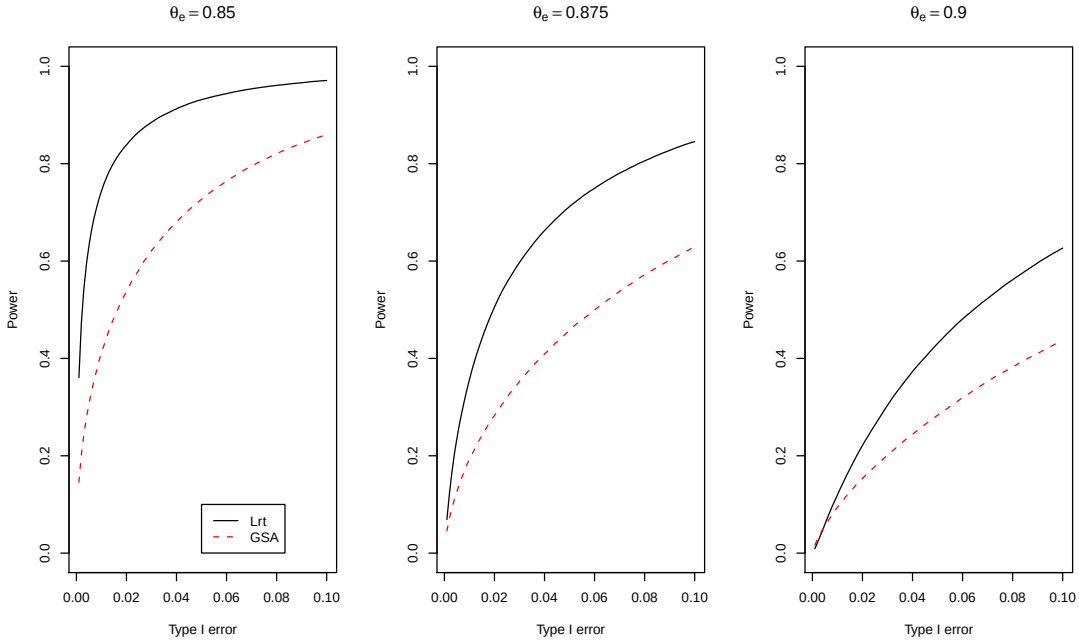


Figure A.15: (scenario 15) $\theta_0 = 0.95, m_e = 100, \rho = 0.7$, n=25

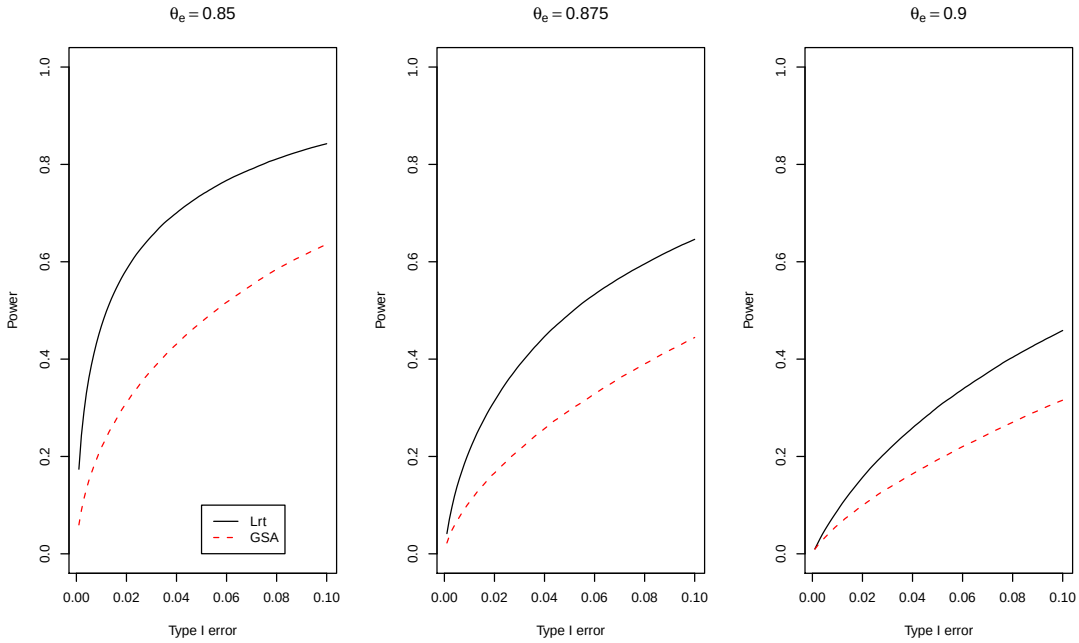
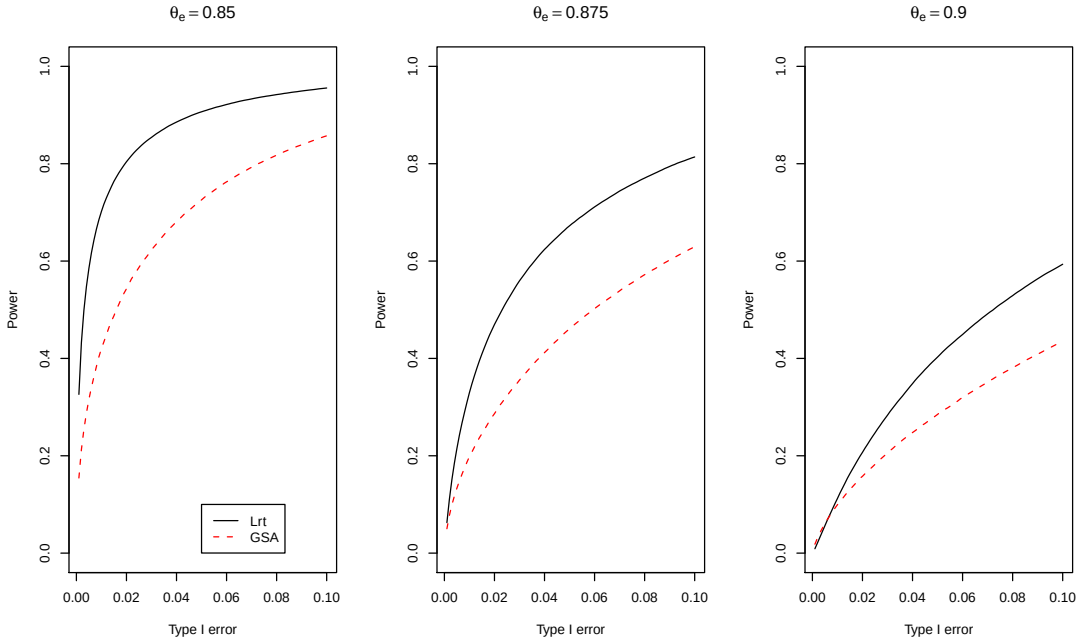


Figure A.16: (scenario 16) $\theta_0 = 0.95, m_e = 100, \rho = 0.7$, n=50



A.2 Onesided GSEA

EM algorithm for gene set model estimation

We discuss the model estimation for gene set A. For gene j in set A, define indicator $w_j \in \{0, 1\}$ following a binomial distribution

$$\Pr(w_j = 1) = p_{10}, \quad 0 \leq p_{10} \leq 1,$$

and conditionally we assume

$$z_j|(w_j = 1) \sim f_0(z_j), \quad z_j|(w_j = 0) \sim f_1(z_j).$$

The complete data likelihood function for gene set A given (f_0, f_1) is

$$\prod_{j \in A} \{p_{10} f_0(z_j)\}^{w_j} \{(1 - p_{10}) f_1(z_j)\}^{(1-w_j)}.$$

In E-step, the conditional expected log likelihood given the current estimate of the parameter $p_{10}^{(t)}$ is

$$\sum_{j \in A} \left\{ \tau_j^{(t)} \log p_{10}^{(t)} + (1 - \tau_j^{(t)}) \log(1 - p_{10}^{(t)}) \right\},$$

where $\tau_j^{(t)} = \frac{p_{10}^{(t)} f_0(z_j)}{p_{10}^{(t)} f_0(z_j) + (1 - p_{10}^{(t)}) f_1(z_j)}$. In M-step, the above conditional expected log likelihood is maximized by

$$p_{10}^{(t+1)} = \frac{1}{m_A} \sum_{j \in A} \tau_j^{(t)}.$$

By combining the E/M-steps, we can obtain

$$p_{10}^{(t+1)} = \frac{1}{m_A} \sum_{j \in A} \frac{p_{10}^{(t)} f_0(z_j)}{p_{10}^{(t)} f_0(z_j) + (1 - p_{10}^{(t)}) f_1(z_j)}.$$

Simulation study

In addition to the simulation presented in the chapter 3 based on the leukemia gene expression data, we also considered 16 different simulation settings based on the following

parameter setup: $\theta_0 = 0.9, 0.95$, $m_e = 100, 300$, $\rho = 0.2, 0.7$ and $n = 25, 50$ as listed in Table A.3.

Table A.4 summarizes the estimated Type I error over 100 simulations. Both Lrt and GSA methods have approximately the right size. Figure A.17 to A.32 show the estimated power versus type I error averaged over 100 simulations. Overall, the proposed Lrt performs better than GSA.

Table A.3: The 16 different simulation settings

θ_0	m_e	ρ	n	scenarios
0.9	100	0.2	25	scenario 1
			50	scenario 2
		0.7	25	scenario 3
			50	scenario 4
	300	0.2	25	scenario 5
			50	scenario 6
		0.7	25	scenario 7
			50	scenario 8
0.95	100	0.2	25	scenario 9
			50	scenario 10
		0.7	25	scenario 11
			50	scenario 12
	300	0.2	25	scenario 13
			50	scenario 14
		0.7	25	scenario 15
			50	scenario 16

Table A.4: For up-regulation, estimated type I error of Lrt and GSA over 100 simulations (listed within parenthesis are the standard errors).

	Lrt				GSA			
α	0.005	0.01	0.05	0.1	0.005	0.01	0.05	0.1
Scenario 1	0.0044	0.0091	0.0438	0.0864	0.0055	0.0108	0.0507	0.0997
	(0.0002)	(0.0003)	(0.0006)	(0.0009)	(0.0002)	(0.0003)	(0.0006)	(0.0010)
Scenario 2	0.0042	0.0084	0.0443	0.0876	0.0063	0.0122	0.0528	0.1020
	(0.0002)	(0.0003)	(0.0007)	(0.0011)	(0.0003)	(0.0003)	(0.0008)	(0.0010)
Scenario 3	0.0039	0.0085	0.0431	0.0859	0.0057	0.0113	0.0508	0.0989
	(0.0002)	(0.0003)	(0.0007)	(0.0010)	(0.0002)	(0.0003)	(0.0007)	(0.0009)
Scenario 4	0.0041	0.0084	0.0438	0.0874	0.0071	0.0129	0.0546	0.103
	(0.0002)	(0.0003)	(0.0007)	(0.0009)	(0.0002)	(0.0004)	(0.0007)	(0.0010)
Scenario 5	0.0043	0.0089	0.0444	0.0882	0.0053	0.0104	0.0494	0.1000
	(0.0002)	(0.0003)	(0.0007)	(0.0009)	(0.0002)	(0.0003)	(0.0007)	(0.0010)
Scenario 6	0.0042	0.0087	0.0454	0.0902	0.0059	0.0114	0.0507	0.0992
	(0.0002)	(0.0003)	(0.0006)	(0.0009)	(0.0002)	(0.0003)	(0.0007)	(0.0010)
Scenario 7	0.0047	0.0086	0.0429	0.0879	0.0057	0.0111	0.0507	0.1000
	(0.0002)	(0.0003)	(0.0006)	(0.0009)	(0.0003)	(0.0004)	(0.0007)	(0.0010)
Scenario 8	0.0043	0.0084	0.0442	0.0895	0.0071	0.0125	0.0534	0.1023
	(0.0002)	(0.0003)	(0.0007)	(0.0010)	(0.0003)	(0.0003)	(0.0007)	(0.0010)
Scenario 9	0.0040	0.0083	0.0407	0.0832	0.0059	0.0112	0.0517	0.1013
	(0.0002)	(0.0003)	(0.0006)	(0.0009)	(0.0003)	(0.0004)	(0.0006)	(0.0009)
Scenario 10	0.0040	0.0076	0.0398	0.0796	0.0071	0.0124	0.0534	0.1019
	(0.0002)	(0.0003)	(0.0007)	(0.0012)	(0.0003)	(0.0003)	(0.0007)	(0.0009)
Scenario 11	0.0040	0.0079	0.0412	0.0792	0.0064	0.0115	0.0519	0.0997
	(0.0002)	(0.0003)	(0.0008)	(0.0013)	(0.0002)	(0.0003)	(0.0008)	(0.0010)
Scenario 12	0.0040	0.0075	0.0386	0.0875	0.0076	0.0134	0.0547	0.1015
	(0.0002)	(0.0003)	(0.0008)	(0.0014)	(0.0003)	(0.0004)	(0.0008)	(0.0010)
Scenario 13	0.0039	0.0082	0.0425	0.0865	0.0051	0.0101	0.0500	0.0999
	(0.0002)	(0.0003)	(0.0006)	(0.0009)	(0.0002)	(0.0003)	(0.0006)	(0.0009)
Scenario 14	0.0040	0.0082	0.0428	0.0844	0.0064	0.0123	0.0524	0.1004
	(0.0002)	(0.0003)	(0.0007)	(0.0010)	(0.0003)	(0.0003)	(0.0007)	(0.0010)
Scenario 15	0.0042	0.0086	0.0438	0.0862	0.0053	0.0103	0.0500	0.0996
	(0.0002)	(0.0003)	(0.0007)	(0.0010)	(0.0002)	(0.0003)	(0.0007)	(0.0009)
Scenario 16	0.0043	0.0089	0.0439	0.0884	0.0072	0.0127	0.0535	0.1005
	(0.0002)	(0.0003)	(0.0008)	(0.0011)	(0.0003)	(0.0004)	(0.0007)	(0.0010)

Figure A.17: (scenario 1) $\theta_0 = 0.9, m_e = 100, \rho = 0.2$, n=25

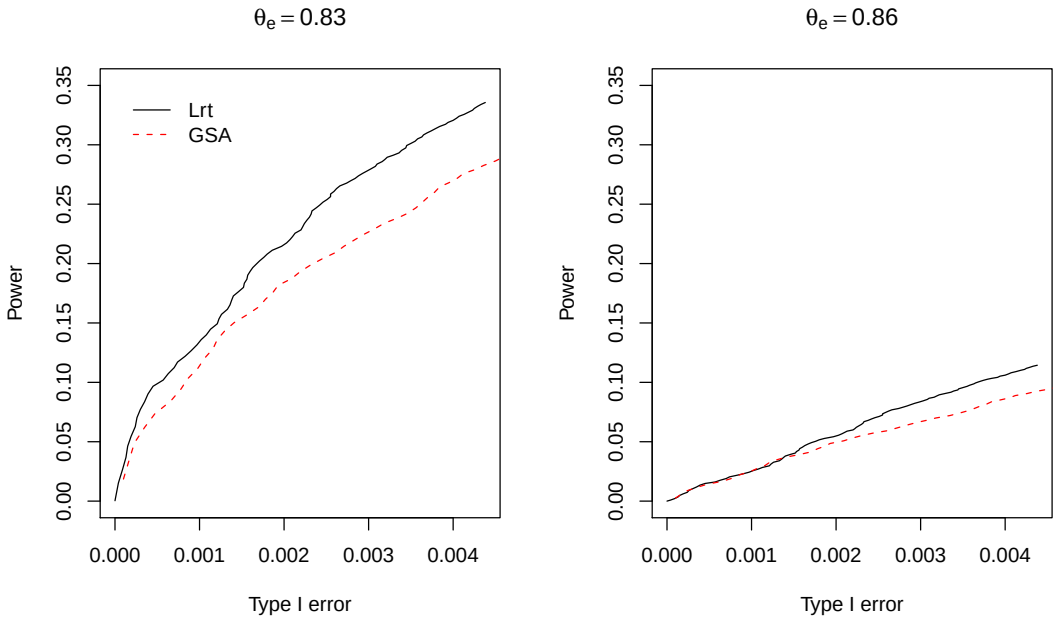


Figure A.18: (scenario 2) $\theta_0 = 0.9, m_e = 100, \rho = 0.2$, n=50

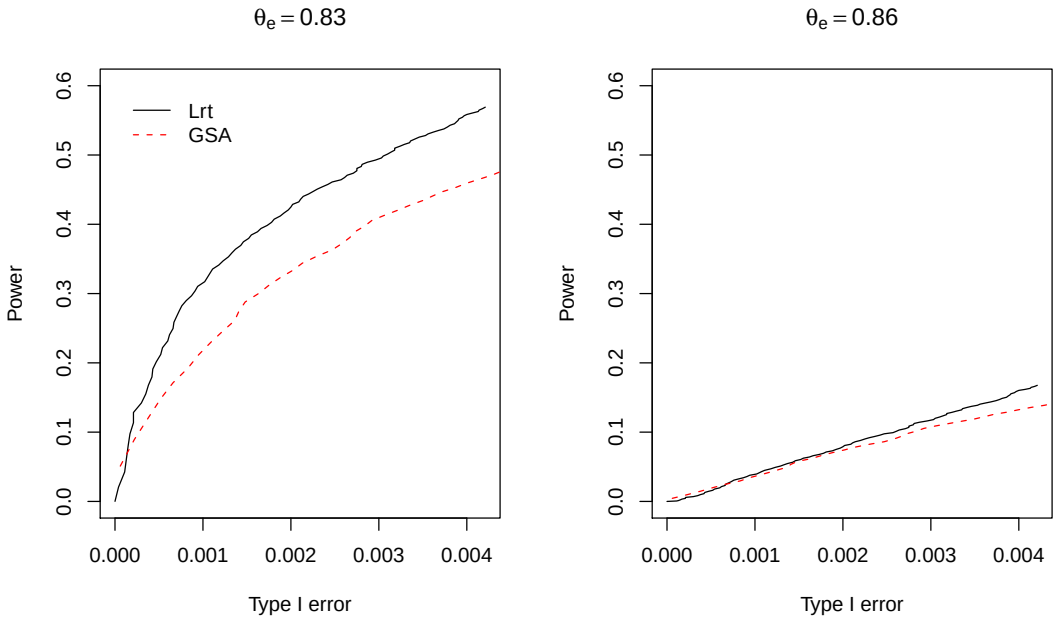


Figure A.19: (scenario 3) $\theta_0 = 0.9, m_e = 100, \rho = 0.7$, n=25

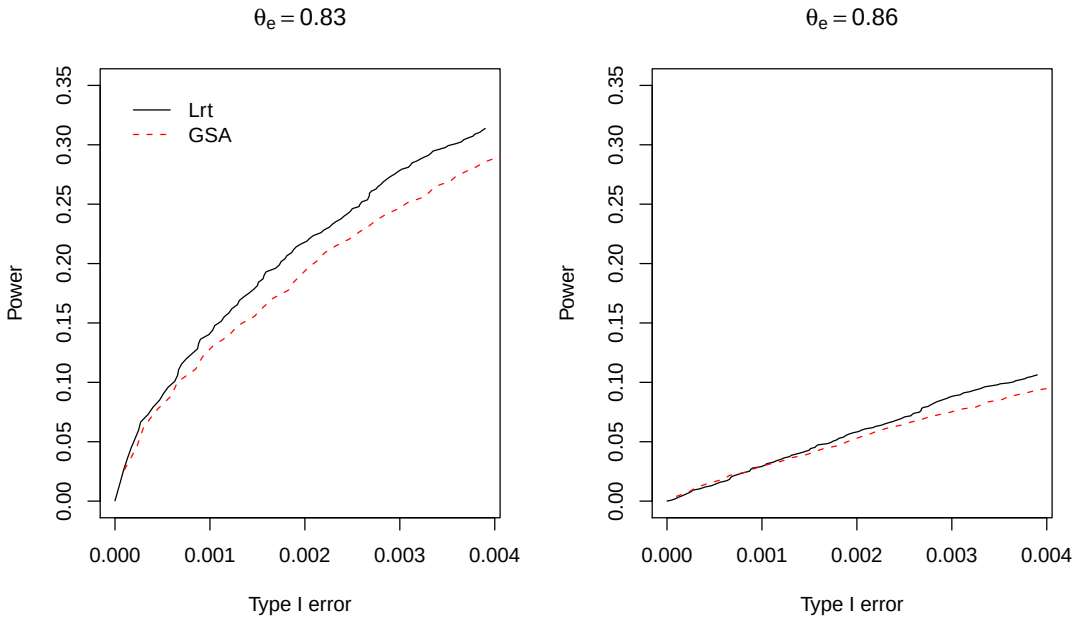


Figure A.20: (scenario 4) $\theta_0 = 0.9, m_e = 100, \rho = 0.7$, n=50

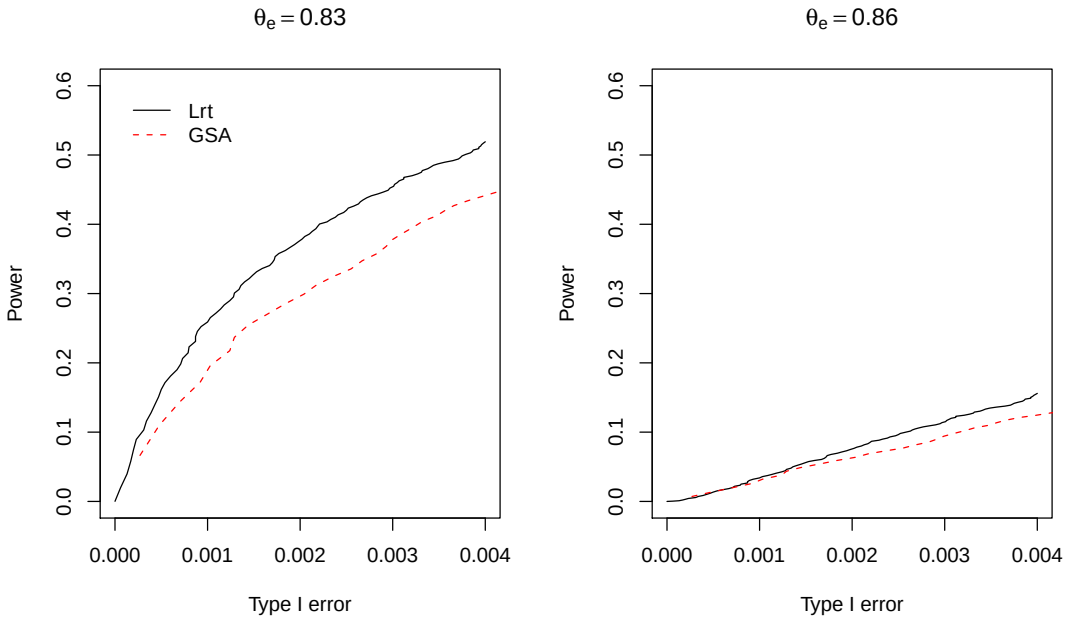


Figure A.21: (scenario 5) $\theta_0 = 0.9, m_e = 300, \rho = 0.2$, n=25

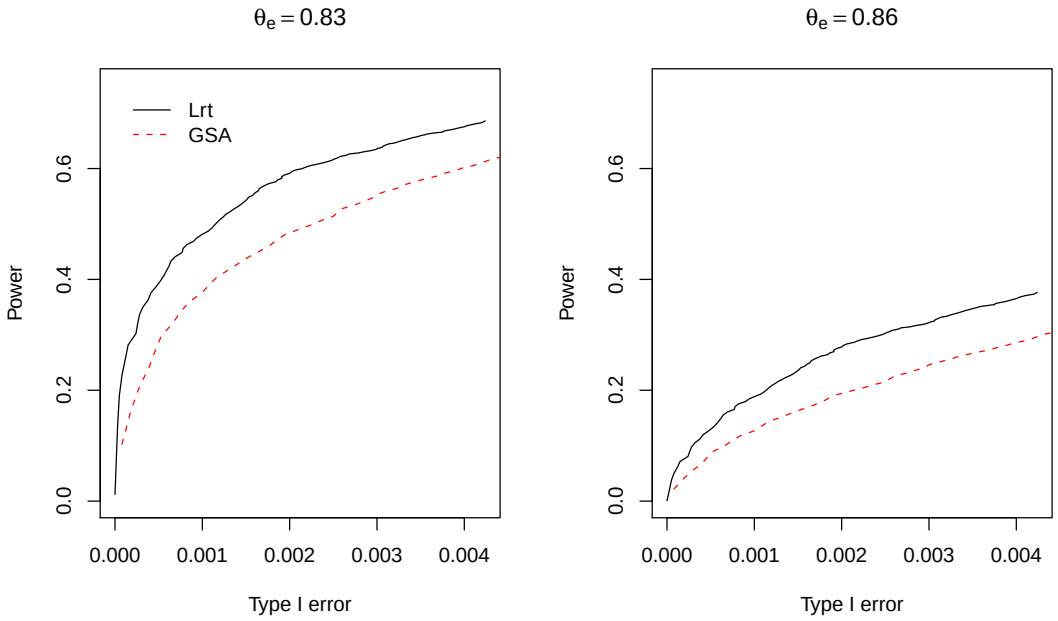


Figure A.22: (scenario 6) $\theta_0 = 0.9, m_e = 300, \rho = 0.2$, n=50

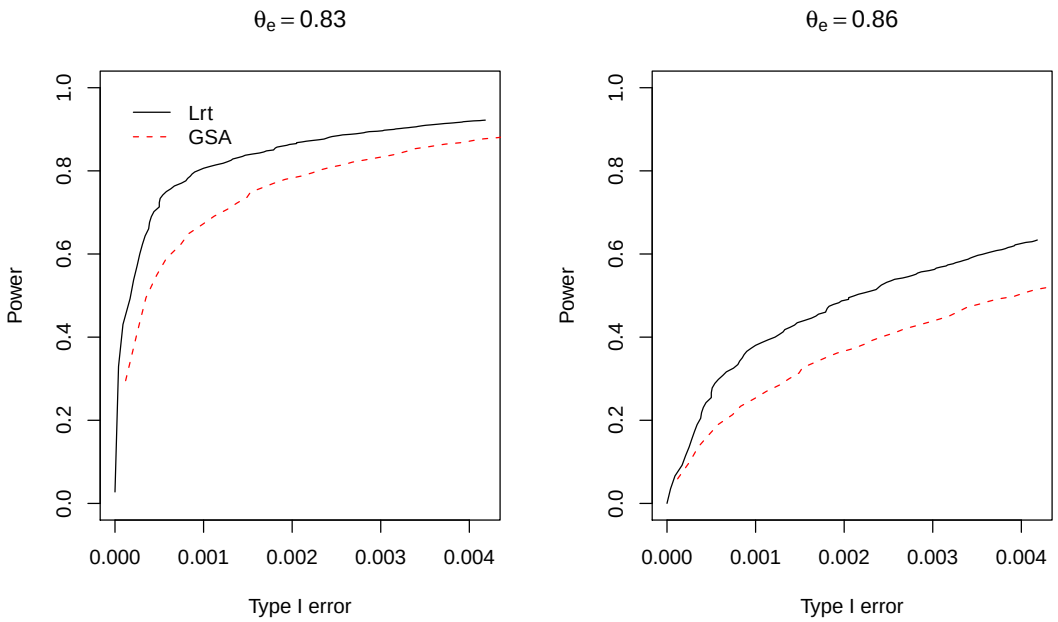


Figure A.23: (scenario 7) $\theta_0 = 0.9, m_e = 300, \rho = 0.7$, n=25

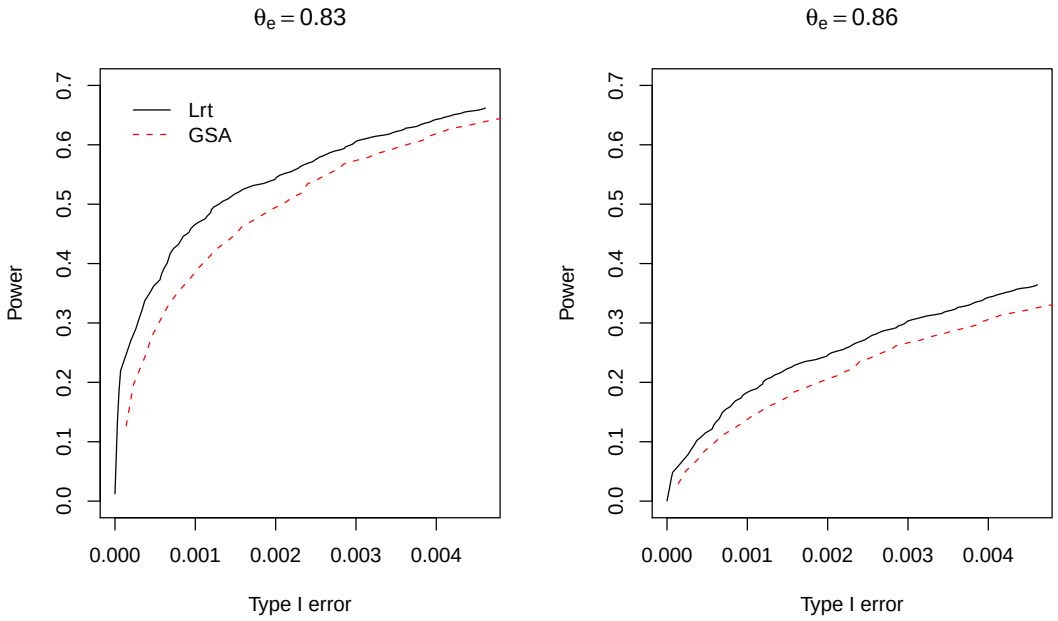


Figure A.24: (scenario 8) $\theta_0 = 0.9, m_e = 300, \rho = 0.7$, n=50

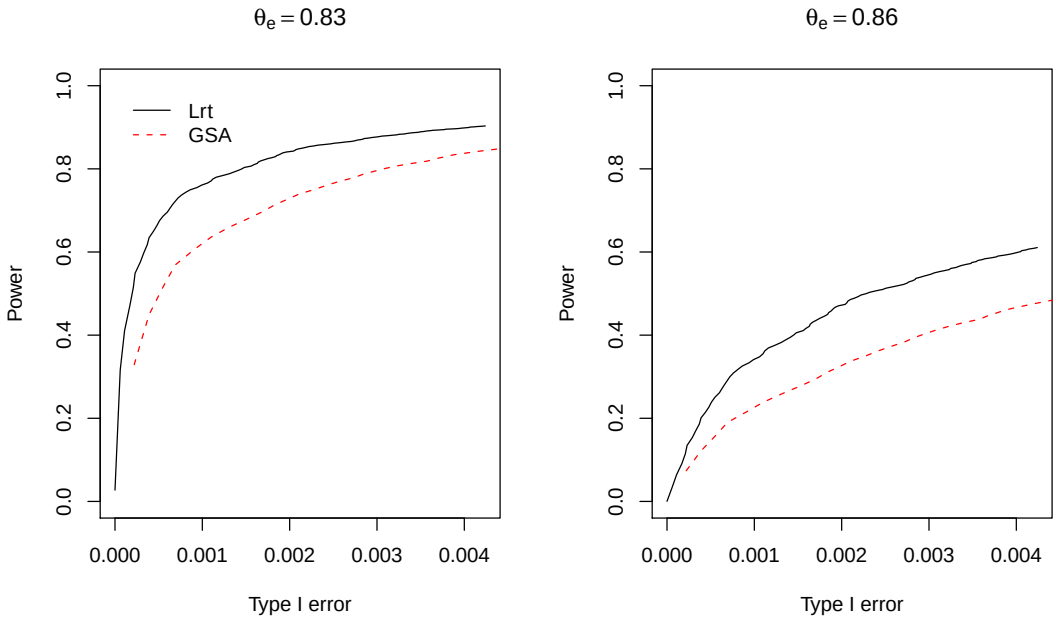


Figure A.25: (scenario 9) $\theta_0 = 0.95, m_e = 100, \rho = 0.2$, n=25

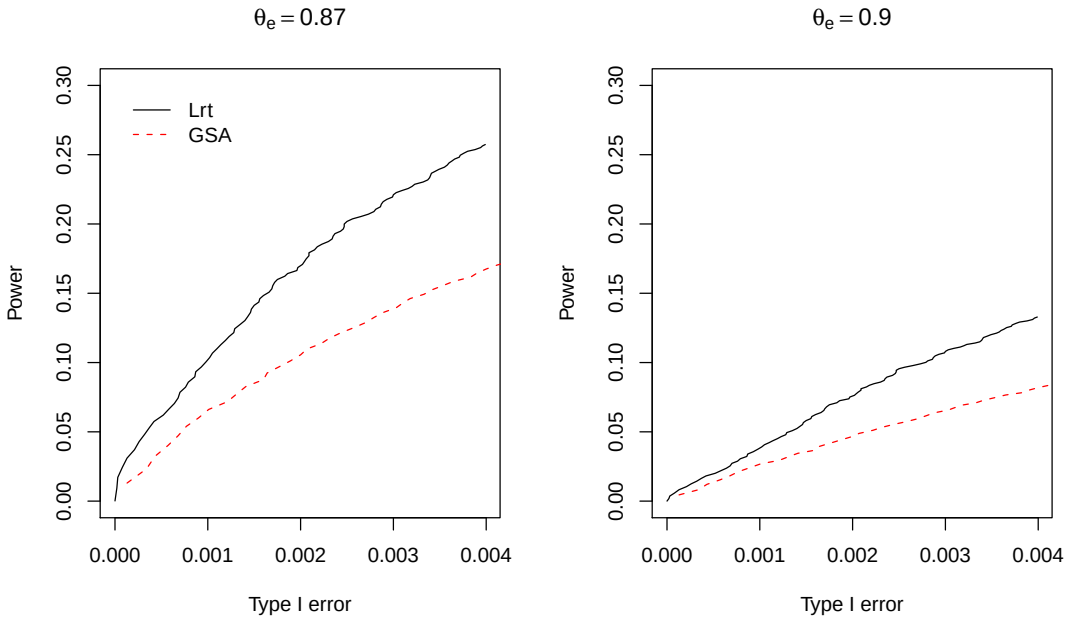


Figure A.26: (scenario 10) $\theta_0 = 0.95, m_e = 100, \rho = 0.2$, n=50

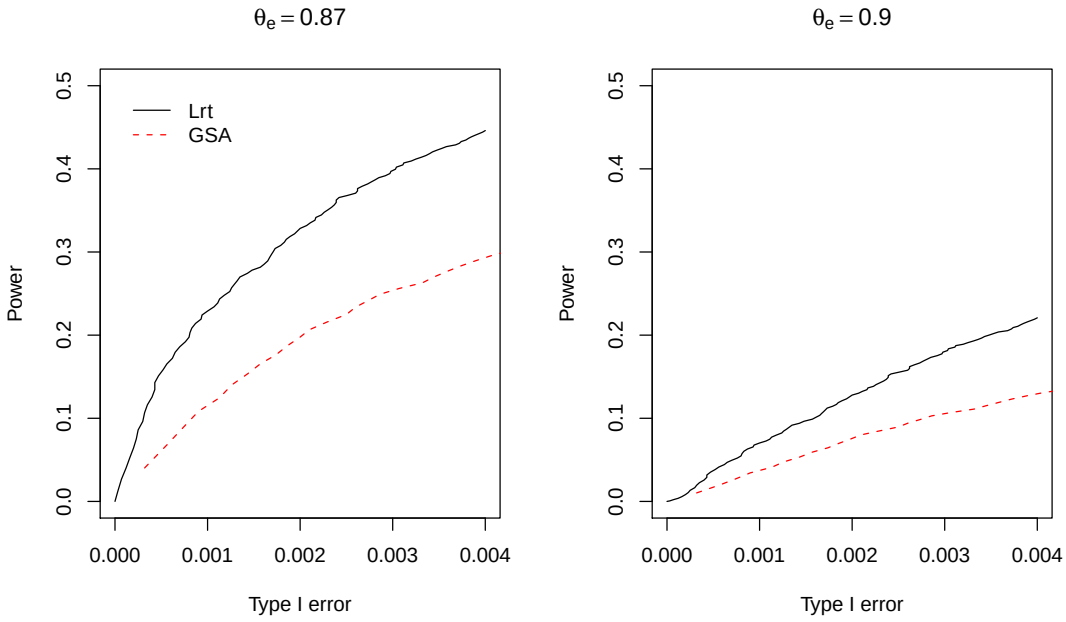


Figure A.27: (scenario 11) $\theta_0 = 0.95, m_e = 100, \rho = 0.7$, n=25

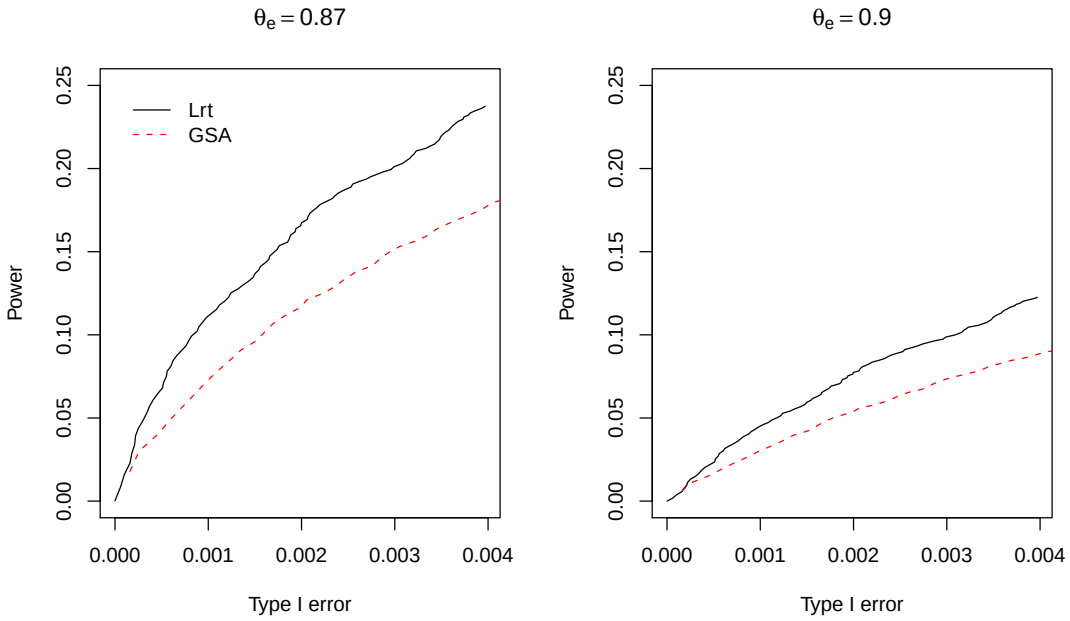


Figure A.28: (scenario 12) $\theta_0 = 0.95, m_e = 100, \rho = 0.7$, n=50

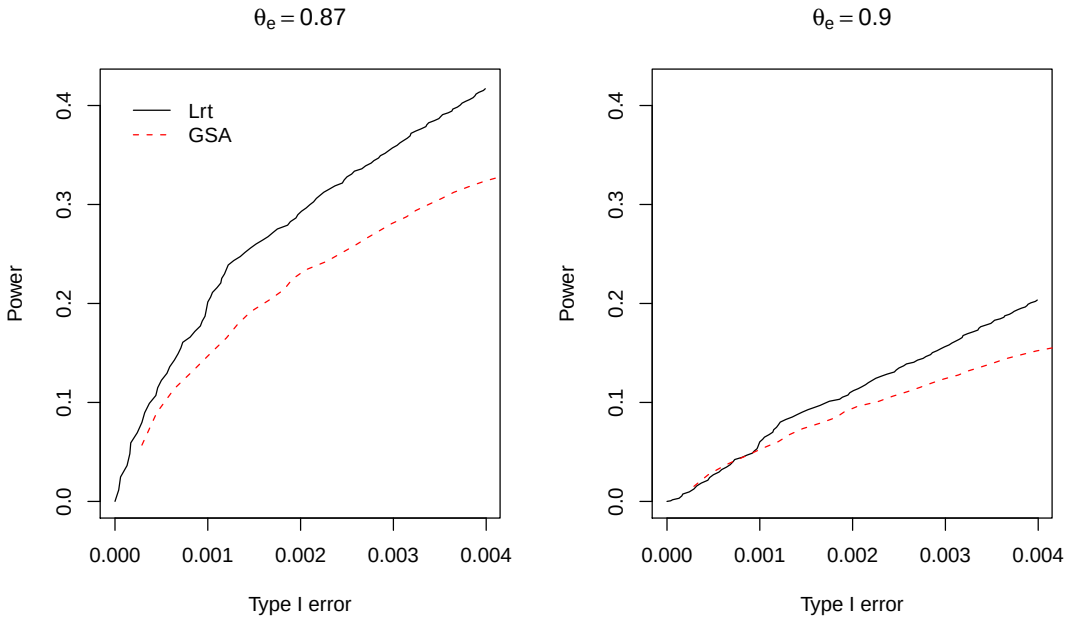


Figure A.29: (scenario 13) $\theta_0 = 0.95, m_e = 300, \rho = 0.2$, n=25

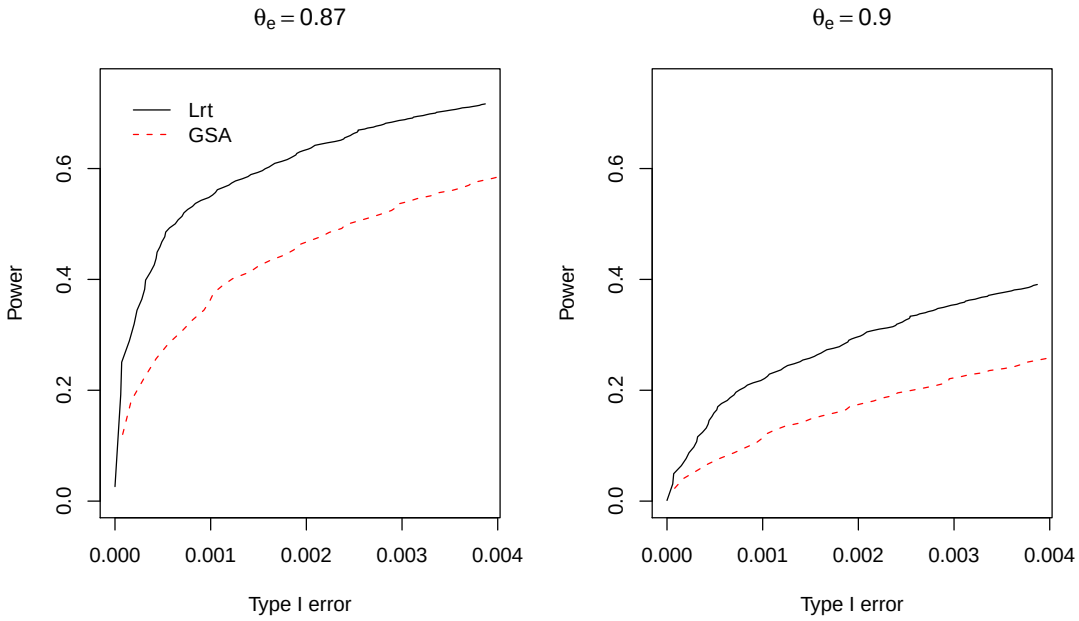


Figure A.30: (scenario 14) $\theta_0 = 0.95, m_e = 300, \rho = 0.2$, n=50

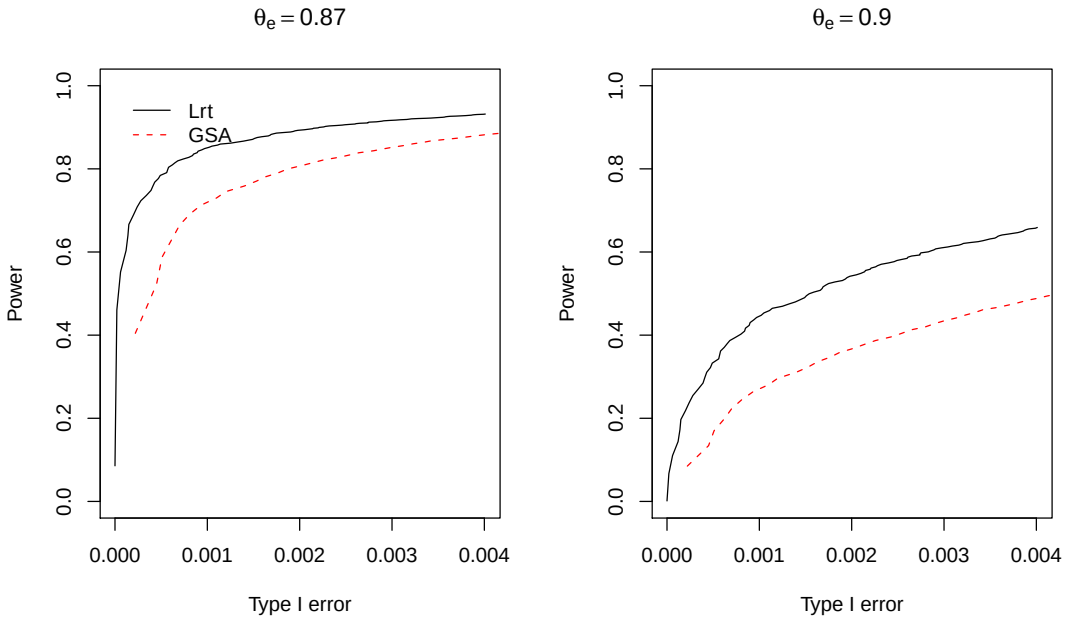


Figure A.31: (scenario 15) $\theta_0 = 0.95, m_e = 300, \rho = 0.7$, n=25

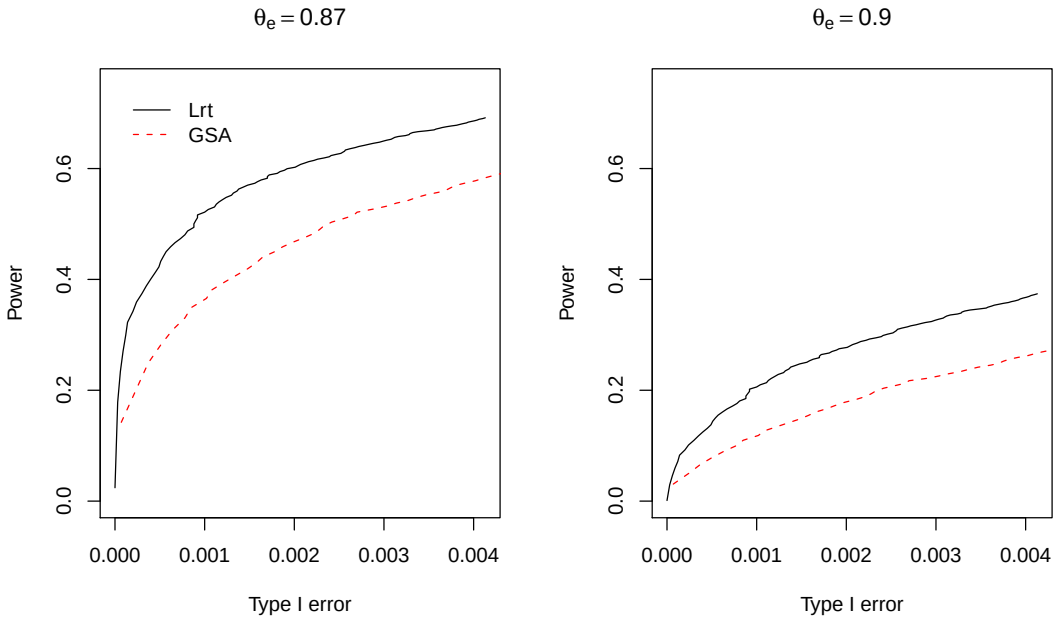
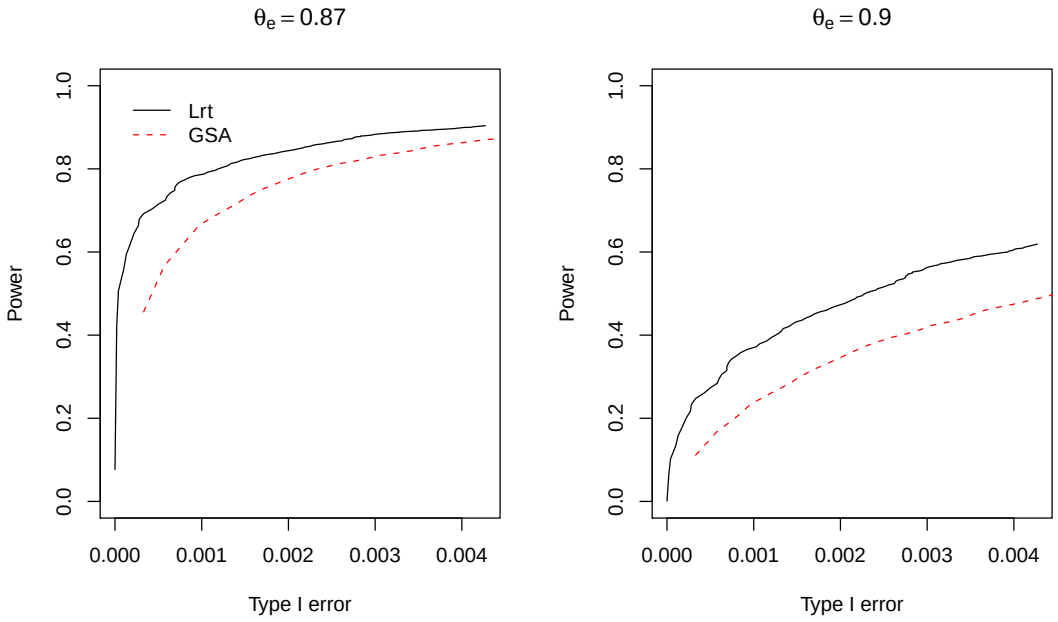


Figure A.32: (scenario 16) $\theta_0 = 0.95, m_e = 300, \rho = 0.7$, n=50



Leukemia data analysis based on Mle fitting null distribution estimation

The full list of 255 significant pathways enriched with down-regulated genes identified by Lrt is provided in Table A.5 to A.7.

Table A.5: The identified pathways for down-regulation by Lrt(FDR$\leq$ 0.05)

Pathway	# genes	p-value	Pathway	# genes	p-value
LEE_TCELLS3_UP	183	0	SERUM_FIBROBLAST_CORE_UP	345	1E-07
LEE_TCELLS2_UP	2008	0	ZHANG_EFT_EWSFLI1_UP	160	3E-07
SERUM_FIBROBLAST_CELLCYCLE	243	0	CHANG_SERUM_RESPONSE_UP	284	3E-07
LE_MYELIN_UP	198	7E-16	UVB_NHEK3_C5	92	4E-07
LI_FETAL_VS_WT_KIDNEY_DN	349	5E-15	GOLDRATH_CELLCYCLE	56	4E-07
STEMCELL_HEMATOPOIETIC_UP	2796	2E-14	HEDVAT_ELF_UP	18	5E-07
HSC_HSCANDPROGENITORS_FETAL	1112	3E-13	NING_COPD_UP	283	7E-07
HSC_HSCANDPROGENITORS_ADULT	1134	6E-13	PROLIFERATION_GENES	657	7E-07
POD1_KO_UP	786	1E-12	ET743PT650_COLONCA_DN	116	8E-07
STEMCELL_NEURAL_UP	3845	1E-12	TAKEDA_NUP8_HOXA9_3D_DN	52	1E-06
TAKEDA_NUP8_HOXA9_8D_DN	326	1E-12	CELL_PROLIFERATION	400	1E-06
MATSUDA_VALPHAINKT_DIFF	1058	3E-12	SANSOM_APC_LOSS4_UP	304	1E-06
HOFFMANN_BIVSBII_BI_TABLE2	437	4E-12	TAKEDA_NUP8_HOXA9_16D_DN	388	2E-06
KUMAR_HOXA_DIFF	789	5E-12	BRCA_PROGNOSIS_NEG	184	2E-06
STEMCELL_EMBRYONIC_UP	2773	8E-12	TSA_CD4_UP	53	2E-06
GAY_YY1_DN	632	8E-12	HOFFMANN_BIVSBII_BI	200	2E-06
CHIARETTI_T_ALL_DIFF	488	1E-11	CMV_HCMV_TIMECOURSE_8HRS_DN	35	2E-06
BRCA_ER_NEG	1671	5E-11	HOFFMANN_BIVSBII_IMVM	178	3E-06
HADDAD_HSC_CD7_UP	97	6E-11	ALCALAY_AML_NPMC_DN	331	3E-06
UVB_NHEK3_ALL	848	6E-11	HDACI_COLON_SUL2HRS_DN	33	4E-06
IDX_TSA_UP_CLUSTER3	208	1E-10	UVC_HIGH_ALL_DN	642	5E-06
TAKEDA_NUP8_HOXA9_10D_DN	207	1E-10	CIS_RESIST_GASTRIC_UP	33	5E-06
DOX_RESIST_GASTRIC_UP	79	2E-10	LU_IL4BCELL	110	5E-06
HADDAD_HPCLYMPHO_ENRICHED	501	9E-10	CORDERO_KRAS_KD_VS_CONTROL_DN	100	6E-06
MIDDLEAGE_DN	33	2E-09	UVC_TTD_ALL_DN	901	7E-06
HADDAD_HSC_CD10_UP	470	2E-09	CALRES_RHESUS_UP	120	7E-06
CROONQUIST_IL6_STARVE_UP	75	2E-09	TARTE_PLASMA_BLASTIC	616	9E-06
P21_P53_ANY_DN	90	2E-09	ERM_KO_SERTOLI_DN	39	1E-05
PRMT5_KD_UP	381	5E-09	GREENBAUM_E2A_UP	72	1E-05
CANCER_UNDIFFERENTIATED_META_UP	120	8E-09	VERHAAK_AML_NPM1_MUT_VS_WT_UP	285	1E-05
ZHAN_MM_CD138_PR_VS_REST	73	2E-08	SHEPARD_GENES_COMMON_BW_CB_MO	134	1E-05
STEMCELL_COMMON_UP	391	2E-08	UVB_NHEK1_DN	622	1E-05
IDX_TSA_DN_CLUSTER1	83	3E-08	HOGERKORP_CD44_UP	56	1E-05
HSC_EARLYPROGENITORS_SHARED	828	4E-08	BROWN_MYELOID_PROLIF_AND_SELF_RENEWAL	239	1E-05
OLDAGE_DN	103	4E-08	UVC_TTD_4HR_DN	762	1E-05
CROONQUIST_IL6_RAS_DN	47	4E-08	HOHENKIRK_MONOCYTE_DEND_UP	196	1E-05
CHIARETTI_T_ALL	449	5E-08	ALZHEIMERS_DISEASE_UP	2784	2E-05
HSC_EARLYPROGENITORS_ADULT	830	5E-08	MOOTHA_VOXPHOS	142	2E-05
YU_CMYC_UP	66	6E-08	UVC_XPCS_8HR_DN	1027	2E-05
BOQUEST_CD31PLUS_VS_CD31MINUS_UP	1227	6E-08	CELL_CYCLE_KEGG	172	2E-05
CARIES_PULP_UP	315	7E-08	BREAST_DUCTAL_CARCINOMA_GENES	43	2E-05
P21_P53_MIDDLE_DN	49	8E-08	NGUYEN_KERATO_DN	183	3E-05
VERHAAK_AML_NPM1_MUT_VS_WT_DN	435	1E-07	RADMACHER_AMLNORMALKARYTYPE_SIG	153	3E-05

Table A.6: The identified pathways for down-regulation by Lrt (FDR$\leq$ 0.05)(*Conti.*)

Pathway	# genes	p-value	Pathway	# genes	p-value
3AB_GAMMA_DN	24	3E-05	HOHENKIRK_MONOCYTE_DEND_DN	199	0.0005
ET743_HELA_UP	92	3E-05	HSA05214_GLIOMA	164	0.0005
UVB_NHEK3_C1	124	4E-05	BRENTANI_DNA_METHYLATION_MODIFICATION	48	0.0005
CROONQUIST_RAS_STROMA_DN	49	4E-05	LINDSTEDT_DEND_DN	127	0.0005
BRENTANI_CELL_CYCLE	177	4E-05	4NQO_UNIQUE_FIBRO_UP	49	0.0005
ROSS_PML_RAR	138	4E-05	VEGF_HUVEC_30MIN_UP	39	0.0005
SMOOTH_MUSCLE_CONTRACTION	318	4E-05	MYC_ONCOGENIC_SIGNATURE	302	0.0005
TAKEDA_NUP8_HOXA9_6H_UP	134	5E-05	VERNELL_PRB_CLSTR1	126	0.0005
HG_PROGERIA_DN	63	6E-05	TAKEDA_NUP8_HOXA9_3D_UP	296	0.0005
UVC_XPCS_ALL_DN	1215	6E-05	HADDAD_HSC_CD7_DN	184	0.0005
STRESS_TPA_SPECIFIC_UP	70	7E-05	BRUNO_IL3_DN	130	0.0005
ADIP_DIFF_CLUSTER5	69	7E-05	HSA00190_OXIDATIVE_PHOSPHORYLATION	209	0.0006
PARP_KO_UP	66	8E-05	TGFBETA_ALL_UP	165	0.0006
MRNA_SPLICING	100	8E-05	ALCALAY_AML_NPMC_UP	254	0.0006
RAY_P210_DIFF	94	9E-05	CMV_HCMV_TIMECOURSE_ALL_DN	812	0.0006
DAC_FIBRO_DN	20	0.0001	SANA_TNFA_ENDOTHELIAL_DN	143	0.0006
WANG_HOXA9_VS_MEIS1_DN	58	0.0001	UV-4NQO_FIBRO_UP	44	0.0008
ROSS_CBF_MYH	82	0.0001	MYOD_NIH3T3_DN	131	0.0008
BRCA1_OVEREXP_DN	212	0.0002	PGC1APATHWAY	70	0.0009
VHL_RCC_UP	174	0.0002	DCPATHWAY	27	0.0009
UVB_NHEK1_C6	310	0.0003	ET743_SARCOMA_6HRS_UP	50	0.0009
PENG_GLUTAMINE_DN	511	0.0003	HCC_SURVIVAL_GOOD_VS_POOR_DN	260	0.0009
ZHAN_MM_CD138_CD1_VS_REST	68	0.0003	TAKEDA_NUP8_HOXA9_16D_UP	244	0.0010
SHEPARD_CRASH_AND_BURN_MUT_VS_WT_DN	310	0.0003	IRITANI_ADPROX_LYMPH	243	0.0010
BASSO_HCL_DIFF	175	0.0003	HSA04640_HEMATOPOIETIC_CELL_LINEAGE	110	0.0010
ERYTHPATHWAY	20	0.0003	BASSO_REGULATORY_HUBS	259	0.0010
HSC_HSC_FETAL	568	0.0003	UVC_XPCS_4HR_DN	611	0.0010
UVC_HIGH_D4_DN	103	0.0003	RUTELLA_HEMATOGFSNDCS_DIFF	1160	0.0010
LINDSTEDT_DEND_8H_VS_48H_UP	110	0.0003	ST_WNT_CA2_CYCLIC_GMP_PATHWAY	44	0.0010
KAMMINGA_EZH2_TARGETS	69	0.0003	UVB_NHEK3_C7	101	0.0010
UVB_SCC_DN	231	0.0003	ET743_SARCOMA_DN	616	0.0011
YE_INTRAMETASTATIC_HCC_UP	41	0.0003	P21_P53_EARLY_DN	27	0.0011
BLYMPHOCYTEPATHWAY	16	0.0003	UVB_NHEK3_C6	61	0.0011
DNA_REPLICATION_REACTOME	84	0.0003	TPA_SENS_LATE_UP	97	0.0011
CELL_CYCLE	162	0.0004	CACAMPATHWAY	47	0.0012
LIZUKA_L1_GR_G1	48	0.0004	BRCA1_SW480_UP	58	0.0013
ET743_SARCOMA_72HRS_DN	498	0.0004	BASSO_GERMINAL_CENTER_CD40_DN	115	0.0013
LEI_MYB_REGULATED_GENES	568	0.0004	TH1TH2PATHWAY	20	0.0013
ROS_MOUSE_AORTA_DN	121	0.0004	GILDEA_BLADDER_UP	44	0.0014
IRITANI_ADPROX_VASC	277	0.0004	MYC_TARGETS	76	0.0014
ELECTRON_TRANSPORT_CHAIN	176	0.0004	MOREAUX_TACI_HI_IN_PPC_UP	111	0.0015
ZHAN_MM_MOLECULAR_CLASSI_DN	97	0.0004	ELONGINA_KO_UP	329	0.0015
ZUCCHI_EPITHELIAL_UP	87	0.0004	CHESLER_HIGHEST_FOLD_RANGE_GENES	98	0.0015

Table A.7: The identified pathways for down-regulation by Lrt (FDR$\leq$ 0.05)(*Conti.*)

Pathway	# genes	p-value	Pathway	# genes	p-value
HIPPOCAMPUS_DEVELOPMENT_PRENATAL	59	0.0015	NAKAJIMA_MCS_UP	143	0.0033
HSA00670_ONE_CARBON_POOL_BY_FOLATE	36	0.0015	TPA_SENS_MIDDLE_UP	121	0.0035
LEE_TCELLS4_UP	69	0.0016	AD12_48HRS_DN	23	0.0036
ET743_RESIST_UP	28	0.0016	ADIP_DIFF_CLUSTER2	66	0.0036
HSA04662_BCELL_RECEPTOR_SIGNALING_PATHWAY	155	0.0018	EXTRINSICPATHWAY	20	0.0037
IGLESIAS_E2FMINUS_UP	316	0.0018	STRIATED_MUSCLE_CONTRACTION	63	0.0040
TAKEDA_NUP8_HOXA9_10D_UP	278	0.0018	OXSTRESS_BREASTCA_DN	21	0.0040
HYPOXIA_RCC_UP	194	0.0018	UVC_HIGH_D7_DN	52	0.0040
SANSOM_APC_LOSS5_UP	166	0.0018	IGF1_NIH3T3_UP	70	0.0040
GLYCINE_SERINE_AND_THREONINE_METABOLISM	49	0.0018	TAKEDA_NUP8_HOXA9_8D_UP	242	0.0042
PENG_LEUCINE_DN	299	0.0018	GH_AUTOCRINE_UP	318	0.0043
HSA00271_METHIONINE_METABOLISM	32	0.0019	AD12_24HRS_DN	34	0.0043
HSA04740_OLFACTORY_TRANSDUCTION	65	0.0020	HDACI_COLON_SUL_DN	457	0.0043
GH_EXOGENOUS_ANY_UP	417	0.0020	JNK_UP	79	0.0044
ET743_HELA_DN	29	0.0020	HUMAN_MITODB_6_2002	709	0.0044
IFN_BETA_UP	142	0.0021	NUCLEOTIDE_METABOLISM	25	0.0045
SHEPARD_BMYB_MORPHOLINO_DN	321	0.0021	UVC_TTD_8HR_DN	408	0.0046
CREBPATHWAY	85	0.0021	EMT_DN	102	0.0050
NEMETH_TNF_UP	191	0.0021	HOGERKORP_ANTI_CD44_UP	57	0.0050
CHEN_HOXA5_TARGETS_DN	81	0.0022	DER_IFNB_UP	191	0.0052
UVC_HIGH_D3_DN	106	0.0023	GERY_CEBP_TARGETS	206	0.0053
CROONQUIST_IL6_STROMA_UP	80	0.0024	LEE_TCELLS5_UP	46	0.0053
PARP_KO_DN	37	0.0024	ESR_FIBROBLAST_UP	98	0.0056
BYSTRYKH_HSC_BRAIN_TRANS_GLOCUS	370	0.0024	UVC_HIGH_D1_DN	24	0.0057
ONE_CARBON_POOL_BY_FOLATE	33	0.0025	YAGI_AML_PROG_FAB	365	0.0058
WANG_MLL_CBP_VS_GMP_DN	111	0.0025	P21_ANY_DN	66	0.0059
TGFBETA_C3_UP	24	0.0026	ZELLER_MYC_UP	42	0.0060
AGEING_KIDNEY_SPECIFIC_UP	298	0.0027	CROMER_HYPOPHARYNGEAL_MET_NON_UP	147	0.0060
GAMMA-UV_FIBRO_UP	75	0.0027	ATMPATHWAY	44	0.0061
CMV_HCMV_TIMECOURSE_14HRS_DN	94	0.0027	MRNA_PROCESSING_REACTOME	224	0.0061
JISON_SICKLECELL_DIFF	610	0.0027	4NQO_ESR_WS_UNREG	65	0.0063
VENTRICLES_UP	384	0.0027	ET743_SARCOMA_24HRS_DN	268	0.0064
SASAKI_TCELL_LYMPHOMA_VS_CD4_UP	317	0.0028	GH_GHRHR_KO_6HRS_UP	179	0.0065
ZHAN_MM_CD138_CD2_VS_REST	54	0.0029	P21_EARLY_DN	20	0.0066
CARIES_PULP_HIGH_UP	117	0.0029	HTERT_UP	131	0.0068
AGEING_BRAIN_UP	449	0.0029	BCRABL_HL60_CDNA_DN	43	0.0070
TGFBETA_LATE_UP	77	0.0030	POMEROY_DESMOPLASIC_CLASSIC_MD_UP	92	0.0070
CMV_IE86_UP	106	0.0030	VEGF_HUVEC_UP	21	0.0071
UVC_LOW_C2_DN	46	0.0030	HSC_HSC_SHARED	538	0.0071
BRG1_ALAB_DN	71	0.0031	MAPKKK_CASCADE	17	0.0071
AGED_MOUSE_CEREBELLUM_UP	105	0.0031	WELCSH_BRCA_DN	27	0.0075
FALT_BCLL_DN	101	0.0032			

Leukemia data analysis based on Central Matching null distribution estimation

Efron (2007) proposed two methods for estimating $(\theta_0, \mu_0, \sigma_0)$ in the empirical null distribution. One is "Central Matching", which approximates the marginal log density with a quadratic curve near zero. "Mle Fitting", is another method based on a truncated normal model by assuming non-null distribution has zero support in a pre-chosen small interval around zero. Central matching yields almost unbiased estimates if θ_0 exceeds 0.9, but it has large variation for estimating μ_0. Mle Fitting generally gives more stable estimates while it depends on the pre-chosen interval. Both methods have been implemented in the R package, *locfdr*.

In chapter 3, we analyze the leukemia data based on Central Matching null distribution (CM). Here we report the enrichment analysis results based on the Mle Fitting method (MF). The estimated values of $(\theta_0, \mu_0, \sigma_0)$ are (0.99,0.18,1.12) by MF and (0.97,0.19,1.08) by CM. Given FDR$<$ 0.05, we detected 27 significant pathways enriched with up- and 157 enriched with down-regulated genes using MF. All identified pathways are overlapped with those detected using CM. Figure A.33 shows the estimated FDR for analyzing enrichment of up- and down-regulation. Table A.8 to A.10 list the selected significant pathways.

Figure A.33: FDR of 30 significant pathways for up- and down-regulation

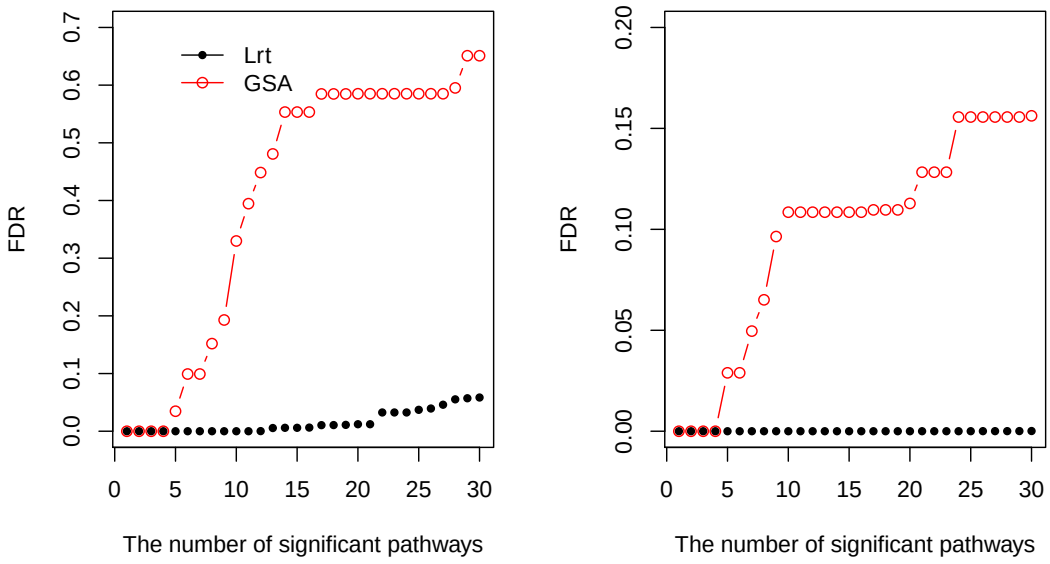


Table A.8: The identified pathways for up-regulation by Lrt and GSA (FDR$\leq$ 0.05)

Lrt

Pathway	# genes	p-value
VERHAAK_AML_NPM1_MUT_VS_WT_UP	285	3E-15
HOX_GENES	83	2E-14
ALCALAY_AML_BY_NPM1_LOCALIZATION_UP	254	1E-12
KUMAR_HOXA_DIFF	789	5E-11
ROSS_MLL_FUSION	129	1E-10
TAKEDA_NUP8_HOXA9_16D_UP	244	2E-10
TAKEDA_NUP8_HOXA9_10D_UP	278	4E-10
WANG_MLL_CBP_VS_GMP_UP	116	6E-10
TAKEDA_NUP8_HOXA9_3D_UP	296	8E-10
HEMATOPOESIS_RELATED_TRANSCRIPTION_FACTORS	170	1E-09
TAKEDA_NUP8_HOXA9_8D_UP	242	3E-09
CHIARETTI_T_ALL_DIFF	488	1E-06
EPHA4PATHWAY	25	4E-05
CHIARETTI_T_ALL	449	5E-05
MONOCYTEPATHWAY	27	5E-05
LAIRPATHWAY	28	6E-05
RIBAVIRIN_RSV_DN	90	0.0001
PASSERINI_SIGNAL	645	0.0001
TAKEDA_NUP8_HOXA9_6H_UP	134	0.0001
HSA04514_CELL_ADHESION_MOLECULES	249	0.0001
GH_AUTOCRINE_DN	235	0.0001
PARK_MSCS_DIFF	83	0.0004
JISON_SICKLECELL_DIFF	610	0.0004
JISON_SICKLE_CELL	58	0.0005
LEE_DENA_DN	112	0.0005
CMV-UV_HCMV_6HRS_DN	190	0.0006
GOLUB_ALL_VS_AML_DN	30	0.0007

GSA

Pathway	# genes	p-value	rank by Lrt
TAKEDA_NUP8_HOXA9_16D_UP	244	0	6
ROSS_MLL_FUSION	129	0	5
HEMATOPOESIS_RELATED_TRANSCRIPTION_FACTORS	170	0	10
WANG_MLL_CBP_VS_GMP_UP	116	0	8
HOX_GENES	83	0.0001	2

Table A.9: The identified pathways for down-regulation by Lrt(FDR$\leq$ 0.05)

Pathway	# genes	p-value	Pathway	# genes	p-value
TAKEDA_NUP8_HOXA9_8D_DN	326	4E-12	CELL_PROLIFERATION	400	1E-05
HSC_HSCANDPROGENITORS_FETAL	1112	1E-11	SHEPARD_CELL_PROLIFERATION	400	1E-05
HSC_HSCANDPROGENITORS_SHARED	1112	1E-11	ET743_HELA_UP	92	1E-05
LEE_TCELLS2_UP	2008	1E-11	CALRES_RHESUS_UP	120	1E-05
HSC_HSCANDPROGENITORS_ADULT	1134	2E-11	ET743PT650_COLONCA_DN	116	2E-05
KUMAR_HOXA_DIFF	789	6E-11	ERM_KO_SERTOLI_DN	39	2E-05
LE_MYELIN_UP	198	1E-10	MIDDLEAGE_DN	33	2E-05
CHIARETTI_T_ALL_DIFF	488	6E-10	TSA_CD4_UP	53	2E-05
TAKEDA_NUP8_HOXA9_10D_DN	207	7E-10	ZHAN_MM_CD138_CD1_VS_REST	68	3E-05
UVB_NHEK3_ALL	848	1E-09	UVC_HIGH_ALL_DN	642	3E-05
SERUM_FIBROBLAST_CELLCYCLE	243	1E-09	DOX_RESIST_GASTRIC_UP	79	3E-05
STEMCELL_HEMATOPOIETIC_UP	2796	2E-09	TAKEDA_NUP8_HOXA9_6H_UP	134	4E-05
HADDAD_HSC_CD7_UP	97	3E-09	HSC_EARLYPROGENITORS_SHARED	828	4E-05
HADDAD_CD45CD7_PLUS_VS_MINUS_UP	97	3E-09	HSC_EARLYPROGENITORS_FETAL	828	4E-05
LI_FETAL_VS_WT_KIDNEY_DN	349	6E-09	UVB_NHEK1_C6	310	4E-05
IDX_TSA_DN_CLUSTER1	83	8E-09	HSC_EARLYPROGENITORS_ADULT	830	4E-05
VERHAAK_AML_NPM1_MUT_VS_WT_DN	435	1E-08	HEDVAT_ELF_UP	18	5E-05
POD1_KO_UP	786	3E-08	CROONQUIST_IL6_STARVE_UP	75	5E-05
UVB_NHEK3_C5	92	4E-08	HOFFMANN_BIVSBII_IMVM	178	5E-05
MATSUDA_VALPHAINKT_DIFF	1058	6E-08	IDX_TSA_UP_CLUSTER3	208	5E-05
HADDAD_HPCLYMPHO_ENRICHED	501	1E-07	UVB_SCC_DN	231	6E-05
CARIES_PULP_UP	315	1E-07	HADDAD_HSC_CD7_DN	184	7E-05
GAY_YY1_DN	632	4E-07	HADDAD_CD45CD7_PLUS_VS_MINUS_DN	184	7E-05
HADDAD_HSC_CD10_UP	470	2E-07	TAKEDA_NUP8_HOXA9_16D_DN	388	7E-05
LEE_TCELLS3_UP	183	4E-07	BRENTANI_DNA_METHYLATION_MODIFICATION	48	8E-05
CHIARETTI_T_ALL	449	7E-07	ROSS_PML_RAR	138	8E-05
BRCA_ER_NEG	1671	7E-07	GOLDRATH_CELLCYCLE	56	8E-05
TAKEDA_NUP8_HOXA9_3D_DN	52	9E-07	TAKEDA_NUP8_HOXA9_16D_UP	244	9E-05
LU_IL4BCELL	110	1E-06	CIS_RESIST_GASTRIC_UP	33	0.0001
HOFFMANN_BIVSBII_BI	200	2E-06	YU_CMYC_UP	66	0.0001
PROLIFERATION_GENES	657	2E-06	ALCALAY_AML_NPMC_DN	331	0.0001
CORDERO_KRAS_KD_VS_CONTROL_DN	100	3E-06	HOHENKIRK_MONOCYTE_DEND_DN	199	0.0001
STEMCELL_COMMON_UP	391	4E-06	UVB_NHEK3_C1	124	0.0001
RADMACHER_AMLNORMALKARYTYPE_SIG	153	5E-06	UVB_NHEK1_DN	622	0.0001
ZHANG_EFT_EWSFLI1_UP	160	6E-06	PRMT5_KD_UP	381	0.0001
3AB_GAMMA_DN	24	7E-06	HOGERKORP_CD44_UP	56	0.0002
CMV_HCMV_TIMECOURSE_8HRS_DN	35	9E-06	ZHAN_MM_MOLECULAR_CLASSI_DN	97	0.0002
BOQUEST_CD31PLUS_VS_CD31MINUS_UP	1227	9E-06	ROSS_CBF_MYH	82	0.0002
HDACI_COLON_SUL2HRS_DN	33	1E-05	HOFFMANN_BIVSBII_BI_TABLE2	437	0.0002

Table A.10: The identified pathways for down-regulation by Lrt (FDR$\leq$ 0.05)(*Conti.*)

Pathway	# genes	p-value	Pathway	# genes	p-value
UVC_HIGH_D4_DN	103	0.0002	CHANG_SERUM_RESPONSE_UP	284	0.0015
PARP_KO_UP	66	0.0003	ZHAN_MM_CD138_PR_VS_REST	73	0.0016
UVC_LOW_C2_DN	46	0.0003	GREENBAUM_E2A_UP	72	0.0018
NING_COPD_UP	283	0.0003	P21_P53_MIDDLE_DN	49	0.0018
VERHAAK_AML_NPM1_MUT_VS_WT_UP	285	0.0003	BRCA_BRCA1_NEG	298	0.0018
TAKEDA_NUP8_HOXA9_10D_UP	278	0.0003	ET743_RESIST_UP	28	0.0018
LEE_TCELLS4_UP	69	0.0004	VHL_RCC_UP	174	0.0019
BASSO_HCL_DIFF	175	0.0004	BRCA_PROGNOSIS_NEG	184	0.0020
ERYTHPATHWAY	20	0.0004	ST_WNT_CA2_CYCLIC_GMP_PATHWAY	44	0.0021
RAY_P210_DIFF	94	0.0004	UVB_NHEK3_C7	101	0.0022
HSA05214_GLIOMA	164	0.0004	OLDAGE_DN	103	0.0023
CMV_HCMV_TIMECOURSE_ALL_DN	812	0.0004	BASSO_GERMINAL_CENTER_CD40_DN	115	0.0024
AGEING_KIDNEY_SPECIFIC_UP	298	0.0004	ZUCCHI_EPITHELIAL_UP	87	0.0026
WANG_HOXA9_VS_MEIS1_DN	58	0.0004	STRESS_GENOTOXIC_SPECIFIC_UP	48	0.0029
BLYMPHOCYTEPATHWAY	16	0.0005	STEMCELL_NEURAL_UP	3845	0.0029
GH_EXOGENOUS_ANY_UP	417	0.0005	UVC_LOW_ALL_DN	122	0.0029
HOHENKIRK_MONOCYTE_DEND_UP	196	0.0006	LEE_TCELLS5_UP	46	0.0030
STRESS_TPA_SPECIFIC_UP	70	0.0006	HSC_HSC_FETAL	568	0.0032
TAKEDA_NUP8_HOXA9_3D_UP	296	0.0006	CACAMPATHWAY	47	0.0032
IRITANI_ADPROX_VASC	277	0.0007	GH_AUTOCRINE_UP	318	0.0032
UVB_NHEK3_C6	61	0.0007	DAC_PANC_UP	555	0.0032
TPA_SENS_LATE_UP	97	0.0007	NGUYEN_KERATO_DN	183	0.0032
CARIES_PULP_HIGH_UP	117	0.0008	ZHAN_MM_CD138_CD2_VS_REST	54	0.0034
HSA04640_HEMATOPOIETIC_CELL_LINEAGE	110	0.0008	HEARTFAILURE_ATRIA_DN	226	0.0034
UV-4NQO_FIBRO_UP	44	0.0008	CMV_HCMV_TIMECOURSE_16HRS_UP	100	0.0035
DCPATHWAY	27	0.0009	MYOD_NIH3T3_DN	131	0.0035
STEMCELL_EMBRYONIC_UP	2773	0.0001	MAPKKK_CASCADE	17	0.0035
PGC1APATHWAY	70	0.0010	VENTRICLES_UP	384	0.0036
ALCALAY_AML_NPMC_UP	254	0.0010	VEGF_HUVEC_30MIN_UP	39	0.0036
SANA_TNFA_ENDOTHELIAL_DN	143	0.0011	LE_MYELIN_DN	222	0.0036
CMV_HCMV_TIMECOURSE_14HRS_DN	94	0.0011	CHESLER_HIGHEST_FOLD_RANGE_GENES	98	0.0038
P21_P53_ANY_DN	90	0.0012	EXTRINSICPATHWAY	20	0.0039
SHEPARD_GENES_COMMON_BW_CB_MO	134	0.0012	ALZHEIMERS_DISEASE_UP	2784	0.0040
SERUM_FIBROBLAST_CORE_UP	345	0.0012	ZHAN_MM_MOLECULAR_CLASSI_UP	116	0.0041
BRUNO_IL3_DN	130	0.0012	GH_EXOGENOUS_MIDDLE_UP	163	0.0042
CROONQUIST_RAS_STROMA_DN	49	0.0013	HYPOXIA_RCC_UP	194	0.0043
SMOOTH_MUSCLE_CONTRACTION	318	0.0014	AGED_MOUSE_CEREBELLUM_UP	105	0.0044
LINDSTEDT_DEND_DN	127	0.0014	CELL_DEATH	21	0.0045
CANCER_UNDIFFERENTIATED_META_UP	120	0.0015	DAC_FIBRO_DN	20	0.0045
SANSOM_APC_LOSS4_UP	304	0.0015			

A.3 GSEA for multi-class microarray data

EM algorithm for gene set model estimation

For gene i in set A_j, define indicator $y_i \in \{0, 1\}$ following a binomial distribution, $\Pr(y_i = 1) = \gamma_{j0}$. Conditionally we assume

$$z_i|(y_i = 1) \sim g_0(z_i), \quad z_i|(y_i = 0) \sim g_1(z_i)$$

The complete likelihood function for gene set A_j is

$$\prod_{i \in A_j} \Big\{\gamma_{j0} g_0(z_i)\Big\}^{y_i} \Big\{(1 - \gamma_{j0}) g_1(z_i)\Big\}^{1-y_i}.$$

The conditional expected log likelihood is proportional to

$$\sum_{i \in A_j} \{\tau_i^{(t)} \log \gamma_{j0}^{(t)} + (1 - \tau_i^{(t)}) \log(1 - \gamma_{j0}^{(t)})\}, \quad \tau_i^{(t)} = \frac{\gamma_{j0}^{(t)} g_0(z_i)}{\gamma_{j0}^{(t)} g_0(z_i) + (1 - \gamma_{j0}^{(t)}) g_1(z_i)}.$$

It is easy to check that

$$\gamma_{j0}^{(t+1)} = \frac{1}{m_j} \sum_{i \in A_j} \tau_i^{(t)},$$

where m_j is the number of genes in set A_j. Combining previous equations, we have a simple recursion rule for estimation

$$\gamma_{j0}^{(t+1)} = \frac{1}{m_j} \sum_{i \in A_j} \frac{\gamma_{j0}^{(t)} g_0(z_i)}{\gamma_{j0}^{(t)} g_0(z_i) + (1 - \gamma_{j0}^{(t)}) g_1(z_i)}, \quad j = 1, 2.$$

Simulation study

In chapter 4 we consider 8 different simulation scenarios listed in Table A.11. For the proposed method, the numerical optimization approach is denoted as Lrt-mle while the moment matching approach is denoted as Lrt-wt.

Table A.12 presents estimated Type I errors averaged over 100 simulations. The Lrt-wt and Lrt-mle yield almost identical results. In general, the proposed Lrt approach

Table A.11: The 8 different simulation settings; the true proportion of null genes θ_0, gene set size m_e and sample size n

θ_0	m_e	n	scenarios
0.9	50	15	scenario 1
		25	scenario 2
	100	15	scenario 3
		25	scenario 4
0.95	50	15	scenario 5
		25	scenario 6
	100	15	scenario 7
		25	scenario 8

is slightly conservative while GSA is slightly optimistic especially for small significance values.

For power comparison, we randomly sample 2000 gene sets each with m_e genes. Assume 10% gene sets are truly enriched with each containing $m_e\theta_e$ genes randomly sampled from the null set and $m_e(1-\theta_e)$ genes from non-null set. We compute true positives and FDR based on estimated Type I error and power.

Figures A.34 to A.41 show the FDR and true positives averaged over 100 simulations. Overall the proposed Lrt-wt and Lrt-mle have comparable results. And both have smaller FDR and can detect more truly enriched gene sets than GSA.

Table A.12: The estimated type I error over 100 simulations (listed within parenthesis are the standard errors)

	Lrt-wt				Lrt-mle				GSA			
α	0.005	0.01	0.05	0.1	0.005	0.01	0.05	0.1	0.005	0.01	0.05	0.1
Scenario 1	0.0044	0.0087	0.0445	0.0898	0.0043	0.084	0.0434	0.0898	0.0090	0.0155	0.0560	0.1028
	(0.0002)	(0.0003)	(0.0007)	(0.0009)	(0.0002)	(0.0003)	(0.0007)	(0.0010)	(0.0003)	(0.0004)	(0.0007)	(0.0009)
Scenario 2	0.0045	0.0091	0.0458	0.0908	0.0043	0.0085	0.0450	0.0912	0.0098	0.0161	0.0583	0.1041
	(0.0002)	(0.0003)	(0.0007)	(0.0009)	(0.0002)	(0.0003)	(0.0006)	(0.0009)	(0.0003)	(0.0004)	(0.0007)	(0.0009)
Scenario 3	0.0044	0.0090	0.0460	0.0918	0.0044	0.0092	0.0466	0.0928	0.0085	0.0148	0.0565	0.1040
	(0.0002)	(0.0003)	(0.0007)	(0.0010)	(0.0002)	(0.0003)	(0.0006)	(0.0009)	(0.0003)	(0.0004)	(0.0008)	(0.0010)
Scenario 4	0.0048	0.0091	0.0470	0.0945	0.0046	0.0092	0.0463	0.0939	0.0090	0.0155	0.0565	0.1032
	(0.0002)	(0.0003)	(0.0007)	(0.0009)	(0.0002)	(0.0003)	(0.0007)	(0.0009)	(0.0003)	(0.0004)	(0.0007)	(0.0009)
Scenario 5	0.0044	0.0085	0.0432	0.0870	0.0044	0.0083	0.0429	0.0856	0.0103	0.0174	0.0576	0.1020
	(0.0002)	(0.0003)	(0.0006)	(0.001)	(0.0002)	(0.0003)	(0.0007)	(0.0011)	(0.0003)	(0.0004)	(0.0007)	(0.0009)
Scenario 6	0.0045	0.0086	0.0433	0.0887	0.0043	0.0087	0.0431	0.0872	0.0130	0.0199	0.0616	0.1038
	(0.0002)	(0.0003)	(0.0007)	(0.0009)	(0.0002)	(0.0003)	(0.0007)	(0.0009)	(0.0004)	(0.0004)	(0.0007)	(0.0009)
Scenario 7	0.0046	0.0089	0.0451	0.0909	0.0043	0.0086	0.0442	0.0898	0.0093	0.0159	0.0575	0.1036
	(0.0002)	(0.0003)	(0.0008)	(0.0010)	(0.0002)	(0.0003)	(0.0007)	(0.0008)	(0.0003)	(0.0004)	(0.0009)	(0.0011)
Scenario 8	0.0049	0.0092	0.0457	0.0906	0.0041	0.0089	0.0450	0.0912	0.0111	0.0180	0.0600	0.1060
	(0.0002)	(0.0003)	(0.0007)	(0.0011)	(0.0002)	(0.0003)	(0.0008)	(0.0010)	(0.0003)	(0.0004)	(0.0007)	(0.0009)

Figure A.34: (scenario 1) $\theta_0 = 0.9, m_e = 50, n = 15$

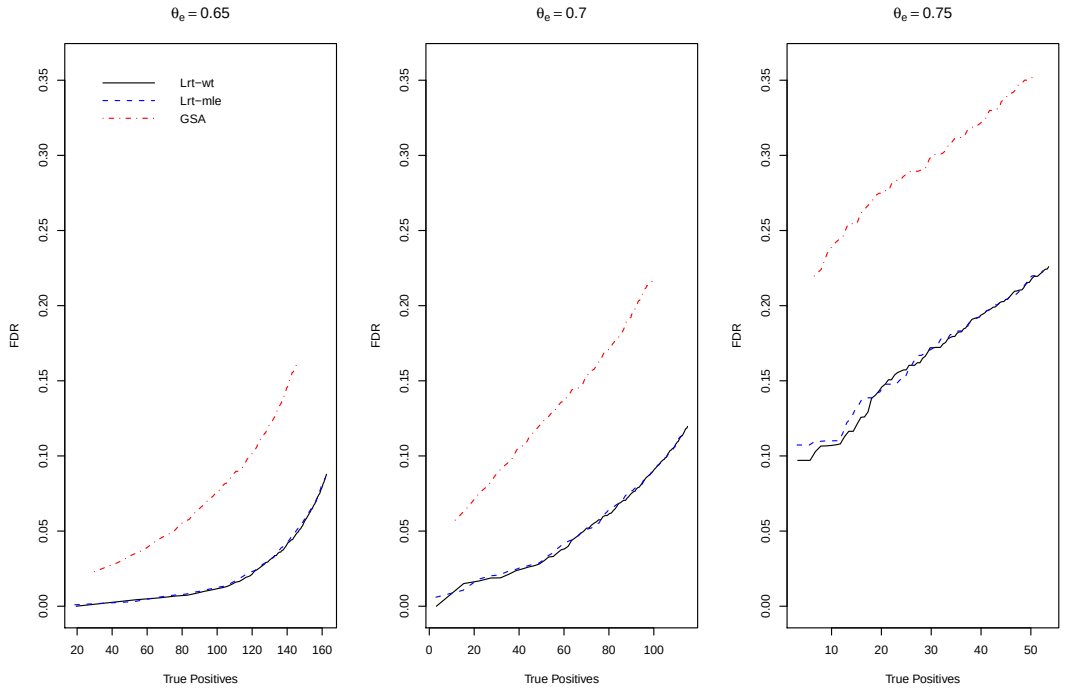


Figure A.35: (scenario 2) $\theta_0 = 0.9, m_e = 50, n = 25$

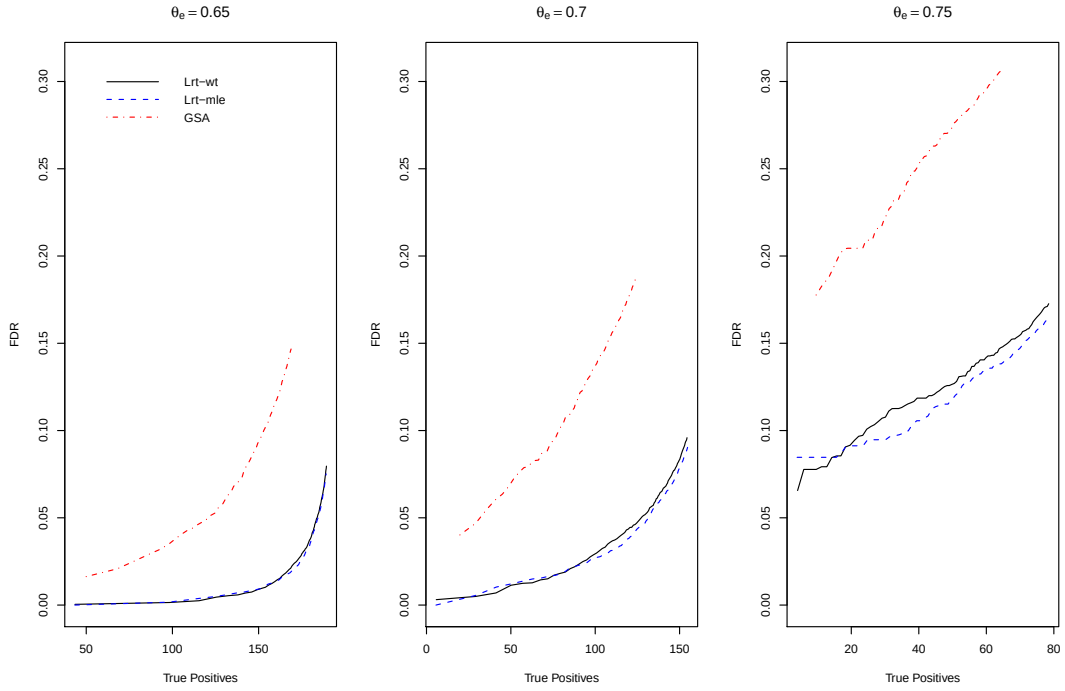


Figure A.36: (scenario 3) $\theta_0 = 0.9, m_e = 100, n = 15$

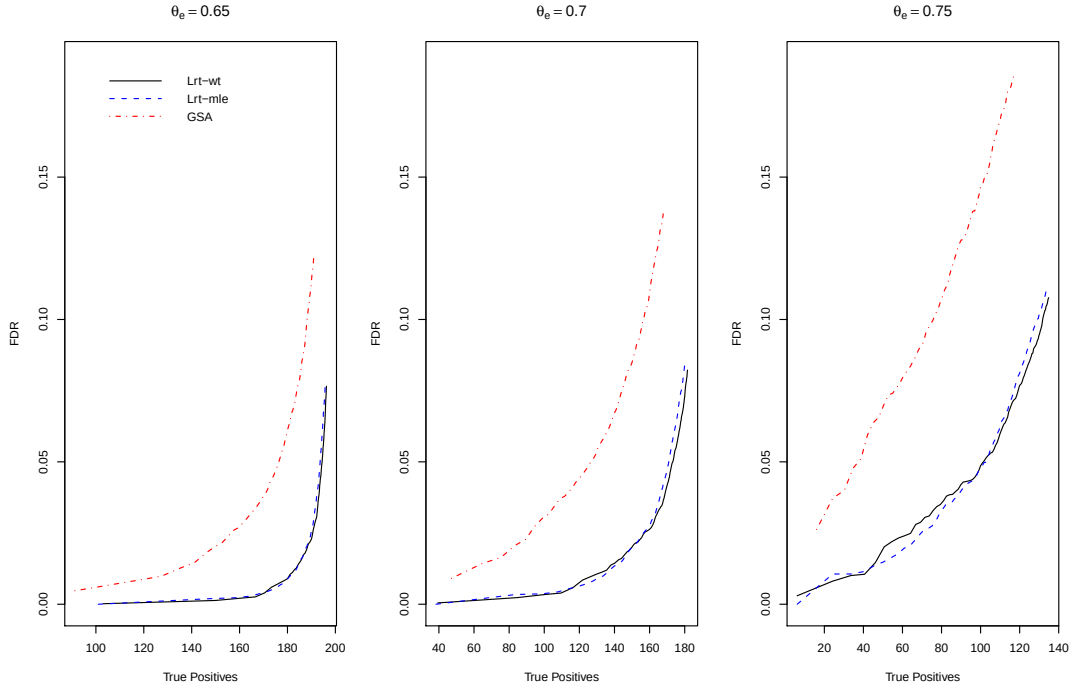


Figure A.37: (scenario 4) $\theta_0 = 0.9, m_e = 100, n = 25$

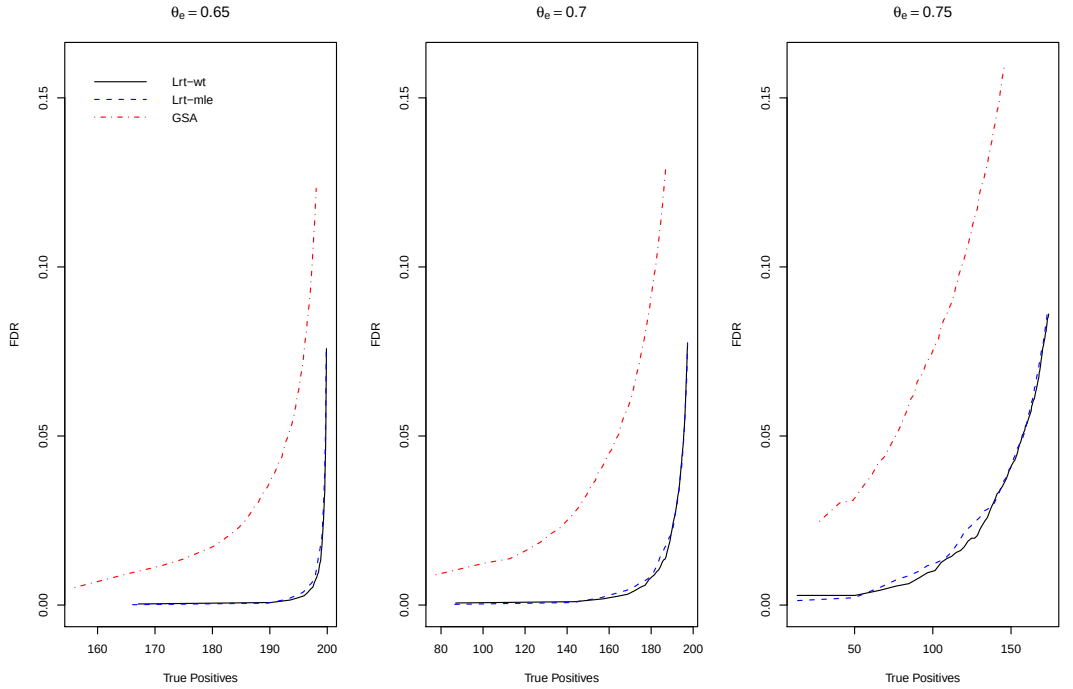


Figure A.38: (scenario 5) $\theta_0 = 0.95, m_e = 50, n = 15$

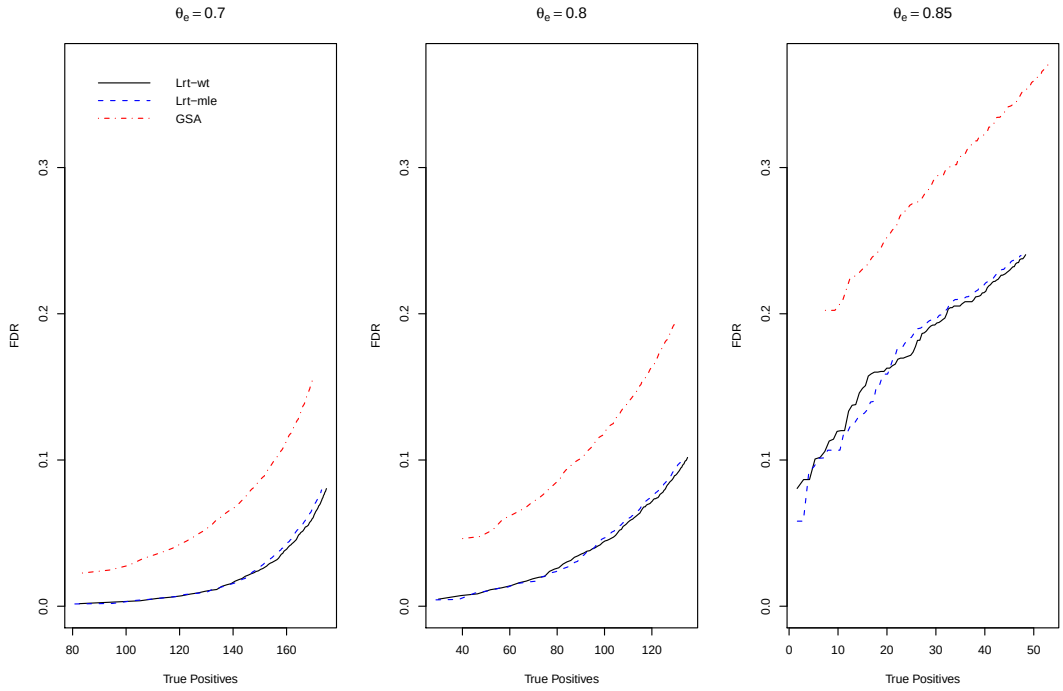


Figure A.39: (scenario 6) $\theta_0 = 0.95, m_e = 50, n = 25$

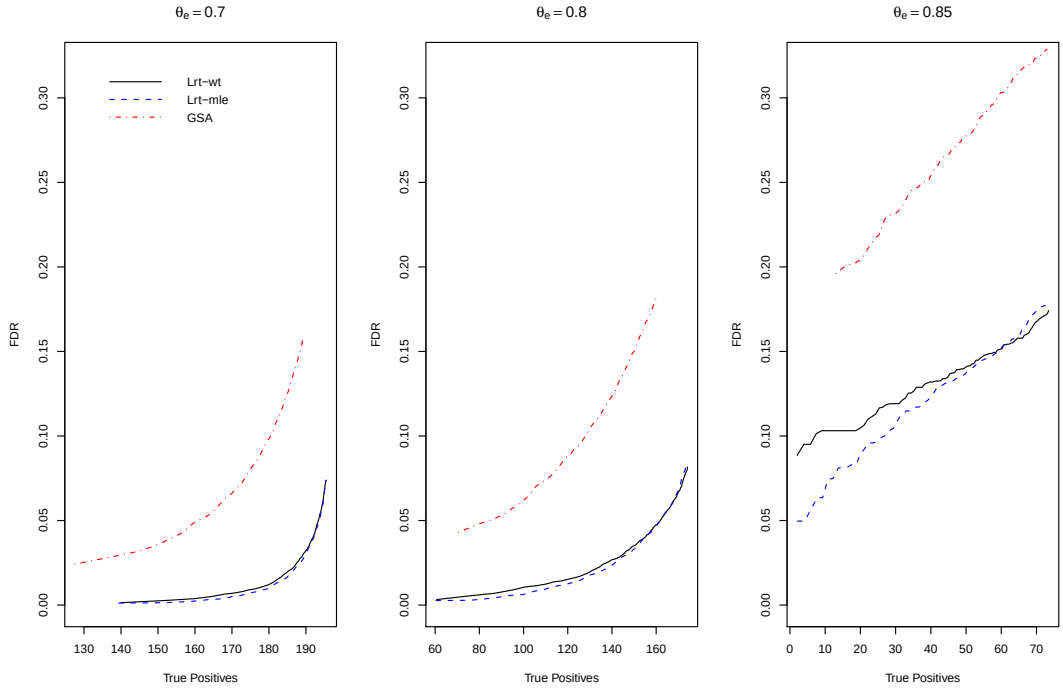


Figure A.40: (scenario 7) $\theta_0 = 0.95, m_e = 100, n = 15$

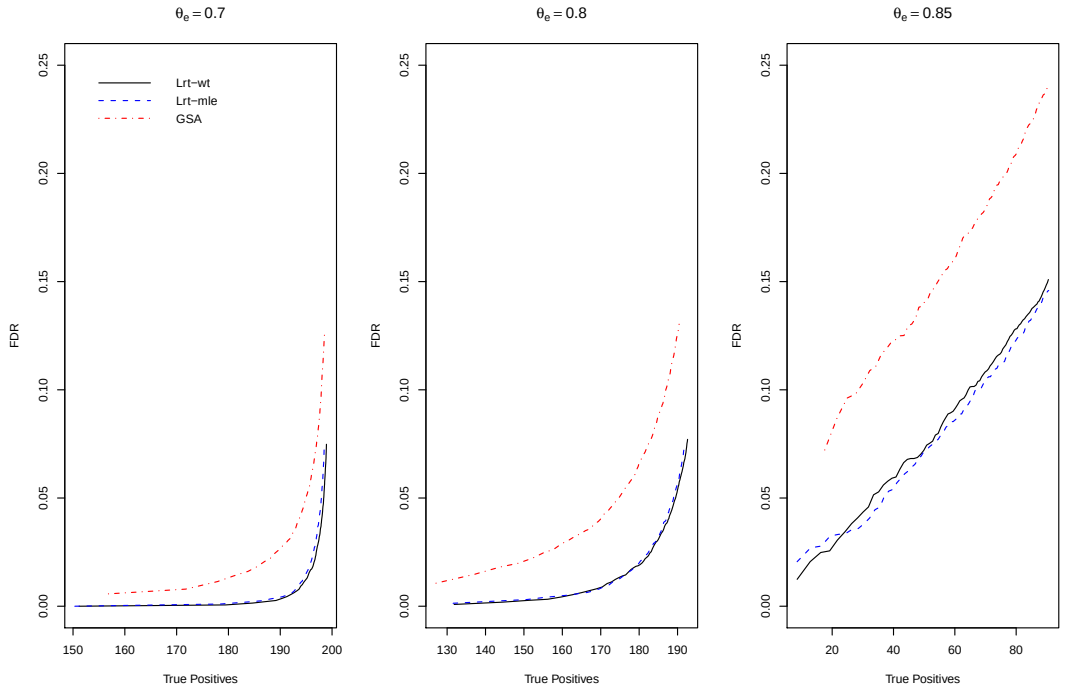


Figure A.41: (scenario 8) $\theta_0 = 0.95, m_e = 100, n = 25$