STATISTICAL ANALYSIS OF MICROARRAY EXPERIMENTS IN PHARMACOGENOMICS

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Youlan Rao, B.S., M.S.

The Ohio State University

2009

Dissertation Committee:

Jason Hsu, Adviser
Yoonkyung Lee, Co-adviser
Steven MacEachern

Approved by

Adviser
Graduate Program in Statistics
Pharmacogenomics is the co-development of a drug that targets a subgroup of the patients, and a device that predicts whether a patient is in the subgroup of responders to the drug. It is a two-stage process, including a training stage and a validation stage. The purpose of the training stage is to identify a biomarker positive (G+) subgroup of patients and its complement, the biomarker negative (G-) subgroup. Typically, subgroups are discovered by comparing the genetic profiles of the responders to the drug with the non-responders. Microarrays could be used to develop such a diagnostic device for identification of subgroups. The purpose of the validation stage is then to prove that the biomarker found in the training stage has sufficient sensitivity and specificity for clinical use, and to independently validate the efficacy and safety of the drug for the target G+ subgroup.

Major statistical problems in the analysis of microarray experiments in pharmacogenomics include normalization of gene expressions, biomarker selection in the training stage and determination of sample sizes for a validation study. Before doing any formal analysis on gene expression data, it is important to normalize the data first to reduce variation between arrays caused by sources of non-biological origin. Then for biomarker selection in the training stage, a re-sampling based multiple testing procedure is proposed by following the generalized partitioning principles. This procedure controls generalized Familywise
Error Rates (gFWER) asymptotically. To plan for a validation study, sample sizes for microarray experiments are determined to meet the pre-specified sensitivity and specificity requirements.

This dissertation is arranged as follows. Chapter 1 introduces the motivation of pharmacogenomics and design considerations of microarray experiments in pharmacogenomics. Chapter 2 compares different normalization techniques for microarray experiments. Chapter 3 focuses on the strong control of gFWER in multiple hypothesis testing. The re-sampling based multiple testing procedures are applied to select differentially expressed genes in the training stage. Chapter 4 formulates sample size determination procedures for validation studies with change of platforms taken into account. Chapter 5 discusses future research.
This is dedicated to my daughter, Rebecca Ruiyi Yao
ACKNOWLEDGMENTS

I would like to express my sincere gratefulness to my advisor, Dr. Jason Hsu and my co-advisor, Dr. Yoonkyung Lee for their patience and inspiring advices.

I would also like to thank Dr. Steven MacEachern, Dr. Elizabeth Stasny, Dr. Doug Wolfe, Dr. Noel Cressie, Dr. Saleh Tanveer from the Department of Mathematics, Dr. David Jarjoura from the Center for Biostatistics and Dr. Weike Wang from Shanghai Jiaotong University for their help and support during my long journey of PhD. study.

I am enormously grateful to my daughter, my husband, Yonggang Yao and my family for their love and support.
VITA

1996 .................................................. B.S. in Mathematics, Wuhan University
1998 .................................................. M.S. in Mathematics, Wuhan University
2000-2002 ............................................. M.S. in Mathematics, the Ohio State University
2003-present ................................. PhD candidate in Mathematics and in Statistics, the Ohio State University

PUBLICATIONS

Research Publications


FIELDS OF STUDY

vi
Major Field: Statistics

Studies in:

  Topic 1  Multiple testing
  Topic 2  Statistical Genetics
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract .............................................................</td>
</tr>
<tr>
<td>Dedication ..........................................................</td>
</tr>
<tr>
<td>Acknowledgments ....................................................</td>
</tr>
<tr>
<td>Vita .................................................................</td>
</tr>
<tr>
<td>List of Tables .......................................................</td>
</tr>
<tr>
<td>List of Figures .....................................................</td>
</tr>
</tbody>
</table>

## Chapters:

1. Microarray Experiment in Pharmacogenomics .......................... 1
   1.1 Purpose of Pharmacogenomics ...................................... 1
   1.2 Two-Stage Process of Pharmacogenomics .......................... 3
   1.3 Design Consideration in Microarray Experiment .................. 4

2. Comparison of Normalization Techniques for Microarray Experiment .... 9
   2.1 Introduction ...................................................... 9
   2.2 Normalization Methods .......................................... 11
   2.3 Description of Data ............................................. 12
   2.4 Analysis .......................................................... 13
   2.5 Conclusion ...................................................... 24
3. GFWER-controlling Multiple Testing Procedures in Microarray Experiment
  
  3.1 Introduction ......................................................... 25
  3.2 Principles in Constructing GFWER-controlling Multiple Tests ........ 30
    3.2.1 Partitioning Principle ........................................ 31
    3.2.2 Generalized Partitioning Principle ................................ 35
  3.3 Step-down Shortcuts .................................................. 37
    3.3.1 Sufficient Condition for Partitioning Step-down Tests .............. 38
    3.3.2 Sufficient Condition for Generalized Partitioning Step-down Tests 42
    3.3.3 Enforced Exchangeability Step-down Tests .......................... 46
  3.4 Re-sampling Based Multiple Tests Controlling GFWER .................. 49
    3.4.1 Linear Model for Microarray Data .................................... 52
    3.4.2 Estimating Dependence by RB Re-sampling ............................ 56
    3.4.3 RB Re-sampling Based Step-down Tests .............................. 57
    3.4.4 Asymptotic Control of GFWER ...................................... 59
  3.5 A Proof of Concept Experiment ........................................ 63
    3.5.1 Normalization Using an External Reference Sample .................. 63
    3.5.2 Multiple Testing for Differential Expressions ...................... 69
  3.6 Software Implementation .............................................. 70
    3.6.1 SAS Procedures .................................................. 71
    3.6.2 R Functions ..................................................... 72

4. Determination of Sample Size for Validation Study in Pharmacogenomics
   ................................................................. 75
  4.1 Introduction ............................................................ 75
  4.2 Formulation of Sample Size Calculation ................................ 79
    4.2.1 Basic Setting for Training and Validation Data ....................... 79
    4.2.2 Minimal Sensitivity Requirement .................................... 80
    4.2.3 Estimation of True Sensitivity ................................... 83
  4.3 Data Analysis .......................................................... 85
    4.3.1 Construction of Diagnostic Rule .................................... 85
    4.3.2 Sample Sizes Determination ....................................... 88
    4.3.3 Confidence Lower Bound for Sensitivity .............................. 89
  4.4 Numerical Study ....................................................... 94
  4.5 Conclusions ............................................................ 97

5. Summary and Future Work ................................................ 99

Bibliography ............................................................... 102
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Outcome of testing multiple hypotheses</td>
<td>26</td>
</tr>
<tr>
<td>3.2</td>
<td>Number of rejections by the step-down method controlling the gFWER at 5%, Holm method, partition method using Markov’s inequality, augmentation method, and resampling method.</td>
<td>71</td>
</tr>
<tr>
<td>4.1</td>
<td>Proportions of $\hat{S}en_v$ greater than pre-specified minimal level of sensitivity $\gamma = 0.8$ for different values of $n_v^n$ and confidence levels $1 - \alpha$ of lower bounds of the true sensitivity.</td>
<td>97</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Hybridization design in the proof of concept experiment. (samples from different groups and UMRS are represented by different shades of gray, for example, MA2\textsubscript{3} is the third cRNA sample from the second mouse from group A, while UMRS\textsubscript{2} is the second UMRS).</td>
<td>8</td>
</tr>
<tr>
<td>2.1 MA and loess plot of expression values for the human brain tissue data. A) without normalization, B) after median normalization, C) after quantile normalization and D) after cyclic loess.</td>
<td>14</td>
</tr>
<tr>
<td>2.2 Weighted smoothed medians of difference of expression values for the human brain tissue data. A) without normalization, B) after median normalization, C) after quantile normalization and D) after cyclic loess.</td>
<td>17</td>
</tr>
<tr>
<td>2.3 Smoothed MADs versus median averages for the human brain tissue data. A) without normalization, B) after median normalization, C) after quantile normalization and D) after cyclic loess.</td>
<td>18</td>
</tr>
<tr>
<td>2.4 Confidence band of the bias for the human brain tissue data. A) without normalization, B) after median normalization, C) after quantile normalization and D) after cyclic loess.</td>
<td>19</td>
</tr>
<tr>
<td>2.5 MSE curves without normalization (black, solid line), after median normalization (green, dashed line), and after quantile normalization (red, dotted line) after cyclic loess (blue, dotted line).</td>
<td>20</td>
</tr>
<tr>
<td>2.6 Mean of MSEs for the difference in expression values without normalization (0 and black), after median normalization (1 and green), after quantile normalization (2 and red) and after cyclic loess (3 and blue).</td>
<td>22</td>
</tr>
</tbody>
</table>
2.7 Mean of CVs for the difference in expression values without normalization (0 and black), after median normalization (1 and green), after quantile normalization (2 and red) and after cyclic loess (3 and blue). 23

3.1 Flow chart of computing critical values by modeling the data and estimating dependence by re-sampling. 51

3.2 Difference estimates between mutated types and wild type under quantile normalization and UMRS-based normalization. 66

3.3 Observed expression levels of genes from inbred mice. Rows, from top to bottom, represent gene 1 to gene 99. Columns correspond to samples from five groups of mice, arranged from left to right: group A, B, C, D, and W (wildtype). There are four mice per group, and four samples per mouse. The 80 columns, from left to right, represent sample 1 of mouse 1 from group A, to sample 4 of mouse 4 from group W. 67

3.4 Estimated expression levels of genes from inbred mice, with Mouse and Sample effects removed. Rows, from top to bottom, represent gene 1 to gene 99. Arranging the columns by unsupervised machine learning (clustering) results in five distinct groups: from left to right, groups C, D, A, B, and W (wildtype). 68

4.1 The probability $P(\hat{S}_{env} \geq \gamma)$ as a function of $m$ and its lower bound $1 - I_{1-\text{senv}}(m - m\gamma, m\gamma + 1)$ when the true sensitivity $\text{senv}$ is 0.90 and the minimal level $\gamma$ is 0.80. The horizontal dashed line indicates the minimal probability of successful validation $1 - \beta$, with $\beta$ fixed at 0.05. $m_H$ is the upper bound given by Hoeffding’s inequality in (4.3), $m_0$ is the root of $I_{1-\text{senv}}(m - m\gamma, m\gamma + 1) = \beta$, and $m^*$ is the desired sample size. $m_H = 74.9$, $m_0 = 36.8$ and $m^* = 30$ in this example. 81

4.2 Sample sizes $m^*$ necessary to meet the minimal sensitivity requirement of $\gamma = 75\%, 80\%$ and $85\%$ with $95\%$ of confidence as the true sensitivity varies. 84

4.3 p-value as a function of $\eta_0$ for testing $H_0 : \eta_{a,b} = \eta_0$ versus $H_a : \eta_{a,b} > \eta_0$ when $n_w = 2$, $n_p = 32$, $n_v = 2$, and thus $df = 46.11$ , $\hat{\omega}_b = 2.15$ and $\eta_{\text{obs}} = 1.94$ from the mice experiment. The solid line is for non-central $t$-distribution and the dashed line is for normal distribution. The $95\%$ confidence lower bounds from non-central $t$ and normal approximation are $\eta_{LB}^t = 1.098$ and $\eta_{LB}^N = 1.177$ respectively, and the former is smaller than the latter. 93
4.4 Sensitivity estimates by the simple plug-in estimator and $100(1 - \alpha)\%$ confidence lower bounds for varied number of replicated samples per subject with $\alpha = 0.10$ and 0.05 respectively, given $n^v_r = 2$ and $n^v_p = 32$ in the mice experiment. ........................................... 94

4.5 The required number of subjects per group and number of replicated samples per subject for various minimal sensitivity levels $\gamma$ when sensitivity is estimated either by the simple plug-in estimator (left panel) or by 95% confidence lower bounds (right panel). $n^v_r = 2$, $n^v_p = 32$ and $\beta = 0.05$. . . . 95
1.1 Purpose of Pharmacogenomics

Pharmacogenomics (pharmaco-genomics) is a word combining pharmacology (drug) with genomics. According to the Food and Drug Administration (FDA)’s voluntary genomics data submission (VGDS) paper and the drug-diagnostic co-development concept paper (FDA, 2005),

The promise of pharmacogenomics lies in its potential to help identify sources of inter-individual variability in drug response (both effectiveness and toxicity); this information will make it possible to individualize therapy with the intent of maximizing effectiveness and minimizing risk.

Although finding a drug working for the whole population of patients with certain disease is highly expected, it is not always achievable. If a drug fails to show efficacy for the whole population, it’s still possible that the drug is efficacious for a subpopulation. The purpose of pharmacogenomics is the co-development of a drug and a device, with the drug targeting a patient subpopulation, and the device predicting whether a patient is in this sub-population of responders to the drug. A good example of pharmacogenomics is the drug Herceptin (Trastuzumab) manufactured by Genentech. Herceptin is the FDA-approved therapeutic for patients with metastatic breast cancer caused by overexpression of some...
protein, named as human epidermal growth factor receptor 2 (HER2) protein. About 25% of breast cancers are caused by tumors that over-express HER2 and the women with these malignancies benefit from Herceptin therapy. To determine if a breast cancer patient has over-expressed HER2, two types of tests are available: Fluorescent in situ hybridization (FISH) and Immunohistochemistry (IHC).

Microarrays is a breakthrough biotechnology developed at the end of the last century, and it allows researchers to measure expression levels of thousands of genes simultaneously. One of the major applications of microarrays is to find signature genes characterizing different samples. We can think of using a well-designed microarray as a device to measure gene expression levels of a patient to predict her/his response to a drug. An example of microarrays for disease prognosis is MammaPrint, a device manufactured by Agendia. MammaPrint is the FDA-approved diagnostic test to predict the likelihood of tumor recurrence of a breast cancer patient based on her expression levels in tumor samples. van’t Veer et al. (2002) and van de Vijver et al. (2002) identified and validated a 70-gene signature that predicts the development of distant metastases in lymph node negative primary breast cancer. The MammaPrint test measures the level of expression of each of these 70 genes in a sample of a woman’s surgically-removed breast cancer tumor and then uses a specific algorithm to produce a score that determines whether the patient is deemed low risk or high risk for spread of the cancer to another site.

Pharmacogenomics provides subtype-specific tailored treatment for patients. It can provide more accurate prognosis and more efficient drug therapy for patients while minimizing risk for patients who do not belong to the subpopulation.
1.2 Two-Stage Process of Pharmacogenomics

The development in pharmacogenomics includes two major stages: a training stage and a validation stage. The purpose of the training stage is to identify a biomarker positive (G+) subgroup of patients and its complement, the biomarker negative (G-) subgroup. Biomarkers (Baker, 2005) historically refer to substances in biological samples or measurements that indicate a person’s disease state or response to a drug. They are crucial for efficient drug development, and there are various types of biomarkers, such as disease biomarkers, surrogate endpoints, efficacy or outcome biomarkers. Disease biomarkers indicate the presence or likelihood of a particular disease in patients. For example, gene expression profiles linked with cancer are such kind of biomarkers. Typically, subgroups are discovered (Wang, O’Neill and Hung, 2007) by comparing the genetic profiles of the responders to the drug with the non-responders. Microarrays could be used to develop such a diagnostic device for identification of subgroups. The purpose of the validation stage is then to prove that the biomarker found in the training stage has sufficient sensitivity and specificity for clinical use, and to independently validate the efficacy and safety of the drug for the target G+ subgroup.

Transition from the training stage to the validation stage typically involves a change of platform as a subset of potential predictors of drug response are identified as a biomarker in the first stage and only those are used in the second stage. For example, in the training stage of MammaPrint, microarrays probing approximately 25,000 genes were used, while in the second stage of validation, microarrays probing 70 genes only were used. Generally, fewer genes are involved in a diagnostic device to be used, and this change allows more replication of probes and more replicated samples from a subject in the validation stage than the training stage. To ensure the diagnostic algorithm derived from the training stage
applies to expressions measured in the validation stage, the platform change needs to be taken into account when planning a validation study and computing sample sizes required for the study. This issue is often overlooked in the process of developing a diagnostic device. As stated in FDA (2005),

> When validating a gene or expression pattern, instead of a set of individual biomarkers, a rigorous statistical approach should be used to determine the number of samples, and the methodology used for validation. It is recommended that the validation strategy be discussed in advance with FDA.

At the end of the training stage, the sensitivity and specificity of the diagnostic algorithm for a validation trial need to be estimated. If both the sensitivity and specificity are significantly high, pharmacogenomic development is recommended to proceed. Otherwise, further pharmacogenomic development is likely to be futile.

To implement the first stage and the successful transition to the second stage in pharmacogenomics development, we would recommend the following steps:

1. Design a training microarray experiment statistically.

2. Select marker genes by multiple testing with properly controlled error rate to train a classification algorithm with.

3. Estimate variance components to calculate sample sizes for the clinical prognostic chip to achieve desired sensitivity and specificity.

### 1.3 Design Consideration in Microarray Experiment

Three basic techniques fundamental to experimental design are replication, blocking and randomization. Replication helps to reliably estimate variabilities and to increase precision in the experiment, blocking helps to avoid confounding with uninterested extraneous
factors, and randomization helps to decrease bias. The key design considerations in the mi-
croarray experiment for training stage are including external reference sample to control for
platform change and using the three basic principles to design the experiment statistically.
The reasons for these design considerations are elucidated as follows:

Platform change For example, the MamaPrint changed from no replication of probes in
the training stage to triplicate of the probes in the validation stage (Glas et al. 2006, Buyse
et al. 2006). There are possible systematic shifts in the expression levels due to platforms
change. A remedy for this is to add external reference sample (ERS) and normalize expres-
sion levels measured from the samples we are interested in relative to those from the ERS
to remove these systematic shifts, since the data measured from external reference sample
remain homogeneous from the first stage to the second stage.

Replication Microarray experiments to measure gene expression levels may contain six
sources of variability: treatment factor or risk groups (e.g. groups treated by a new drug
or not), subject (patient) effect, sample effect, gene effect, probe effect and noise effect
coming from various non-biological sources. To estimate these variabilities, we suggest to
replicate subjects within each group, replicate samples within each subject and replicate
probes within each gene, if sample quantity and manufacturing technology allow. In this
case, we can reliably estimate the difference of gene expression levels between different
risk groups since the variance of the difference between two risk groups depends on the
variances of subjects, samples and noise, which can be estimated separately.

Blocking Gene expression measurements from microarrays are potentially affected by
extraneous array processing effects. In this context, a statistical block is a condition under
which measured gene expressions are likely to be equally affected by these confounding factors. A block might, therefore, be an array, or a batch of arrays processed together. For example, NimbleGen 12-well arrays can have 12 biological samples hybridized on the same array, so each array can conveniently form a block. We suggest to keep proportion of samples from the groups to be the same for each block, so that one can expect the collection of expression levels of the genes to be approximately the same across blocks and normalize all the blocks together by quantile normalization (Bolstad et al. 2003). Such a design can avoid confounding due to extraneous effects and increase sensitivity and specificity of a prognostic algorithm because it eliminates array/batch variability.

**Randomization**  We suggest to randomize samples hybridization and to randomize probes placement in the array. If the placement of the biological samples into microarray is not randomized, it may have bias in expression levels because of the batch processing effect (for a microarray allowing one sample placed on one array) or the position effect (for a microarray allowing multiple samples placed on one array). If the placement of probes into microarrays is not randomized, measurements from the training stage to validation stage may have different biases. In most microarray experiments, stock genomewide scan microarrays, with systematic placement of probes in the microarray and with no replication of probe sets, such as Affymetrix HG-U95 or HG-U133 chips, are used to profile each patient by analyzing one sample from each patient using one chip. With one patient’s sample per array, it is impossible to separate array to array variability from sample to sample variability. With systematic placement of probes, it can cause bias on the estimate of difference of gene expression levels between two risk groups as platform changes. With no replication of probe sets, it can cause increase of error rate in the prognostic algorithm.
Hsu et al. (2009) describes such a microarray experiment to prove the concept that a training experiment can be designed statistically to reliably estimate the variance components of subject, sample, and noise separately. To simulate the comparison of phenotype groups, the experiment compared tissues from normal mice (wild type, labeled group W) with tissues from four groups of mice (labeled groups A, B, C and D) with four different mutations of the microphthalmia transcription factor (Mitf) gene.

Four mice were sampled from each of the five groups. From each mouse, four cRNA samples were prepared after total RNA was isolated and biotin labeled cDNA was synthesized from a spleen tissue sample.

To prove the concept that external reference samples can be used to control for the platform change, we also prepared 16 samples of Universal Mouse Reference Sample (UMRS) from StrataGene.

To demonstrate statistical design of hybridization of samples to microarrays, we utilized NimbleGen microarrays with 12 mini-microarrays on each array. The 96 samples were hybridized to eight arrays, with the samples placed in the mini-microarrays according to the three rows by four columns patterns shown in Figure 1.1, following the statistical principles of randomization, replication, and blocking.

A total of 99 genes thought to be regulated by the Mitf gene were selected as probes. The probe set for each gene consisted of thirty-two 24-mer probes. To demonstrate the utility of statistically designing microarrays according to the principles of randomization and replication, each probe set was replicated twice in each of the mini-microarrays, and placements of the probes were completely randomized in each mini-microarray.
Figure 1.1: Hybridization design in the proof of concept experiment. (samples from different groups and UMRS are represented by different shades of gray, for example, MA23 is the third cRNA sample from the second mouse from group A, while UMRS2 is the second UMRS).
CHAPTER 2

COMPARISON OF NORMALIZATION TECHNIQUES FOR MICROARRAY EXPERIMENT

2.1 Introduction

In microarray experiments, variation of expression measurements among arrays can be attributed to many sources, such as differences in sample RNA preparation, cDNA labeling, image intensity and microarray hybridization/wash efficiency. Normalization of expression levels applied to microarray data can help in removing this error. Different methods, including cyclic loess, quantile normalization (Bolstad et al. 2003) and median or mean normalization (Churchill 2002, Churchill 2003, Churchill and Oliver 2001, Kerr and Churchill 2001, and Wolfinger et al. 2001), have been utilized to normalize microarray data. Briefly, cyclic loess makes the MA plot of probe intensities from every pair of arrays scatter about the $M = 0$ axis, quantile normalization makes the distributions of expression levels the same across arrays, and median or mean normalization shifts the individual log-intensities on each array so that the median or mean log-intensities, respectively, are the same across arrays. These normalization algorithms can be applied either globally to an entire data set or locally to some physical subset of the data (Quackenbush 2002). Irizarry et al. (2003) applied the quantile normalization procedure to normalize dilution data and spike-in data.
from Affymetrix arrays, and showed how quantile normalization removed bias as compared to no normalization. Their analysis was unique in that they knew the true expression levels and could therefore determine the degree of bias reduction from quantile normalization.

In this chapter, we compare the performance of median, cyclic loess, quantile, and no normalization for microarray experiment. We investigate these methods over a MicroRNA (miRNA) microarray data, which includes 72 microarrays obtained from RNA from 26 human and 10 mouse tissues that were hybridized as technical replicates. Hence, each RNA sample was hybridized to two independent microarrays. Since replicate samples should, in theory, have almost identical values for expressions, one can compare different normalization techniques in terms of the closeness of normalized measurements in the replicated samples. The differences between these paired expression levels with and without normalization can be divided into a bias and variance components by expression level. Both of these miRNA-by-miRNA differences components should be reduced after applying normalization methods. We used these differences to provide direct evidence of the capability of each method of reducing these two components. It was critical to examine the effects on both quantities because the complexity of a transformation may increase the error variance over and above its bias reduction. To resemble how normalization is typically applied to samples, normalization was done globally across all 72 samples.

Section 2.2 describes the normalization methods in detail. Section 2.3 describes the miRNA data used in this chapter. Section 2.4 compares normalization methods.
2.2 Normalization Methods

Three commonly used normalization techniques are reviewed. Suppose that we have the (log base 2 transformed) probe level expression values from $p$ miRNAs and $n$ arrays in a $p \times n$ matrix $X = \{X_{ij}\}$ with $i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, n$.

**Median normalization** shifts miRNAs expressions on each array by additive constants so that the medians of miRNAs expressions are the same across arrays by the following steps:

- Take the median of each column of $X$ and generate an $n$-dimensional median vector $M$, $M = (\text{median}_i \{X_{i1}\}, \text{median}_i \{X_{i2}\}, \ldots, \text{median}_i \{X_{in}\})^T$;
- Calculate the overall median of the vector $M$ and denote $c_M = \text{median}_j \{M_j\}$;
- Shift miRNAs expression values of each array by subtracting the difference between the median of each array and the overall median from them. The normalized miRNA expressions will then be $X_{ij} - (\text{median}_i \{X_{ij}\} - c_M)$ for $j = 1, 2, \ldots, n$.

Instead of matching the median only across the arrays, **quantile normalization** makes the distributions of expression levels the same across arrays by the following steps:

- Sort each column of $X$ separately to generate a sorted $p \times n$ matrix $Y = \{Y_{ij}\}$ with $i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, n$;
- Take the mean of each row of $Y$ and generate a $p$-dimensional vector $A_b$, then $A_b = (\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_p)^T$. Assign $A_b$ to each column of $Y$ to get a matrix denoted as $X_{\text{sort}}$;
- Get the normalized miRNAs expressions for each array by rearranging each column of $X_{\text{sort}}$ to have the same ordering as the corresponding column of the matrix $X$ so
that empirical distributions of the normalized miRNA expressions are the same as that of $A_b$ across arrays.

**Cyclic loess** considers the MA plot of probe intensities from every pair of arrays $(X_{ij}, X_{ij'})$, with fixed $j \neq j'$ and $i = 1, 2, ..., p$, and makes the M and A pairs scattered around the $M = 0$ axis by the following steps:

- Compute $M_i = X_{ij} - X_{ij'}$ and $A_i = \frac{1}{2}(X_{ij} + X_{ij'})$;
- Fit a loess curve by regression of $M$ on $A$, and denote the fitted vector by $\hat{M}$;
- Setting the vector $D = (M - \hat{M})/2$, get the normalized miRNAs expressions for $(X_{ij}, X_{ij'})$ by modifying $X_{ij}$ to $X_{ij} + D_i$ and $X_{ij'}$ to $X_{ij'} - D_i$, $i = 1, 2, ..., p$.

### 2.3 Description of Data

MicroRNAs (miRNAs) are noncoding RNAs of 19-24 nucleotides that are negative regulators of gene expression. Recently implicated as important in development and normal physiology, microRNAs are abnormally expressed in many human cancers (Volinia et al. 2006, Lu et al. 2005). Moreover, aberrant microRNA expression has been shown to initiate and promote carcinogenesis (reviewed in Hagan and Croce 2007). These microRNA expression signatures may reveal new oncogenetic pathways in human cancers. For systematic investigation of microRNA expression, oligonucleotide-based microarrays for microRNAs in human and mouse tissues have been developed recently (Liu et al. 2004) and several commercial platforms are now available.

The Ohio State University Comprehensive Cancer Center Version 3.0 microRNA microarray was used, and it contains 3790 probes spotted in duplicate. The probes are 40 nucleotides in length, consisting of the genomic sequence that has the mature microRNA
sequence and additional flanking bases. With the exception of six probes designed against *Arabidopsis thaliana* microRNAs, the rest of the probes are derived from known and predicted human and mouse microRNAs. This design allows for the detection of mature as well as precursor miRNAs and is particularly helpful in determining if computationally predicted miRNAs are real.

Although microRNA microarrays are lower density spotted arrays than mRNA microarrays, they are not “boutique” arrays. For example, microRNA arrays do not meet the following criteria: “more than half the probes might be differentially expressed between any two samples and that the differential expression might be predominately in one direction” (Oshlack et al. 2007). We also do not expect global differences across miRNA arrays.

### 2.4 Analysis

Background-corrected median signals for duplicate probes on an array were averaged. After normalization across all 72 arrays, let $X_i$ be the log base 2 transformed expression value of the $i$th miRNA for a certain tissue, and let $Y_i$ be the log base 2 transformed expression value of the $i$th miRNA for the replicate of the tissue.

**Bias.** The average $A_i = (X_i + Y_i)/2$ and the difference $M_i = X_i - Y_i$ of expression values for each miRNA can then be computed. The MA plot of the two vectors $X_i$ and $Y_i$ is a 45-degree rotation and axis scaling of their scatter plot. This plot is particularly useful for array data because $M_i$ represents the log fold change and $A_i$ represents the average log intensity for the $i$th miRNA. When the loess curves of the MA plot deviate from the horizontal line at $M = 0$, this demonstrates differences in the intensity levels between two arrays from the same tissue (Gentleman et al. 2005). In contrast, if the loess curves align with $M = 0$, the normalization method is considered to exhibit little bias at all levels of
expression. When MA plots and loess curves were made for the replicate array data from human brain tissue using no normalization, median normalization, quantile normalization and cyclic loess, we observed that the quantile normalization method removed bias the best (Figure 2.1C), the loess curve closely followed the horizontal line at $M = 0$. No normalization, median normalization and cyclic loess behaved similarly in that their loess curves are not aligning with $M = 0$ closely enough (Figure 2.1A, 2.1B and 2.1D).

Figure 2.1: MA and loess plot of expression values for the human brain tissue data. A) without normalization, B) after median normalization, C) after quantile normalization and D) after cyclic loess.
**Binning.** To compare the normalization methods in how much they reduced error variance in addition to reducing bias, we formally modeled the mean and variance of differences in replicate arrays as a function of their expression levels. In order to obtain reliable estimates of the expression levels, we binned duplicates according to their average expression level first and then proceeded by modeling the mean and variance based on the binned data.

We created equally-sized bins containing 34 miRNAs probes. For each bin, we summarized the differences in the replicate arrays by median absolute deviation (MAD) of the differences and median of the differences to obtain robust estimates of variance and bias, respectively (Lin et al. 2002). The smoothed MADs and medians of the differences were used to detect systematic effects due to the different normalization methods as a function of expression levels. Lower values of smoothed MADs and smoothed medians closer to zero across average expressions correspond to a superior normalization method.

As stated above, each bin consisted of 34 miRNAs probes. For fixed \( k \) \((1 \leq k \leq K)\), let \( X_{(i)k} \) \((i = 1, 2, ..., 34)\) be the expression value of the \( i \)th miRNA in the \( k \)th bin for a specific tissue, and let \( Y_{(i)k} \) \((i = 1, 2, ..., 34)\) be the expression value of the \( i \)th miRNA in the \( k \)th bin for the replicate of the tissue. The difference between the replicate arrays expression values for each miRNA in the \( k \)th bin can be denoted by \( D_{(i)k} = X_{(i)k} - Y_{(i)k} \) \((i = 1, 2, ..., 34)\), and the corresponding observations by \( d_{(i)k} \). We assume that for fixed \( k \),

\[
D_{(i)k} \sim \text{i.i.d.} N(\mu_k, \sigma_k^2) \quad i = 1, 2, ..., 34
\]

and use

\[
md_k = \text{median}(d_{(i)k})
\]
as a robust location (center) estimate of $\mu_k = \mathbb{E}[D_{(1)}k]$, and

$$MADd_k = \text{median}_{1 \leq i \leq 34}|d_{(i)k} - \text{median}_{1 \leq i \leq 34}(d_{(i)k})|,$$

as a robust estimate of scale (spread), which is proportional to $\sigma_k = \sqrt{\text{var}[D_{(1)}k]}$ under normality.

For the average expression values of miRNAs in the $k$th bin across certain tissue replicates, let $A_{(i)k} = (X_{(i)k} + Y_{(i)k})/2$ ($i = 1, 2, ..., 34$) and $a_{(i)k}$ be the $i$th observation. Similarly, for estimation of the center of the average expression values in each bin, we consider

$$ma_k = \text{median}_{1 \leq i \leq 34}(a_{(i)k}).$$

As Figure 2.1A suggests, it is sensible to model $\mu_k$ and $\sigma_k$ as a function of the center of the average expression values of miRNA replicates in the $k$th bin.

For the paired observations $(ma_1, md_1), (ma_2, md_2), ..., (ma_K, md_K)$, we modeled the median difference as a smooth function of the median average

$$md_k = \eta(ma_k) + \epsilon_k, \quad k = 1, 2, ..., K$$

with $\epsilon_k \sim N(0, \sigma_m^2)$ and with a different variance for each bin. The smoothed relationship $\eta$ was obtained by the weighted smoothing spline with weights equal to the reciprocal of the squared MAD of difference. Quantile normalization gave the best results when comparing the weighted smoothed curves for the median difference in expression values using the human brain tissue data (Figure 2.2).

Similarly, for the paired observations $(ma_1, MADd_1), (ma_2, MADd_2), ..., (ma_K, MADd_K)$, we considered the following model with unequal variance

$$MADd_k = \xi(ma_k) + \epsilon_k, \quad k = 1, 2, ..., K$$

16
Figure 2.2: Weighted smoothed medians of difference of expression values for the human brain tissue data. A) without normalization, B) after median normalization, C) after quantile normalization and D) after cyclic loess.

and $\epsilon_k \sim N(0, \sigma^2_{MAD})$. The smoothed MAD of differences $\xi$ can again be obtained by smoothing splines with the smoothing parameter selected by generalized maximum likelihood (GML) (Gu 2002). It was difficult to see differences in the relationship between $MAD_d$ and $ma$ among the normalization methods (Figure 2.3), but they became more apparent if the bias and variance were combined into a mean-squared error statistic.

**Confidence intervals.** The fitted medians of differences $\eta$ is the smoothed estimate of bias parameter $\mu_k$, and the fitted MAD of differences $\xi$ is the smoothed estimate of
Figure 2.3: Smoothed MADs versus median averages for the human brain tissue data. A) without normalization, B) after median normalization, C) after quantile normalization and D) after cyclic loess.

scale parameter. We used the fitted MAD to estimate confidence intervals around bias and obtained a pointwise confidence interval for the bias by binned expression values as

\[ \hat{\eta}(ma_k) \pm \frac{3.98}{\sqrt{34}} \hat{\xi}(ma_k), \]

(see Hoaglin et al. 2000). The confidence band after quantile normalization encompasses the horizontal line at \( M = 0 \), while those using no normalization, median normalization or cyclic loess do not include zero for larger expression values (Figure 2.4).
Figure 2.4: Confidence band of the bias for the human brain tissue data. A) without normalization, B) after median normalization, C) after quantile normalization and D) after cyclic loess.

**Mean Squared Error.** We obtained the mean squared error (MSE) of the difference in expression values (including variance and squared bias)

$$\text{MSE}_k = \mathbb{E}[D_{(1)k}^2] = \text{var}[D_{(1)k}] + \mathbb{E}[D_{(1)k}]^2 = \sigma_k^2 + \mu_k^2,$$

which can be estimated by the smoothed estimates

$$\frac{\hat{\xi}(ma_k)}{0.6745}^2 + \hat{\eta}(ma_k)^2,$$
(see Huber 2003). The estimated MSE for quantile normalization is smallest when average expression values are greater than noise levels of measurements, and the estimated MSE for cyclic loess is slightly larger than that of quantile normalization across all average expression values. Median normalization performed similarly to no normalization (Figure 2.5).

![MSE curves](image)

Figure 2.5: MSE curves without normalization (black, solid line), after median normalization (green, dashed line), and after quantile normalization (red, dot-dashed line) after cyclic loess (blue, dotted line).
To evaluate the global bias and variance for each method, we averaged MSEs across expression levels greater than 4.5; the value 4.5 (log base 2 transformed) was selected because 95% of the blanks (spots lacking oligonucleotide probes) gave intensities less than this value. The average MSEs for no normalization, median normalization, quantile normalization and cyclic loess using the brain tissue data were 0.278, 0.274, 0.225, 0.270 respectively. These results were found consistently across the other 35 tissue types (Figure 2.6), where the MSEs were lower for quantile normalization (coded 2) in almost all tissue samples compared to no normalization (coded 0), median normalization (coded 1) and cyclic loess (coded 3), except for human lung, human liver, human thymus, mouse liver and mouse lung. When the normalization methods were applied to each tissue type separately, instead of to all 72 arrays together, the results were similar.

**Checking for Scale Compression.** It is possible that the superior results for quantile normalization is the result of the compression of the scale downward after transformation. To check this, we first calculated coefficients of variation (CV) as the ratio of an estimate of the standard deviation of measurement ($\sqrt{\text{MSE}}$) for each bin to the mean expression for that bin and then average the ratios across bins. We found the CVs followed the same pattern as the MSEs, that is, typically lower values for quantile normalization across tissues (Figure 2.7). It is also possible that the superior results for quantile normalization is the result of compressing the scale from both ends after transformation; thereby reducing spread and sensitivity of transformed measurements. To check this, we calculated the average variance of expression levels across the 36 tissues for each miRNA. This variance consists of true variance across tissues and measurement error as obtained with the MSE. Averaging the variance across miRNAs and the MSEs across tissues, we found the ratios of signal
Figure 2.6: Mean of MSEs for the difference in expression values without normalization (0 and black), after median normalization (1 and green), after quantile normalization (2 and red) and after cyclic loess (3 and blue).

(true) variance to noise (measurement error) variance were 12.0, 14.0, 16.3 and 16.3 for no, median, quantile and cyclic loess normalization respectively.

Comparative Study We compare real-time RT-PCR miRNA data (Lee et al. 2008) with our microarray miRNA data, since twenty-one tissues were common to both datasets. Specifically, we focused on brain and heart, since these tissues are quite biologically distinct and have substantial differences in their miRNA expression profiles. If a normalization technique was overly aggressive, then there would be an “averaging-out” effect, leading to
Figure 2.7: Mean of CVs for the difference in expression values without normalization (0 and black), after median normalization (1 and green), after quantile normalization (2 and red) and after cyclic loess (3 and blue).

a significant decrease in the number of differentially expressed miRNAs. A well known difference between microarray and RT-PCR data is that the fold changes observed by microarray tend to be compressed in comparison with fold changes observed by RT-PCR. We found 51 miRNAs were characterized by a four fold difference in expression by RT-PCR. For the microarray data on identical miRNAs, we found that 36, 35, 35, 35 miRNAs were two fold differentially expressed for no, median, cyclic loess and quantile normalization respectively. This set of miRNAs was found to have roughly a 70% overlap with the RT-PCR
data. The observed values for fold changes varied little with respect to the normalization method used. In this respect, we could not conclude any superior normalization method based strictly on this analysis, but we could at least conclude that quantile normalization is not worse than other methods in terms of its sensitivity.

2.5 Conclusion

We showed that the quantile normalization method works best in reducing differences in miRNA expression values for duplicate tissue samples, cyclic loess works slightly worse than quantile normalization, whereas no normalization and median normalization behave similarly and seem to be inferior to quantile normalization and cyclic loess with regard to bias. This is not surprising because quantile normalization adjusted better for differential bias across the scale of expression values. By showing that the total MSE was lower across almost all 36 tissue samples, we were assured that the bias correction provided by quantile normalization was not outweighed by additional error variance that can arise from a more complex normalization method. Furthermore, we showed that quantile normalization does not achieve smaller replication error by compressing the scale downward or by compressing the scale from both ends.
CHAPTER 3

GFWER-CONTROLLING MULTIPLE TESTING PROCEDURES IN MICROARRAY EXPERIMENT

3.1 Introduction

Multiple testing is the general problem of simultaneously testing a family of hypotheses \( \{H_0_i, i = 1, 2, ..., K\} \), with the size of the family \( K \) to be finite or infinite. It is different from F-test, the homogeneity test; it requires telling specifically which \( H_0 \)'s are rejected, and which are not. We usually assume that tests for individual hypotheses \( H_01, H_02, ..., H_0K \) are available, and the main problem is how to combine them into a simultaneous test procedure.

If we simply test each individual hypothesis at level \( \alpha \), disregarding the multiplicity adjustment, then the probability of one or more false rejections for the family \( \{H_0_i, i = 1, 2, ..., K\} \) rapidly increases with \( K \) and is typically much larger than \( \alpha \) for large \( K \).

For example, assume the test statistics used for testing individual hypotheses \( H_01, H_02, ..., H_0K \) are independent (very strong assumption), and the level at which each of these hypotheses is tested equals \( \alpha = 0.05 \). When all of the hypotheses \( H_01, H_02, ..., H_0K \) are true, we can calculate the probability of one or more false rejections, which is defined as weak control of familywise error rate later.
In fact, let \( A_i \) denote the event of rejecting \( H_i \) with \( i = 1, 2, \ldots, K \). By assumption on individual test, \( P(A_i|H_i \text{ is true}) = 0.05 \). Take \( K = 4 \) for example.

\[
P(\text{one or more false rejection})
\]
\[
= P(\text{reject one or more hypotheses from } H_{0i}, i = 1, 2, 3, 4 | \text{all } H_{0i}, i = 1, 2, 3, 4 \text{ are true})
\]
\[
= P(A_1 \cup A_2 \cup A_3 \cup A_4 | \text{all } H_{0i}, i = 1, 2, 3, 4 \text{ are true})
\]
\[
= P((A_1^c \cap A_2^c \cap A_3^c \cap A_4^c) | \text{all } H_{0i}, i = 1, 2, 3, 4 \text{ are true})
\]
\[
= 1 - P(A_1^c \cap A_2^c \cap A_3^c \cap A_4^c | \text{all } H_{0i}, i = 1, 2, 3, 4 \text{ are true})
\]

By independence among \( A_i, i = 1, 2, 3, 4 \) (if suppress the conditional statement)

\[
P(\text{one or more false rejection}) = 1 - P(A_1^c) \cdot P(A_2^c) \cdot P(A_3^c) \cdot P(A_4^c)
\]
\[
= 1 - (1 - 0.05)^4 = 0.185 > 0.05.
\]

Generally, for any fixed \( K \), \( P(\text{one or more false rejection}) = 1 - 0.95^K \).

Table 3.1 lists the possible outcomes of testing \( K \) hypotheses simultaneously. The family of \( K \) hypotheses includes \( K_0 \) true null hypotheses and \( K_1 \) false null hypotheses. \( V \) denotes the number of false rejections, \( R \) denotes the total number of rejections, and \( S \) denotes the number of false hypotheses rejected.

|                  | Accepted | Rejected | total 
|------------------|----------|----------|------
| True null hypotheses | \( \bar{U} \) | \( \bar{V} \) | \( K_0 \)
| False null hypotheses | \( T \) | \( S \) | \( K_1 \)
| All hypotheses | \( \bar{W} \) | \( \bar{R} \) | \( K \)

Table 3.1: Outcome of testing multiple hypotheses
In high scale multiple testing problem with large $K$, the most commonly used type I error rates are the generalized familywise error rate (gFWER) and false discovery rate (FDR). gFWER is the probability that the number of false rejections exceeds a given threshold $m$, that is,

$$gFWER = P(V > m).$$

For $m = 0$, the type I error rate is called familywise error rate (FWER). FDR is the expected proportion of false rejections among all rejections, which is,

$$FDR = E\left(\frac{V}{R I_{R>0}}\right).$$

Let $I_0 = \{i, H_{0i} \text{ is true}\}$ denote the indices corresponding to the true null hypotheses, $K_0 = \text{card}(I_0)$. Since $I_0$ is unknown, the definition of error rates is ambiguous. For gFWER, weak control means control of $P(V > m)$ under the assumption that all null hypotheses are true, that is, $K_0 = K$; strong control means control of $P(V > m)$ under any possible null configurations.

In the setting of selecting genes which are differentially expressed between two groups, the individual (sometimes also called marginal) null hypotheses to be tested are

$$H_{0i} : \theta_i = 0,$$

for $i = 1, ..., K$, where $K$ is the number of genes probed on microarrays in the training stage, and $\theta_i$ is the expected mean difference of gene expressions between two groups for the $i$th gene. Multiple testing of $\{H_{0i} : \theta_i = 0|i = 1, 2, ..., K\}$ is different from the intersection hypothesis of $\{\theta_1 = 0, \text{ and } \theta_2 = 0, ..., \text{ and } \theta_K = 0\}$, as there are many possible configurations of true null hypotheses for multiple testing, while we assume $K_0 = K$ for the intersection hypothesis.
Let $\theta = (\theta_1, \ldots, \theta_K)$, and let $\Sigma$ denote generically all nuisance parameters that the observed expression levels depend on (including, for example, covariance of the expression levels for each of the two groups). Strong control of the gFWER keeps the probability of rejecting more than a fixed number $m$ of true null hypothesis at a pre-specified low level,

$$gFWER = \max_{I_0 \subseteq \{1, 2, \ldots, K\}} \sup_{\theta, \Sigma} P_{\theta, \Sigma} \{V > m\}, \tag{3.1}$$

where for any given configuration of null hypotheses $I_0 \subseteq \{1, 2, \ldots, K\}$, the supremum is taken over all possible $\theta$ and $\Sigma$ subject to $\theta_i = 0$ for $i \in I_0$.

In order to choose an appropriate multiple testing method, it is critical to select the definition of correct decisions that reflect the objective of the experiment. The FWER and gFWER are exceedance probabilities of the number of false discoveries. The FDR is the expectation of $V/R$ for $R > 0$, which is an expectation of the proportion of true null hypotheses among all rejected null hypotheses. Reporting of the expectation of some quantity measuring errors in multiple testing may be inadequate, if the quantity being controlled (or a component of it) turns out to be highly variable for a statistical method. Gordon et al. (2007) showed that, in terms of the number of false discoveries, the Benjamini and Hochberg (1995) method is less stable than the Bonferroni method. On the other hand, FDR is an unconditional error rate. It may be tempting for investigators to report conditional error rates in practice. It has been showed by Efron (2007) that conditional FDR (the false discover proportion) may be substantially higher or lower than the nominal (unconditional) FDR.

In this chapter, we focus on multiple testing procedures controlling gFWER strongly. Strong control of FWER or gFWER implies control of the long run relative frequency of incorrect decision across different studies.
Different procedures of controlling gFWER have been proposed in the literature. van der Laan, Dudoit, and Pollard (2004) proposed augmentation procedure to control gFWER. The procedure rejects the null hypotheses that would be rejected by any FWER-controlling multiple testing procedure; and then augments the rejections by automatically rejecting the null hypotheses associated with the next extreme test statistics. Lehmann and Romano (2005) proposed a Bonferroni-type single-step and a Holm-type step-down procedures that control the gFWER by Bonferroni inequality and Markov’s inequality respectively. These methods do not take joint distribution of the test statistics into account, and they are generally conservative.

Xu and Hsu (2007) proposed the generalized partitioning principle as a systematic technique of constructing gFWER-controlling procedures that can take the joint distribution of test statistics into account. It partitions the parameter space into disjoint subspaces and tests each hypothesis corresponding to the disjoint subspaces accordingly. A particular application of the generalized partitioning principle is to use Markov’s inequality to provide a gFWER-controlling test for each disjoint hypothesis.

Assuming the test statistics have an exchangeable distribution under each disjoint hypothesis, Xu and Hsu (2007) also constructed step-down methods that control gFWER while taking dependence among the test statistics into account. Step-down procedures that control the gFWER adjust for multiplicity conservatively, but only to the extent that some subset (but not all) of the null hypotheses of interest might be true, conditional on how many of them have been rejected. This is in contrast to a single-step procedure which typically adjusts for multiplicity under the scenario that all the null hypotheses are true. The critical value (threshold) used by a step-down procedure, in effect, is the one corresponding to the maximum subset hypothesis that could be true, conditional on data. Such conditional tests
keep the true error rate as close to the desired error rate as possible, while still guaranteeing conservatism.

General strategies for taking dependence across different individual test statistics (or genes in genomic data) into account are to model dependence as arising from a latent variable or to estimate dependence by re-sampling. When the test statistics have a multivariate normal or a multivariate $t$ distribution, and the correlation is assumed to be exactly, or approximately of a one-factor structure, the factor analytic technique of Hsu (1992) applies. When the exact form or an approximation of the joint distribution of test statistics cannot be derived analytically, re-sampling methods, such as bootstrap and subsampling are applied directly over the observed gene expression data (Ramano and Wolf, 2007).

We think re-sampling residuals after modeling data appropriately, e.g. the bootstrap based on residuals method (P. 289, Shao and Tu, 1995), is a more appropriate procedure to estimate the dependence across test statistics. Such a resampling-based step-down procedure is constructed by generalized partitioning principle. It provides strong control of gFWER asymptotically and will be illustrated in a proof-of-concept Icelandic mice experiment. For a well-designed microarray experiment, we model the gene expression data appropriately by a gene-by-gene linear mixed effect model, and then the residuals after modeling are bootstrap re-sampled to estimate the joint distribution of test statistics. This method is compared with other existing gFWER-controlling procedures. It finds more truly differentially expressed genes than other methods.

3.2 Principles in Constructing GFWER-controlling Multiple Tests

We review general principles of constructing multiple testing procedures. The partition principle (Stefansson et al., 1988, Finner and Strassburger, 2002) is a general principle of
constructing multiple tests to control FWER strongly. It has been extended to the generalized partition principle (Xu and Hsu, 2007), which gives guidelines to construct multiple tests controlling gFWER strongly.

### 3.2.1 Partitioning Principle

To illustrate the partitioning principle, consider simultaneously testing hypotheses

\[ H_{0i} : \theta_i = 0, \ i = 1, \ldots, K. \]  

(3.2)

The partition testing proceeds as follows:

- **P1:** For each \( I \subseteq \{1, \ldots, K\} \), \( I \neq \emptyset \), form \( H^*_{0I} : \theta_i = 0 \) for all \( i \in I \) and \( \theta_j \neq 0 \) for \( j \notin I \). There are \( 2^K - 1 \) hypotheses to be tested.

- **P2:** Test each \( H^*_{0I} \) at level-\( \alpha \) so that

\[
\sup_{\theta, \Sigma} P_{\theta, \Sigma} (\text{reject } H^*_{0I}) \leq \alpha, \quad (3.3)
\]

where the supremum is taken over all possible \( \theta \) and \( \Sigma \) subject to \( \theta_i = 0 \) for \( i \in I \).

- **P3:** For each \( i \), infer \( \theta_i \neq 0 \) if and only if all \( H^*_{0I} \) with \( i \in I \) are rejected, because \( H_{0i} \) is the union of \( H^*_{0I} \) with \( i \in I \).

Since the null hypotheses \( H^*_{0I} \)'s are disjoint, at most one \( H^*_{0I} \) can be true. Therefore, there is no need for multiplicity adjustment among the \( H^*_{0I} \)'s for partition testing to control the FWER strongly.

**Example** Take \( K = 4 \) for example, then
P1 partitions the parameter space $\Theta = \{\theta_1 \in R, \theta_2 \in R, \theta_3 \in R, \theta_4 \in R\}$ into sixteen disjoint subspaces:

\[
\begin{align*}
\Theta^*_{\{1234\}} &= \{\theta_1 = 0 \text{ and } \theta_2 = 0 \text{ and } \theta_3 = 0 \text{ and } \theta_4 = 0\} \\
\Theta^*_{\{123\}} &= \{\theta_1 = 0 \text{ and } \theta_2 = 0 \text{ and } \theta_3 = 0 \text{ and } \theta_4 \neq 0\} \\
\Theta^*_{\{124\}} &= \{\theta_1 = 0 \text{ and } \theta_2 = 0 \text{ and } \theta_3 \neq 0 \text{ and } \theta_4 = 0\} \\
\Theta^*_{\{134\}} &= \{\theta_1 = 0 \text{ and } \theta_2 \neq 0 \text{ and } \theta_3 = 0 \text{ and } \theta_4 = 0\} \\
\Theta^*_{\{234\}} &= \{\theta_1 \neq 0 \text{ and } \theta_2 = 0 \text{ and } \theta_3 = 0 \text{ and } \theta_4 = 0\} \\
\Theta^*_{\{1\}} &= \{\theta_1 = 0 \text{ and } \theta_2 \neq 0 \text{ and } \theta_3 \neq 0 \text{ and } \theta_4 \neq 0\} \\
\Theta^*_{\{2\}} &= \{\theta_1 \neq 0 \text{ and } \theta_2 = 0 \text{ and } \theta_3 \neq 0 \text{ and } \theta_4 \neq 0\} \\
\Theta^*_{\{3\}} &= \{\theta_1 \neq 0 \text{ and } \theta_2 \neq 0 \text{ and } \theta_3 = 0 \text{ and } \theta_4 \neq 0\} \\
\Theta^*_{\{4\}} &= \{\theta_1 \neq 0 \text{ and } \theta_2 \neq 0 \text{ and } \theta_3 \neq 0 \text{ and } \theta_4 = 0\} \\
\Theta^*_\emptyset &= \{\theta_1 \neq 0 \text{ and } \theta_2 \neq 0 \text{ and } \theta_3 \neq 0 \text{ and } \theta_4 \neq 0\}.
\end{align*}
\]
P2 tests fifteen hypotheses

\[ H^*_0\{1234\} : \theta_1 = 0 \text{ and } \theta_2 = 0 \text{ and } \theta_3 = 0 \text{ and } \theta_4 = 0 \]
\[ H^*_0\{123\} : \theta_1 = 0 \text{ and } \theta_2 = 0 \text{ and } \theta_3 = 0 \text{ and } \theta_4 \neq 0 \]
\[ H^*_0\{124\} : \theta_1 = 0 \text{ and } \theta_2 = 0 \text{ and } \theta_3 \neq 0 \text{ and } \theta_4 = 0 \]
\[ H^*_0\{134\} : \theta_1 = 0 \text{ and } \theta_2 \neq 0 \text{ and } \theta_3 = 0 \text{ and } \theta_4 = 0 \]
\[ H^*_0\{234\} : \theta_1 \neq 0 \text{ and } \theta_2 = 0 \text{ and } \theta_3 = 0 \text{ and } \theta_4 = 0 \]
\[ \ldots \]
\[ H^*_0\{1\} : \theta_1 = 0 \text{ and } \theta_2 \neq 0 \text{ and } \theta_3 \neq 0 \text{ and } \theta_4 \neq 0 \]
\[ H^*_0\{2\} : \theta_1 \neq 0 \text{ and } \theta_2 = 0 \text{ and } \theta_3 \neq 0 \text{ and } \theta_4 \neq 0 \]
\[ H^*_0\{3\} : \theta_1 \neq 0 \text{ and } \theta_2 \neq 0 \text{ and } \theta_3 = 0 \text{ and } \theta_4 \neq 0 \]
\[ H^*_0\{4\} : \theta_1 \neq 0 \text{ and } \theta_2 \neq 0 \text{ and } \theta_3 \neq 0 \text{ and } \theta_4 = 0 \]

at level-\(\alpha\).

P3 infers \(\theta_i \neq 0\) if and only if all \(H^*_0\{I\}\) involving \(\theta_i = 0\) are rejected. For example, \(\theta_1 \neq 0\) is inferred if and only if all \(H^*_0\{1\}, \ H^*_0\{12\}, \ H^*_0\{13\}, \ H^*_0\{14\}, \ H^*_0\{123\}, \ H^*_0\{124\}, H^*_0\{134\}\) and \(H^*_0\{1234\}\) are rejected. \(\square\)

For simplicity, the so-called weak partitioning principle (Finner and Strassburger, 2002) tests the less restrictive hypotheses

33
\[ H_{0\{1234\}} : \theta_1 = 0 \text{ and } \theta_2 = 0 \text{ and } \theta_3 = 0 \text{ and } \theta_4 = 0 \]
\[ H_{0\{123\}} : \theta_1 = 0 \text{ and } \theta_2 = 0 \text{ and } \theta_3 = 0 \]
\[ H_{0\{124\}} : \theta_1 = 0 \text{ and } \theta_2 = 0 \text{ and } \theta_4 = 0 \]
\[ H_{0\{134\}} : \theta_1 = 0 \text{ and } \theta_3 = 0 \text{ and } \theta_4 = 0 \]
\[ H_{0\{234\}} : \theta_2 = 0 \text{ and } \theta_3 = 0 \text{ and } \theta_4 = 0 \]
\[ \ldots \]
\[ H_{0\{1\}} : \theta_1 = 0 \]
\[ H_{0\{2\}} : \theta_2 = 0 \]
\[ H_{0\{3\}} : \theta_3 = 0 \]
\[ H_{0\{4\}} : \theta_4 = 0, \]

which still guarantees strong control of FWER because a level-\(\alpha\) test for \(H_{0I}\) is also a level-\(\alpha\) test for \(H_{0I}^*\).

Note that \(H_{0\{1\}}\) is different from \(H_{01}\), since \(H_{0\{1\}}\) is a null hypothesis for \(\theta = (\theta_1, \theta_2, \theta_3, \theta_4) \in \mathbb{R}^4\) assuming \(\theta_1 = 0\) and \(H_{01}\) is a (marginal) individual null hypothesis for \(\theta_1 \in \mathbb{R}\) only assuming \(\theta_1 = 0\). This difference brings complication for testing \(H_{0\{1\}}\) since test statistic may be not only depending on \(\theta_1\) which is 0, but also on \(\theta_2 \in \mathbb{R}, \theta_3 \in \mathbb{R}\) and \(\theta_4 \in \mathbb{R}\), while when testing \(H_{01}\), the test statistic only depends on \(\theta_1 = 0\). In testing \(H_{0\{1\}}\), to control (3.3), we need to set the boundary of the rejection region at the least favorable configuration under \(H_{0\{1\}}\) with \(\theta_i \in \mathbb{R}\) for \(i \neq 1\) and nuisance parameters. This issue is often overlooked especially in constructing hypothesis testing by re-sampling technique.
Partitioning principle gives guidelines on how to construct multiple testing procedures to control FWER strongly, such procedures are not unique though, depending on how we choose a test statistic for $H_{0I}^*$ and calculate the critical value accordingly. Assume $T_i$ with $i = 1, 2, ..., K$ are test statistics for individual hypotheses $H_{0i}$ with $i = 1, 2, ..., K$ respectively. One appropriate way is to choose the test statistic for $H_{0I}^*$ to be of max $T$ type:

$$\text{if } \max_{i \in I} |T_i| > c_I, \text{ then reject } H_{0I}^*. $$

where $c_I$ satisfies

$$\sup_{\theta} P_{\theta}(\max_{i \in I} |T_i| > c_I) \leq \alpha. \quad (3.4)$$

The supremum is again taken over all possible $\theta$ (and $\Sigma$) subject to $\theta_i = 0$ for $i \in I$. Hereafter, the nuisance parameter $\Sigma$ is suppressed for brevity.

Assuming individual hypotheses $H_{0i}$ are roughly of equal interest, with the multiple test being of max $T$ type, one should guard against the rejection of $H_{0I}^*$ being mostly based on a few highly variable $T_i$. This would be the case if the test statistics $T_i$ have $t$-distributions and one of them has a very small degree of freedom, much smaller than the others, for example. To avoid such a situation, one can convert each $T_i$ to its $p$-value, and use the $p$-values as the test statistics. The test statistic for $H_{0I}^*$ will be of min $P$ type.

### 3.2.2 Generalized Partitioning Principle

The partitioning principle has been extended to a so-called generalized partitioning principle by Xu and Hsu (2007), which is a systematic approach to construct gFWER-controlling multiple testing procedure. To illustrate this principle, consider simultaneously testing (3.2) as well. The generalized partition testing proceeds as follows:
GP1: For each $I \subseteq \{1, \ldots, k\}, I \neq \emptyset$, form $H^*_{0I} : \theta_i = 0$ for all $i \in I$ and $\theta_j \neq 0$ for $j \notin I$. There are $2^k$ parameter subspaces and $2^k - 1$ hypotheses to be tested.

GP2: In each parameter space $\Theta^*_I$, reject all $H_{0i}, i \notin I$, and test $\{H_{0i} : \theta_i = 0, i \in I\}$ at gFWER level $\alpha$, controlling

$$\sup_{\theta} P_{\theta}(V_I > m) \leq \alpha, \quad (3.5)$$

where the supremum is taken over all possible $\theta$ subject to $\theta_i = 0$ for $i \in I$ and $V_I$ is the number of falsely declaring $\theta_i \neq 0$ for $i \in I$.

GP3: Reject $H_{0i}$ if $H_{0i}$ is rejected in $\Theta^*_I$ for all $I$ including $i$.

Note that the error rate (3.5) in multiple testing $\{H_{0i} : \theta_i = 0, i \in I\}$ within the parameter space $\Theta^*_I$ is different from the gFWER (3.1) in multiple testing $\{H_{0i} : \theta_i = 0, i \in \{1, 2, \ldots, K\}\}$. Recall that in (3.1),

$$gFWER = \max_{I_0 \subseteq \{1, 2, \ldots, K\}} \sup_{\theta} P_{\theta}\{V > m\}.$$

Since we do not know the true configuration of null hypotheses, we control the maximum of the probabilities of making more than $m$ mistakes over all possible configurations. While in this case, we know the true configuration, which is $\{\theta_i = 0 \text{ for } i \in I \text{ and } \theta_i \neq 0 \text{ for } i \notin I\}$.

The generalized partitioning principle first partitions the parameter space into disjoint sets, so that exactly one $\Theta^*_I$ contains the true parameter. It controls gFWER by controlling gFWER within each $\Theta^*_I$.

Likewise, the generalized partitioning principle gives guidelines on how to construct multiple testing procedures controlling gFWER strongly, such procedures are not unique as well, depending on how we choose a test statistic for multiple testing $\{H_{0i} : \theta_i = 0, i \in I\}$.
within $\Theta_I^*$ and calculate the critical value accordingly. One appropriate way is to choose the test statistic for the multiple testing to be of $(m+1)$th largest $T$ type. Suppose $\{T_i, i \in I\}$ is a collection of individual test statistics indexed by a finite set $I$ having $|I|$ elements, $T_{|I|-m}^I$ is used to denote the $(m+1)$th largest value of the $|T_i|$ with $i \in I$. Then, the $(m+1)$th largest $T$ test of $\{H_{0i} : \theta_i = 0, i \in I\}$ within $\Theta_I^*$ is: when $|I| > m$, always reject the $H_{0i}$'s corresponding to the first $m$ largest $T_i$ values ($i \in I$) within $\Theta_I^*$, and

$$
\text{if } T_{|I|-m}^I > c_I, \text{ reject all other } H_{0i}'s \ (i \in I) \text{ within } \Theta_I^*,
$$

where $c_I$ satisfies

$$
sup_{\theta \in \Theta_I^*} P(\theta | T_{|I|-m}^I > c_I) \leq \alpha.
$$

(3.6)

This procedure was also called as $I$-reject generalized partitioning test in Xu and Hsu (2007). Ramano and Wolf (2007) chose $(m+1)$th largest $T$ as a test statistic to construct their multiple tests controlling gFWER as well. Alternative test statistics can be used if desired.

If each $T_i$ is converted to its $p$-value, and $p$-values are used as test statistics, the multiple testing $\{H_{0i} : \theta_i = 0, i \in I\}$ within $\Theta_I^*$ is of $(m+1)$th smallest $P$ type.

### 3.3 Step-down Shortcuts

With $K$ individual hypotheses to be tested, the partitioning principle calls for $2^K - 1$ disjoint tests and the generalized partitioning principle calls for $\sum_{j=m+1}^{K} \binom{K}{j}$ disjoint tests. In microarray experiment, the number of hypotheses of interest $K$ is large, full implementation of partitioning testing and generalized partitioning testing is impossible.
Under some conditions, there are shortcuts to both principles which reduce the number of tests to at most $K - 1$ and $K - m$ respectively.

These shortcuts are in the form of step-wise, either step-down or step-up, tests. In step-down tests, the most possible significant null hypotheses (with the largest observed test statistics in a $maxT$ type partition testing) are considered successively, with further tests depending on the outcomes of earlier ones. Step-up tests process the other way around, the least possible significant null hypotheses (with the smallest observed test statistics) are considered successively, with further tests depending on the outcomes of earlier ones. Holm’s step-down test (Holm, 1979) and Hochberg’s step-up test (Hochberg, 1988) are popular step-wise multiple tests. As shown in Huang et al. (2007), Holm’s method and Hochberg’s method are both special cases of shortcuts of partition testing.

In this section, we focus on step-down tests. The step-down shortcuts for partition testing and generalized partition testing (Huang et al. 2007, Xu and Hsu, 2007) are reviewed. Step-down shortcut for partition principle exists if critical values are monotone, and step-down shortcut for generalized partition principle also exists if critical values are both monotone and exchangeable. We will prove that shortcuts for generalized partition testing do not work without exchangeability, and further propose a exchangeable-enforcement step-down procedure.

### 3.3.1 Sufficient Condition for Partitioning Step-down Tests

For $maxT$-type partition testing, if the critical values $c_I$ in (3.4) satisfy a monotone condition:

\[(M): \quad c_J \leq c_I, \text{ for } J \subset I,\]

38
the step-down test exists. Under monotonicity assumption, if \( H_{0,t}^* \) is rejected, then most of \( H_{0,J}^* \) with \( J \subset I \) can be rejected without actual testing.

**Example** Take \( K = 4 \) for example. If we observe \( T_1 < T_4 < T_3 < T_2 \), a step down test is executed as following steps:

**Step 1:** If

\[
T_2 = \max_{i \in \{1,2,3,4\}} T_i > c_{1,2,3,4},
\]

reject \( H_{02} \) and go to next step, otherwise stop.

In this step, \( H_{0(1234)}^* \) is rejected because of above testing. \( H_{0(124)}^* \), \( H_{0(123)}^* \), \( H_{0(234)}^* \), \( H_{0(12)}^* \), \( H_{0(23)}^* \) and \( H_{0(2)}^* \) are rejected without actually testing since

\[
T_2 = \max_{i \in \{1,2,3\}} T_i > c_{1,2,3} \quad \text{by} \quad \{1,2,3,4\} \supset \{1,2,3\} \quad \text{then} \quad c_{1,2,3,4} \geq c_{1,2,3}
\]

\[
T_2 = \max_{i \in \{1,2,3\}} T_i > c_{1,2,3} \quad \text{by} \quad \{1,2,3,4\} \supset \{1,2,3\} \quad \text{then} \quad c_{1,2,3,4} \geq c_{1,2,3}
\]

\[
T_2 = \max_{i \in \{2,3,4\}} T_i > c_{2,3,4} \quad \text{by} \quad \{1,2,3,4\} \supset \{2,3,4\} \quad \text{then} \quad c_{1,2,3,4} \geq c_{2,3,4}
\]

\[
T_2 = \max_{i \in \{1,2\}} T_i > c_{1,2} \quad \text{by} \quad \{1,2,3,4\} \supset \{1,2\} \quad \text{then} \quad c_{1,2,3,4} \geq c_{1,2}
\]

\[
T_2 = \max_{i \in \{2,3\}} T_i > c_{2,3} \quad \text{by} \quad \{1,2,3,4\} \supset \{2,3\} \quad \text{then} \quad c_{1,2,3,4} \geq c_{2,3}
\]

\[
T_2 = \max_{i \in \{2,4\}} T_i > c_{2,4} \quad \text{by} \quad \{1,2,3,4\} \supset \{2,4\} \quad \text{then} \quad c_{1,2,3,4} \geq c_{2,4}
\]

\[
T_2 = \max_{i \in \{2\}} T_i > c_{2} \quad \text{by} \quad \{1,2,3,4\} \supset \{2\} \quad \text{then} \quad c_{1,2,3,4} \geq c_{2}
\]

i.e. we can reject all \( H_{0(1234)}^* \), \( H_{0(124)}^* \), \( H_{0(123)}^* \), \( H_{0(234)}^* \), \( H_{0(12)}^* \), \( H_{0(23)}^* \), \( H_{0(24)}^* \) and \( H_{0(2)}^* \), then we can reject \( H_{02} \) since \( H_{02} \) is the union of the previous eight disjoint hypotheses.

**Step 2:** If

\[
T_3 = \max_{i \in \{1,3,4\}} T_i > c_{1,3,4},
\]

39
then reject $H_{03}$ and go to next step, otherwise stop.

In this step, $H_{0\{134\}}^*$ is rejected because of above testing. $H_{0\{13\}}^*$, $H_{0\{34\}}^*$ and $H_{0\{3\}}^*$ are rejected without actually testing, since

$$T_3 = \max_{i \in \{1,3\}} T_i > c_{\{1,3\}} \text{ by } \{1,3,4\} \supseteq \{1,3\} \text{ then } c_{\{1,3,4\}} \geq c_{\{1,3\}}$$

$$T_3 = \max_{i \in \{3,4\}} T_i > c_{\{3,4\}} \text{ by } \{1,3,4\} \supseteq \{3,4\} \text{ then } c_{\{1,3,4\}} \geq c_{\{3,4\}}$$

$$T_3 = \max_{i \in \{3\}} T_i > c_{\{3\}} \text{ by } \{1,3,4\} \supseteq \{3\} \text{ then } c_{\{1,3,4\}} \geq c_{\{3\}}.$$  

From Step 1, we’ve already rejected $H_{0\{1234\}}^*$, $H_{0\{123\}}^*$, $H_{0\{234\}}^*$, $H_{0\{23\}}^*$. We can thus reject $H_{03}$ since $H_{03}$ is the union of the previous eight disjoint hypotheses.

Step 3: If

$$T_4 = \max_{i \in \{1,4\}} T_i > c_{\{1,4\}},$$

then reject $H_{04}$ and go to next step, otherwise stop.

In this step, $H_{0\{14\}}^*$ is rejected because of above testing. $H_{0\{4\}}^*$ is rejected without actually testing, since

$$T_4 = \max_{i \in \{4\}} T_i > c_{\{4\}} \text{ by } \{1,4\} \supseteq \{4\} \text{ then } c_{\{1,4\}} \geq c_{\{4\}}.$$  

From previous two steps, we’ve already rejected $H_{0\{1234\}}^*$, $H_{0\{124\}}^*$, $H_{0\{134\}}^*$, $H_{0\{234\}}^*$, $H_{0\{24\}}^*$, $H_{0\{34\}}^*$. We can thus reject $H_{04}$ since $H_{04}$ is the union of the previous eight disjoint hypotheses.

Step 4: If

$$T_1 = \max_{i \in \{1\}} T_i > c_{\{1\}},$$

then reject $H_{01}$ and stop.
In this step, $H_{0\{1\}}^*$ is rejected because of above testing. From previous three steps, we’ve already rejected $H_{0\{1234\}}^*$, $H_{0\{123\}}^*$, $H_{0\{134\}}^*$, $H_{0\{124\}}^*$, $H_{0\{13\}}^*$, $H_{0\{14\}}^*$, $H_{0\{12\}}^*$, $H_{0\{14\}}^*$.

We can thus reject $H_{01}$ since $H_{01}$ is the union of the previous eight disjoint hypotheses.

Note that in Step 1, $T_2$ is the maximum for the complete set $\{1, 2, 3, 4\}$, it is also the maximum of any subset of the complete set. By the monotonicity of the critical values $c$, as long as $T_2 > c_{\{1,2,3,4\}}$, all the other 7 tests can be omitted. Also observe that in Step 2, $T_3$ is the maximum of the set $\{1, 3, 4\}$, which is the largest subset not including 2. For any subset $I$ with $3 \in I$, it has two cases: it does not include 2 or it does include 2. In the former case, we discuss similarly as in Step 1, $T_3$ is the maximum of any subset $I$ including 3 (but not including 2). By the monotonicity of the critical values $c$, as long as $T_3 > c_{\{1,3,4\}}$, all the 3 tests are omitted (for subset including 3 but not 2). In the latter case, for any subset including 2 (and also including 3), the test can be omitted, since in Step 1, $T_2 > c_{\{1,2,3,4\}}$, by the monotonicity of the critical values $c$, they’ve been rejected from previous step. □

Generally speaking, consider simultaneously testing (3.2). Assume monotonicity (M) of critical values, step-down shortcut works with at most $K$ disjoint tests. Let $[1], [2],..., [K]$ be the indices such that $T_{[1]} \leq T_{[2]} \leq ... \leq T_{[K]}$, and denote the critical values for the $K$ disjoint tests as $c_1, c_2,...c_K$ for brevity, where $c_j := c_I$ with $I = \{[j],[j-1],...,[1]\}$. The FWER-controlling $\max T$ step down test can now be executed as following:

Step 1: If $T_{[K]} > c_K$, reject $H_{0[K]}$ and go to next step, otherwise stop.

Step 2: If $T_{[K-1]} > c_{K-1}$, reject $H_{0[K-1]}$ and go to next step, otherwise stop.

:  

41
Step K: If $T_{[1]} > c_1$, reject $H_{[1]}$ and stop, otherwise stop.

A FWER-controlling $\min P$ step down test can be derived similarly, when $p$-values are used as test statistics.

### 3.3.2 Sufficient Condition for Generalized Partitioning Step-down Tests

For a $(m + 1)$th largest $T$ generalized partition testing, if the critical values $c_I$ in (3.6) satisfy both a monotonicity condition and an exchangeability condition:

\[(\text{GM}): \quad c_J \leq c_I, \text{ for } J \subset I,\]
\[(\text{GE}): \quad c_I = c_{I'}, \text{ if } \text{card}(I) = \text{card}(I').\]

the step-down test exists.

Give $i = 1, 2, ..., K$, generalized partitioning principle rejects $H_{0i}$ if and only if $H_{0i}$ is rejected in all $\Theta^*_{0I}$ with $I$ including $i$. According to the $(m + 1)$th largest $T$ generalized partition testing, all $H_{0i}$’s with $i \in I$ corresponding to the first $m$ observed largest test statistics are rejected in $\Theta^*_{0I}$ without testing, and all other $H_{0i}$’s with $i \in I$ are rejected in $\Theta^*_{0I}$, if $T^I_{[|I|-m]} > c_I$.

**Example** Take $K = 4$ and $m = 1$ as example, and assume observe $T_1 < T_4 < T_3 < T_2$ again. The step down test is executed as following steps:

Step 0: Reject $H_{02}$ and go to next step.

$H_{02}$ is rejected in $\Theta^*_{[2]}$ without testing. In $\Theta^*_{[2]}$, there is only one mistake to make, no matter what our decisions are, the number of false discoveries can not be greater than $m = 1$. 

42
In \( \Theta_I^* \) with \( I = \{1, 2\}, I = \{2, 3\}, I = \{2, 4\}, I = \{1, 2, 3\}, I = \{1, 2, 4\}, \)
\( I = \{2, 3, 4\}, I = \{1, 2, 3, 4\} \) respectively, \( H_{02} \) is always rejected without testing since it corresponds to (first) largest test statistic.

We can thus reject \( H_{02} \), since \( H_{02} \) was rejected in every \( \Theta_I^* \) with \( 2 \in I \).

Step 1: If

\[ T_3 = 2 - \max_{i \in \{1,2,3,4\}} T_i > c_{\{1,2,3,4\}}, \]

then reject \( H_{03} \) and go to next step, otherwise stop.

\( H_{03} \) (and \( H_{01}, H_{04} \)) are all rejected in \( \Theta_{\{1234\}}^* \) by above testing. \( H_{03} \) are rejected in
\( \Theta_{\{123\}}^*, \Theta_{\{234\}}^* \) and \( \Theta_{\{23\}}^* \) without testing, since

\[ T_3 = 2 - \max_{i \in \{1,2,3\}} T_i > c_{\{1,2,3\}} \quad \text{by} \quad \{1, 2, 3, 4\} \supset \{1, 2, 3\} \quad \text{then} \quad c_{\{1,2,3,4\}} \geq c_{\{1,2,3\}}, \]

\[ T_3 = 2 - \max_{i \in \{2,3,4\}} T_i > c_{\{2,3,4\}} \quad \text{by} \quad \{1, 2, 3, 4\} \supset \{2, 3, 4\} \quad \text{then} \quad c_{\{1,2,3,4\}} \geq c_{\{2,3,4\}}, \]

\[ T_3 = 2 - \max_{i \in \{2,3\}} T_i > c_{\{2,3\}} \quad \text{by} \quad \{1, 2, 3, 4\} \supset \{2, 3\} \quad \text{then} \quad c_{\{1,2,3,4\}} \geq c_{\{2,3\}}. \]

(In fact, \( H_{01} \) is rejected in \( \Theta_{\{123\}}^* \) without testing and \( H_{04} \) is rejected in \( \Theta_{\{234\}}^* \) without testing because of the above inequalities. )

In \( \Theta_{\{3\}}^* \), \( H_{03} \) is rejected without testing.

In \( \Theta_I^* \) with \( I = \{1, 3, 4\}, I = \{1, 3\}, I = \{3, 4\} \) respectively, \( H_{03} \) is always rejected without testing since it corresponds to the largest test statistic from these sets \( I \).

We can thus reject \( H_{03} \), since \( H_{03} \) was rejected in every \( \Theta_I^* \) with \( 3 \in I \). In this step, only monotonicity of the critical values is required.

Step 2: If

\[ T_4 = 2 - \max_{i \in \{1,3,4\}} T_i > c_{\{1,3,4\}}, \quad (3.7) \]
then reject $H_{04}$ and go to next step, otherwise stop.

$H_{04}$ (and $H_{01}$) are rejected in $\Theta^*_I$ by above testing. By the monotonicity condition for critical values, $H_{04}$ is rejected in $\Theta^*_I$ without testing since

$$T_4 = 2\max_{i \in \{3, 4\}} T_i > c_{\{3, 4\}} \text{ by } \{1, 3, 4\} \supset \{3, 4\} \text{ then } c_{\{1, 3, 4\}} \geq c_{\{3, 4\}}.$$  

Furthermore by exchangeability of critical values, $H_{04}$ (and $H_{01}$) are rejected in $\Theta^*_I$ without testing and $H_{04}$ are rejected in $\Theta^*_I$ without testing since

$$T_4 = 2\max_{i \in \{1, 2, 4\}} T_i > c_{\{1, 2, 4\}} \text{ by } \text{card}(\{1, 3, 4\}) = \text{card}(\{1, 2, 4\}) \text{ then } c_{\{1,3,4\}} = c_{\{1,2,4\}} \quad (3.8)$$

$$T_4 = 2\max_{i \in \{2, 4\}} T_i > c_{\{2, 4\}} \text{ by } \{1, 2, 4\} \supset \{2, 4\} \text{ then } c_{\{1,2,4\}} \geq c_{\{2,4\}}.$$  

In $\Theta^*_I$, $H_{04}$ is rejected without testing.

In $\Theta^*_I$, $H_{04}$ is rejected without testing since it corresponds to the largest observed test statistic from set $\{1, 4\}$. From Step1, $H_{04}$ is already rejected in $\Theta^*_I$ and $\Theta^*_I$.

We can reject $H_{04}$, since $H_{04}$ was rejected in all $\Theta^*_I$ with $4 \in I$. In this step, both monotonicity and exchangeability of critical values are necessary.

Step 3: If

$$T_1 = 2\max_{i \in \{1, 4\}} T_i > c_{\{1, 4\}}, \quad (3.9)$$

then reject $H_{01}$ and stop.
$H_{01}$ is rejected in $\Theta^*_\{14\}$ by above testing. By exchangeability of critical values, $H_{01}$ is rejected in $\Theta^*_\{12\}$ and in $\Theta^*_\{13\}$ respectively without testing since

\[ T_1 = 2 \cdot \max_{i \in \{1,2\}} T_i > c_{\{1,2\}} \text{ by } \text{card}(\{1,4\}) = \text{card}(\{1,2\}) \text{ then } c_{\{1,4\}} = c_{\{1,2\}} \]

(3.10)

\[ T_1 = 2 \cdot \max_{i \in \{1,3\}} T_i > c_{\{1,3\}} \text{ by } \text{card}(\{1,4\}) = \text{card}(\{1,3\}) \text{ then } c_{\{1,4\}} = c_{\{1,3\}}. \]

(3.11)

In $\Theta^*_\{1\}$, $H_{01}$ is rejected without testing.

From Step 1, $H_{01}$ is rejected in $\Theta^*_\{1234\}$ and $\Theta^*_\{123\}$ respectively. From Step 2, $H_{01}$ is rejected in $\Theta^*_\{134\}$ and in $\Theta^*_\{124\}$ respectively.

We can reject $H_{01}$, since $H_{01}$ was rejected in all $\Theta^*_I$ with $1 \in I$.

Note that in Step 1, $T_3$ is the second largest observed test statistic for the complete set $\{1,2,3,4\}$. In any subset $I$ with $3 \in I$ and $2 \in I$, $H_{03}$ always corresponds to the second largest test statistic. By monotonicity of critical values $c$, as long as $T_3 > c_{\{1,2,3,4\}}$, all $H_{03}$ are rejected in $\Theta^*_I$ with $3 \in I$ and $2 \in I$ without testing since all such $I$ are subset of the complete set. For any subset $I$ containing $3$ but not $2$, $H_{03}$ are rejected in $\Theta^*_I$ without testing since $H_{03}$ corresponds to the largest test statistic.

We can now show that without the exchangeability assumption on critical values, the short-cut doesn’t work in Step 2 and Step 3. In Step 2, $T_4$ is the second largest observed test statistic for $\{1,3,4\}$, which is the largest set not including $2$. For any subset $I$ including $4$ and $3$ but not $2$, by monotonicity of critical values, as long as $T_4 > c_{\{1,3,4\}}$, $H_{04}$ is rejected in $\Theta^*_I$ without testing since such $I$ are subsets of $\{1,3,4\}$. While for any subset $I$ including $4$ and $2$ but not $3$, testing of $H_{04}$ in $\Theta^*_I$ can not be skipped without exchangeability of critical values since such $I$ are not subset of $\{1,3,4\}$. $\square$
Generally, consider the multiple testing problem (3.2) allowing at most \( m \) false discoveries. Assume monotonicity (GM) and exchangeability (GE) of critical values, \( c_I \) only depend on cardinality of the set \( I \), \( c_I = c_{|I|} \), the step-down shortcut works with at most \( K - m \) tests. The gFWER-controlling \((m + 1)\)th largest \( T \) step down test is executed as following:

---

**Step 0:** Reject \( H_{0[K]}, H_{0[K−1]}, \ldots, H_{0[K−m+1]} \).

**Step 1:** If \( T_{[K−m]} > c_{K−m} \), reject \( H_{0[K−m]} \) and go to next step, otherwise stop.

**Step 2:** If \( T_{[K−m−1]} > c_{K−m−1} \), reject \( H_{0[K−m−1]} \) and go to next step, otherwise stop.

:  

**Step K-m:** If \( T_{[1]} > c_{1} \), reject \( H_{[1]} \) and stop, otherwise stop.

---

When \( p \)-values are used as test statistics, a gFWER-controlling \((m + 1)\)th smallest \( P \) step down test can be derived as well. Both \((m + 1)\)th largest \( T \) and \((m + 1)\)th smallest \( P \) step down tests control gFWER strongly as proved in Xu and Hsu (2007).

### 3.3.3 Enforced Exchangeability Step-down Tests

Although shortcuts do not exist for generalized partitioning principle, when exchangeability (GE) assumption of critical values \( c_I \) is not held. Above discussion gives us an idea to modify the critical values for each test \( H_0^{*I} \) so that the new critical values are exchangeable. In this case, the shortcuts reduce number of generalized partition testing from \( \sum_{j=m+1}^{K} \binom{K}{j} \) to at most \( \sum_{l=m}^{K-1} \binom{l}{m} \) and can still be executed as step-down tests.
Let $c_I$ represent the critical values in (3.6) and $d_I$ represent new critical values modified from $c_I$. Assume $d_I$ are exchangeable, then $d_I = d_{\{I\}}$. In a step-down test, let $R_j$ represent indices corresponding to individual hypotheses that have been rejected in previous $j - 1$ steps and $A_j$ represent complement of $R_j$, that is, indices corresponding to individual hypotheses that have not yet been rejected in previous $j - 1$ steps. $d_{K - m + j}$ in the $j$th step is then set to be

$$
 d_{K - m + j + 1} := \max_{i_1, i_2, \ldots, i_m} \{ c_J : J = A_j \cup \{ i_1, i_2, \ldots, i_m \} \}.
$$

(3.12)

**Example** Take $K = 4$ and $m = 1$ as example, and assume observe $T_1 < T_4 < T_3 < T_2$.

The gFWER-controlling step-down test is now executed as in following steps:

**Step 0:** reject $H_{02}(H_{0[4]})$.

**Step 1:** Now $R_1 = \{ 2 \}$, $A_1 = \{ 1, 3, 4 \}$ and set $d_3 := c_{\{1,2,3,4\}}$. If

$$
 T_3 = 2\cdot \max_{i \in \{1,2,3,4\}} > d_3,
$$

then reject $H_{03}(H_{0[3]})$ and go to next step, otherwise stop.

**Step 2:** $R_2 = \{ 2, 3 \}$, $A_2 = \{ 1, 4 \}$ and set $d_2 := \max(c_{\{1,2,4\}}, c_{\{1,3,4\}})$ as in (3.12). If

$$
 T_4 = 2\cdot \max_{i \in \{1,3,4\}} > d_2,
$$

then reject $H_{04}$ and go to next step, otherwise stop.

In this step, both inequalities (3.7) and (3.8) in previous example are hold since $d_2 \geq c_{\{1,3,4\}}$ and $d_2 \geq c_{\{1,2,4\}}$.

**Step 3:** $R_3 = \{ 2, 3, 4 \}$, $A_3 = \{ 1 \}$ and $d_1 = \max(c_{\{1,2\}}, c_{\{1,3\}}, c_{\{1,4\}})$. If

$$
 T_1 = 2\cdot \max_{i \in \{1,4\}} T_i > d_1,
$$

47
then reject $H_{01}$ and stop.

In this step, inequalities (3.9)-(3.11) in previous example are all satisfied since $d_1 \geq c_{\{1,4\}}$, $d_1 \geq c_{\{1,2\}}$ and $d_1 \geq c_{\{1,3\}}$. □

We have shown that, without exchangeability (GM) assumption on critical values $c_I$, the generalized partition testing can still be executed as if it was a step-down test with at most $K - m$ steps after some modification on critical values. However, at most $\sum_{l=m}^{K-1} \binom{l}{m}$ hypotheses are actually tested.

Although $\sum_{l=m}^{K-1} \binom{l}{m}$ is far less than $\sum_{j=m+1}^{K} \binom{K}{j}$, it is still computationally infeasible since it becomes very large for large values of $K$ and $m > 1$. For example, as $K = 99$ and $m = 5$, $\sum_{j=m+1}^{K} \binom{K}{j} \approx 6 \times 10^{29}$ and $\sum_{l=m}^{K-1} \binom{l}{m} \approx 10^9$. From now on, we call such a shortcut as an enforced exchangeability step down test, which is a quasi step down test.

Generally, consider simultaneously testing (3.2) again. Assume monotonicity (GM) only for critical values $c_I$ in (3.6). If we modify the critical values as in (3.12), the $(m+1)$th largest $T$ enforced exchangeability step down test is executed as following:

**Step 0**: Reject $H_{0[K]}$, $H_{0[K-1]}$, ..., $H_{0[K-m+1]}$.

**Step 1**: $R_1 = \{[K], [K-1], ..., [K-m+1]\}$, $A_1 = \{[K-m], [K-m-1], ..., [1]\}$ and $d_{K-m} = c_{A_1 \cup R_1}$. If $T_{[K-m]} > d_{K-m}$, reject $H_{0[K-m]}$ and go to next step, otherwise stop.

**Step 2**: $R_2 = \{[K], [K-1], ..., [K-m+1], [K-m]\}$, $A_2 = \{[K-m-1], ..., [1]\}$ and $d_{K-m-1} = \max_{i \in R_2} (c_{A_2 \cup (R_2 - \{i\})})$. If $T_{[K-m-1]} > d_{K-m-1}$, reject $H_{0[K-m-1]}$ and go to next step, otherwise stop.
Step K-m: $R_{K-m} = \{1, 2, ..., K\} - \{[1]\}$, $A_{K-m} = \{[1]\}$ and

$$d_1 = \max_{\{i_1, i_2, ..., i_m\} \in R_2} (c([1]) \cup \{i_1, i_2, ..., i_m\}).$$

If $T_{[1]} > d_1$, reject $H_{[1]}$ and stop, otherwise stop.

If the critical values $c_I$ are exchangeable but not monotone, $c_I = c_{|I|}$, then any $c_j$ can be modified by

$$c_{j+1} = \max(c_j, c_{j+1}),$$

so that they are monotone.

Similarly, a $(m + 1)$th smallest $P$ enforced exchangeability step down test can be illustrated as well, when $p$-values are used as test statistics. Both $(m + 1)$th largest $T$ and $(m + 1)$th smallest $P$ enforced exchangeability step down tests control gFWER strongly by generalized partitioning principle and its sufficient conditions for shortcuts.

### 3.4 Re-sampling Based Multiple Tests Controlling GFWER

To construct more powerful multiple tests than aforementioned inequality-based procedures, calculation of critical values for each $H_{0I}$ needs take dependence across individual test statistics into account. When the joint distribution of the test statistics cannot be derived analytically, re-sampling methods, such as bootstrap (Romano and Wolf, 2007) and permutation, are applied to estimate the dependence.
We suggest applying bootstrap based on residuals (abbreviated as RB) re-sampling approach (P289, Shao and Tu, 1995) to estimate the dependence across different test statistics. Romano and Wolf (2007) re-sample the observed data by bootstrap directly without any model assumption. They further assume a consistency of distribution estimator from the bootstrap, to prove that their resampling-based multiple tests control gFWER asymptotically. In our case, the consistency of distribution estimator from RB does hold. If we re-sample residuals after fitting data by an appropriate linear model, and construct a re-sampling based multiple testing procedure accordingly by following the generalized partitioning principles, we can then prove the asymptotic control of our approach.

Figure 3.1 shows steps of computing critical values by RB re-sampling approach. They are discussed with details in Section 3.4.1 and 3.4.2 for microarray experiment. Our resampling based multiple testing procedure is then described in Section (3.4.3) followed by a formal proof of its asymptotic control of gFWER.

**Permutation tests may not control error rates** Permuting raw data across groups to be compared is often used as a tool to generate reference distributions under the null hypotheses to be tested. It is thought of as capable of taking dependence into account, as well as producing more exact inferences for small samples than methods based on asymptotic properties.

It turns out that, depending on applications, there may be an assumption needed for permutation tests to control multiple testing error rates. At issue is, in comparing parameters of the marginal distributions of two sets of multivariate observations, the validity of permutation testing is affected by all the parameters in the joint distributions of the observations.

Calian, Li, and Hsu (2008) showed the surprising fact that, in the case of a linear model with i.i.d. errors, permuting raw data across groups to be compared turns out to control the
FWER, if the test statistic for each $H_0^*$ is based on ordinary least squares estimates and of $maxT$-type.

On the other hand, as shown in Xu and Hsu (2007), in comparing the mean expression levels of genes between two groups of subjects, permuting raw data across groups may not generate the correct reference distributions under the null hypotheses, unless equalities of mean expression levels for a set of genes automatically imply equality of joint distributions of expressions levels for this set of genes (including equalities of variances, covariances, and higher cumulants). Our view is, unless such an assumption can be made on biological grounds, it is safer to take a modeling and then re-sampling the residuals approach.

Figure 3.1: Flow chart of computing critical values by modeling the data and estimating dependence by re-sampling.
3.4.1 Linear Model for Microarray Data

Suppose that a microarray experiment is conducted properly, adhering to the statistical design principles and addressing the issues on different sources of variability in gene expression measurements discussed in Chapter 1. We can analyze the microarray data by proper modeling. A concrete example of such a design and experiment was given in Section 1.3.

Let $y_{igmspr}$ denote the background corrected, log transformed and normalized probe intensity for the $i$th gene ($i = 1, \ldots, n_i$), $s$th sample ($s = 1, \ldots, n_{s(m,g)}$) from the $m$th subject ($m = 1, \ldots, n_{m(g)}$) in group $g$ ($g = 1, \ldots, n_g$), $p$th probe ($p = 1, \ldots, n_{p(i)}$), and $r$th replicate ($r = 1, \ldots, n_r$) from the experiment.

We assume, for each $i$, $y_{igmspr}$ follows a linear effect model:

$$
Y_{igmspr} = \mu_i + \tau_{ig} + M_{im(g)} + S_{is(m(g))} + \pi_{p(i)} + \epsilon_{igmspr},
$$

(3.13)

where

- $\mu_i =$ mean gene expression for $i$th gene,
- $\tau_{ig} =$ group $g$ effect on $i$th gene,
- $M_{im(g)} =$ effect of subject $m$ in group $g$ on $i$th gene,
- $S_{is(m(g))} =$ effect of $s$th sample from $m$th subject in group $g$ on $i$th gene,
- $\pi_{p(i)} =$ effect of $p$th probe on $i$th gene,
- $\epsilon_{igmspr} =$ measurement error.

Replicate measurement errors are assumed to be independent, identically distributed with variance $\sigma_{\epsilon}^2$. If they are normally distributed, then estimated group differences have a
multivariate normal distribution, from which multiple tests can be derived. If they are not normally distributed, then multiple tests based on re-sampling of the residuals can be used.

Subject and sample effects can be considered as either fixed or random depending on the purpose of the microarray analysis. If the subject and sample effects are random, their variances are assumed to be $\sigma_{iM}^2$ and $\sigma_{iS}^2$ respectively. In this case, \( \text{var}(Y_{igm.spr}) = \sigma_{ie}^2 + \sigma_{iM}^2 + \sigma_{iS}^2 \), the covariance between gene expressions for different replicates from the same sample is $\sigma_{iM}^2 + \sigma_{iS}^2$ and the covariance between gene expressions for different samples from the same subject is $\sigma_{iM}^2$.

Model (3.13) is a marginal model in the sense that it does not specify the joint distribution of $y_{igm.spr}$ across the genes. Modelling expression levels from all genes simultaneously with subject by gene interaction included would generally require too much computer memory.

Other models exist for gene expression data from cDNA spotted arrays and stock Affymetrix arrays (e.g., Wolfinger et al., 2001, Chu et al., 2002, 2004, Smyth, 2004, Lee et al., 2002). However, the model (3.13) differs significantly from the existing models in that the design underlying it allows for the separate estimation of the subject and sample effects.

Consider, for example, the model Chu et al. (2004) used to analyze the ionizing radiation data in Tusher, Tibshirani, and Chu (2001). Each of two samples from each of four groups (treatment and cell line combinations with two levels each) was hybridized to a stock Affymetrix array. On a gene-by-gene basis, their linear mixed effects model was:

$$ Y_{igpl} = \mu + \tau_{ig} + \pi_{p(i)} + A_{l(g)} + \epsilon_{igpl} $$

(3.14)
where $Y_{igp}$ is the log transformed perfect match values for probe-level data from the $i$th gene, $\mu_i$ is the mean gene expression for the $i$th gene, $\tau_{ig}$ is the $g$th group (treatment and cell line combination) effect on the $i$th gene, $\pi_{p(i)}$ is the $p$th probe effect in the $i$th gene, $A_{l(g)}$ is a random array effect, and $\epsilon_{igpl}$ is measurement error.

Comparing model (3.14) with model (3.13), we see the array effect $A_{l(g)}$ in model (3.14) includes both the subject effect $M_{im(g)}$ and sample effect $S_{is(m(g))}$ in model (3.13). That is, with stock Affymetrix arrays, sample and subject effects are completely confounded with array effects, and cannot be estimated. Therefore, treatment or group effect is confounded with array effect in this case.

Smyth (2004) applied the following simple linear fixed effect model to fit log-transformed intensities. For the $i$th gene, assume

$$E(Y_i) = X\alpha_i \quad \text{(3.15)}$$
$$\text{var}(Y_i) = W_i \sigma_i^2, \quad \text{(3.16)}$$

where $Y_i$ is a vector containing all the transformed intensities from different samples for the $i$th gene, $X$ is the design matrix and $\alpha_i$ is a vector containing all the parameters $\mu$, $\tau$ and $\pi$ for the $i$th gene. $W_i$ is assumed to be a known non-negative definite matrix. It is not entirely clear how to set values for the matrix $W_i$ in real applications.

Models (3.15) and (3.13) have difference in the variance matrix of the vector $Y_i$. In particular, the variances of the elements of $Y_i$ in the model (3.15) are $\sigma_i^2$ times the diagonals of the matrix $W_i$, while they are given by $\sigma_{ii}^2 + \sigma_{iM}^2 + \sigma_{iS}^2$ (a combination of separate variance components from subject effects, sample effects, and measurement error) in the model (3.13). Also, the covariances among gene expression levels are modelled differently in (3.13) and (3.15). Model (3.13) specifies the covariance between gene expression levels
of different samples from the same subject to be $\sigma_{iM}^2$ and the covariance between gene expression levels of the same sample to be $\sigma_{iM}^2 + \sigma_{iS}^2$. In other words, (3.13) models the additional covariance due to measurements being from the same sample additively. Model (3.15), on the other hand, assumes both of the covariance between gene expression levels from the same subject and that from the same sample to be multiples of $\sigma_i^2$. These multiples are to be specified in the matrix $W_i$. However, proper specification of the multiples does not seem to be straightforward, especially when the sample and subject effects are additive.

In order to borrow information from the ensemble of genes to assist in estimation of variance of each individual gene, Smyth (2004) assumes a prior distribution on $\sigma_i^2$,

$$\frac{1}{\sigma_i^2} \sim \frac{1}{d_0s_0^2} \chi^2 d_0.$$  

With this prior specification, (3.15) is not a gene-by-gene model any more. The unknown variance $\sigma_i^2$ in model (3.15) can then be estimated by the posterior mean of $\sigma_i^2$ given $s_i^2$, i.e.

$$\hat{\sigma}_i^2 = \frac{d_0}{d_0 + d_i} s_0^2 + \frac{d_i}{d_0 + d_i} s_i^2.$$  

The estimate $\hat{\sigma}_i^2$ shrinks the observed variances $s_i^2$ towards the prior values $s_0^2$ with the extent of shrinkage determined by the relative sizes of the observed and the prior degrees of freedom $d_i$ and $d_0$. This is similar in principle to Tusher, Tibshirani, and Chu (2001)’s idea in SAM of modifying the estimate of $\sigma_i$ by an offset parameter, i.e. $\hat{\sigma}_i = s_i + s_0$. Smyth’s offset estimate is motivated by a hierarchical model, whereas $s_0$ in SAM is empirically chosen to be a particular percentile of all $s_i$ values without a model or an associated distribution theory.
The two-stage ANOVA model in Lee et al. (2002) is basically the same as model (3.14) except that all effects are assumed to be fixed. They also adjust the mean square error by an offset quantity in testing for significance using $F$-statistics.

### 3.4.2 Estimating Dependence by RB Re-sampling

After fitting data by the model (3.13), we can now describe the algorithm of estimating dependence by RB re-sampling to find differentially expressed genes between a mutant group (say, group M) and a control group (C), when $T_i$, $i = 1, 2, ..., n$, are used as the test statistics.

**Step 0:** Calculate the test statistic of the comparison between two groups. The observed test statistic $t_i = \hat{\tau}_{im} - \hat{\tau}_{ic}$ for the $i$th gene is the difference of the (weighted) average intensities, computed as follows. First average the (background corrected, externally-normalized) probe intensities within each subject. Then these average intensities are averaged over the subjects within each group. Finally, $\hat{\tau}_{im} - \hat{\tau}_{ic}$ is computed as the difference of the group average intensities.

**Step 1:** To build a reference distribution of the test statistics, we fit the data to model (3.13) with both subject and sample effects as fixed, to obtain the residuals $\hat{\epsilon}_{igmprs}$.

**Step 2:** Re-sample independently with replacement the residual vectors, pooled across groups $M$ and group $C$. Note that, to account for the potential dependence among the measurement errors across genes, the residuals are re-sampled vectors at a time, with each vector consisting of residuals from within each mini-microarray. For each re-sampled data set, compute the test statistic

$$T^b_i = \hat{\tau}_{im} - \hat{\tau}_{ic} = \bar{\epsilon}_{im} - \bar{\epsilon}_{ic}$$

(3.17)
of each gene.

Step 3: Repeat Step 2 sufficiently many times (say \( B \) times) and store the test statistic \( T_i^b \) with \( i = 1, 2, \ldots, n_i \) and \( b = 1, 2, \ldots, B \) as a \( n_i \times B \) matrix, denoted as \( T \). We have an estimated reference distribution for the test statistic \( \hat{\tau}_{iM} - \hat{\tau}_{iC} \) of each gene.

### 3.4.3 RB Re-sampling Based Step-down Tests

After obtaining the estimated reference distribution (3.17) of \( T_i \)'s across all genes by re-sampling residuals after modeling, we are now in the position of describing a re-sampling version of the gFWER-controlling \((m + 1)\)th largest \( T \) step down test described in Section 3.3.2. Hereafter, we call this re-sampling based method as RB-based \((m + 1)\)th largest \( T \) step down test. This test controls gFWER asymptotically without exchangeability assumption, its proof is delayed until next section.

Before giving the formal description of the RB-based \((m+1)\)th largest \( T \) step down test, we need clarify two commonly used terms, raw \( p \)-value and adjusted \( p \)-value, in multiple testing problems.

**Raw \( p \)-value versus adjusted \( p \)-value** Given \( i \), for a single hypothesis \( H_i \), one can reject the hypothesis by either \( T_i \) greater than a critical value or the corresponding \( p \)-value, \( p_i \) less than \( \alpha \). The \( p \)-value for a single test is defined as the smallest nominal Type I error level of the single hypothesis testing procedure at which one would reject the single hypothesis. The notion of \( p \)-value extends directly to multiple testing problems. The adjusted \( p \)-value (P27, Dudoit and van der Laan, 2008) for null hypothesis \( H_i \) in multiple testing problem (3.2) is the smallest nominal gFWER level of the multiple testing procedure at which one would rejected \( H_i \). To differentiate them, the aforementioned \( p \)-values are the so-called raw \( p \)-value from now on.
Now, we can describe the so-called RB-based \((m+1)\)th largest \(T\) step down test in terms of adjusted \(p\)-values. Let \([1], [2], \ldots, [n_i]\) be the indices such that \(T[1] \leq T[2] \leq \ldots \leq T[n_i]\). Without monotonicity nor exchangeability assumption on critical values \(c_I\) in (3.6), the test is executed as follows:

**Step 0:** reject null hypotheses \(H_0[n_i], H_0[n_i-1], \ldots, H_0[n_i-m+1]\).

**Step 1:** The observed test statistics \(t_i\) are sorted in decreasing order, denoted as \(t[n_i], t[n_i-1], \ldots, t[1]\), and the matrix \(T\) is sorted accordingly.

**Step 2:** To test hypothesis \(H_{[n_i-m]}\) corresponding to \(t[n_i-m]\), estimate the adjusted \(p\)-value by

\[
\text{adj}_p[n_i-m] = \frac{\# \{b : |T^b_{[n_i-m]}| \geq t[n_i-m]\}}{B},
\]

where \(|T^b_{[n_i-m]}|\) is the \((m+1)\)th largest absolute value for the \(b\)th column in matrix \(T\). If \(\text{adj}_p[n_i-m] < \alpha\), reject \(H_{[n_i-m]}\) and go to Step 3; otherwise stop.

**Step 3:** Remove the first row of the matrix \(T\), still denote it as matrix \(T\). Repeat Step 2 to get \(\text{adj}_p[n_i-m-1]\). If \(\text{adj}_p[n_i-m-1] < \alpha\), reject \(H_{[n_i-m-1]}\) and go to Step 4; otherwise stop.

If monotonicity condition is not satisfied, modify \(\text{adj}_p[n_i-m-1]\) by

\[
\text{adj}_p[n_i-m-1] = \max(\text{adj}_p[n_i-m], \text{adj}_p[n_i-m-1]),
\]

then compare it with \(\alpha\).

**Step 4** Repeat Step 3 until the adjusted \(p\)-value greater than \(\alpha\).
Note that the matrix $T$ was sorted according to the decreasing order of observed values of test statistics, $t_i$, and the first row of the matrix $T$ was removed in Step 3 in above algorithm. We need to compute the adjusted $p$-value based on the reference distribution of the test statistic $T_{n_i-m}$ for $H_{0I}^*$ with $I = \{[n_i], [n_i - 1], ..., [2]\}$ by the generalized partitioning principle.

As $m = 0$, it gives the RB-based $\text{max}_T$ step down test to control FWER asymptotically. We can derive the RB-based $(m + 1)$th largest $T$ enforced exchangeability step down test as well, and the details for this test are omitted for brevity.

As raw $p$-values are served as test statistics, RB versions of $\text{min}_P$ step down test, $(m + 1)$th smallest $P$ step down test and $(m + 1)$th smallest $P$ enforced exchangeability step down test can all be derived accordingly.

### 3.4.4 Asymptotic Control of GFWER

The main purpose of this section is to show that the re-sampling based step down procedure we constructed in Sections 3.4.1-3.4.3 controls gFWER asymptotically as sample size $n$ (number of subject per group $n_{m(g)}$ and number of replicated samples per subject $n_{s(m,g)}$) goes to infinity,

$$\limsup_{n \to \infty} P(\text{m or more false rejections}) \leq \alpha.$$

For simplicity, let’s focus on the RB versions of $(m + 1)$th largest $T$ step down test and $(m + 1)$th largest $T$ enforced exchangeability step down test only. The exact versions of both tests are described in section 3.3, where the $(m + 1)$th largest $T$ enforced exchangeability step down test controls gFWER strongly as critical values $c_I$ in (3.6) are monotone and $(m + 1)$th largest $T$ step down test controls gFWER strongly as critical values $c_I$ in (3.6) are
both monotone and exchangeable. Under linear model assumption, the re-sampling technique we used here are the so-called bootstrap based on residuals approach. The asymptotic strong control on gFWER of these procedures are essentially based on the consistency of the distribution estimator based on the RB re-sampling approach (P320, Shao and Tu, 1995) and generalized partitioning principle.

To state and prove the above statement strictly, we need specify some notation in advance. Let $Q$ denote the true distribution generating the data, $Q = (Q_1, Q_2, ..., Q_n)^\top$, where

$$Q_i = X\beta_i + \epsilon_i$$

(3.18)

is fitted as a linear model in (3.13). In these models, subject and sample effects are fixed and replicate measurement errors are i.i.d across different subjects, with mean zero and unknown variances for each gene $i, i = 1, 2, ..., n_i$. Let $\hat{\beta}_{n,i}$ denote the OLS estimator of $\beta_i$ in the model, given observed data with sample size $n$. $T_{n,i} = h'\hat{\beta}_i$ denotes the test statistics for each individual hypothesis $H_0$ and $T^*_{n,i}$ is the bootstrap replication (Efron, 1993) of $T_{n,i}$ based on residuals. $T_{n,i} = \hat{\tau}_{iM} - \hat{\tau}_{iC}$ for gene selection in Section 3.4.2.

For $I \subseteq \{1, 2, ..., K\}$, let $J_{n,I}(x, Q)$ denote the joint distribution of $\{\sqrt{n}T_{n,i} : i \in I\}$ under the true data generating distribution $Q$, and $J_{n,I}(x, boot)$ be its bootstrap estimator, which is the RB bootstrap joint distribution of $\{\sqrt{n}T^*_{n,i} : i \in I\}$ under the empirical estimation of $Q$. By extending the result in Shao and Tu (1995), we have,

**Proposition 1.** Under the above model assumption in (3.18),

$$||J_{n,I}(Q) - J_{n,I}(boot)||_{L_\infty} \rightarrow 0 \text{ as } n \rightarrow \infty,$$

(3.19)

where $||h||_\infty$ is the sup-norm of a function $h$ on $R^{|I|}$, with $||h||_\infty = \sup_x|h(x)|$. 

60
Let $F_n,I(x, Q)$ denote the cumulative distribution of $\sqrt{n}T_{|I| - m}$ under the true data generating distribution $Q$, and $F_n,I(x, \text{boot})$ be its RB bootstrap estimator, then

**Proposition 2.** Under the above model assumption in (3.18),

$$||F_n,I(Q) - F_n,I(\text{boot})||_{L_\infty} \to 0 \text{ as } n \to \infty.$$  \hspace{1cm} (3.20)

In fact, as sample size goes to infinity, the joint distribution of $\{\sqrt{n}T_{n,i} : i \in I\}$, $J_{n,I}(Q)$, has a limiting distribution, which is multivariate normal. Since $T_{|I| - m}$ is a continuous function on $T_{n,i}$, (3.20) is true by (3.19) and continuous mapping theorem.

Let $c_{n,I}$ denote the critical values in (3.6) for the test statistics $T_{|I| - m}$ and $\hat{c}_{n,I}$ its RB bootstrap estimator as described in Section 3.4.2, then $c_{n,I} = F_{n,I}^{-1}(\alpha, Q)/\sqrt{n} = \inf\{x, F_{n,I}(x, Q) \geq \alpha\}/\sqrt{n}$ and $\hat{c}_{n,I}$ is the empirical estimate of $F_{n,I}^{-1}(\alpha, \text{boot})/\sqrt{n}$ and

$$\lim_{n \to \infty} |\hat{c}_{n,I} - F_{n,I}^{-1}(\alpha, \text{boot})/\sqrt{n}| \to 0.$$

The RB-based $(m + 1)$th largest $T$ step down test and RB-based $(m + 1)$th largest $T$ enforced exchangeability step down test are the tests described in Section 3.3 with $c_{n,I}$ replaced by $\hat{c}_{n,I}$ and with an extra step to enforce monotonicity.

$$\hat{c}_{n,I} = \max(\hat{c}_{n,I}, \hat{c}_{n,J}) \text{ for } I \supseteq J.$$  \hspace{1cm} (3.21)

We can now prove the asymptotic control of gFWER of both procedures.

**Theorem 3.** Under the above model assumption in (3.18), without monotonicity assumption on $c_{n,I}$, the RB-based $(m + 1)$th largest $T$ enforced exchangeability step down test controls the gFWER asymptotically,

$$\limsup_{n \to \infty} P(\text{m or more false rejections}) \leq \alpha.$$
Proof. Within the parameter space $\Theta^*_I$, by consistency of $F_{n,I}(Q)$ (3.20) and law of large numbers,

$$\limsup_{n \to \infty} P(T^l_{|I| - m} > c_{n,I}) = \limsup_{n \to \infty} \hat{P}(T^l_{|I| - m} > \hat{c}_{n,I}) \leq \alpha,$$

(3.22)

where $\hat{P}(T^l_{|I| - m} > \hat{c}_{n,I})$ is the empirical estimate of $P(T^l_{|I| - m} > \hat{c}_{n,I})$.

From (3.22), gFWER is controlled asymptotically at level $\alpha$ within each configuration of multiple null hypotheses, gFWER is then controlled asymptotically for the procedure by the generalized partitioning principle. The shortcuts works since $\hat{c}_{n,I}$ is monotone from (3.21). \qed

**Theorem 4.** Under the above model assumption in (3.18), without monotonicity or exchangeability assumption on $c_{n,I}$, the RB-based $(m+1)$th largest $T$ step down test controls the gFWER asymptotically,

$$\limsup_{n \to \infty} P(m \text{ or more false rejections}) \leq \alpha.$$

Proof. For fix sample size $n$, $c_{n,I}$ is not required to be exchangeable. For any set with the same cardinality, the critical values in this procedure were calculated using the set $I$ containing indices corresponding to the least possible significant hypotheses, compared with other sets with the same cardinality $|I|$, denote it as $c_{n,|I|}$ and $\hat{c}_{n,|I|}$ as its estimator. In $j$th step of the step-down test, $j = 1, 2, ..., K - m$, we have

$$\hat{c}_{n,K-m-j-1} \geq \hat{d}_{n,K-m-j-1}$$

has probability tending to one as $n \to \infty$, where $\hat{d}_{n,K-m-j-1}$ is the estimate of $d_{K-m-j-1}$ defined in (3.12). Actually,

$$\hat{d}_{n,K-m-j+1} = \max_{\{i_1, i_2, ..., i_m\} \subseteq R_j} (\hat{c}^q_{n,J} \text{ with } J = A_j \cup \{i_1, i_2, ..., i_m\}),$$

62
where $\hat{c}_{n,J}^d$ denotes RB estimators of $c_{n,J}^d$, which are the critical values in the $(m + 1)$th largest $T$ enforced exchangeability step down test, to differentiate them from $c_{|J|}$ in this procedure.

Since the asymptotical control of gFWER for RB-based $(m + 1)$th largest $T$ enforced exchangeability step down test has been proved in previous theorem, this step-down procedure also controls gFWER asymptotically as sample size $n$ goes to infinity.

The RB-based $(m + 1)$th largest $T$ step down test then provides a computationally feasible re-sampling based multiple testing procedure, which controls gFWER asymptotically.

### 3.5 A Proof of Concept Experiment

The proposed gFWER-controlling multiple testing procedures will be applied on the mice experiment described in Section 1.3 to select marker genes for a validation study. Recall that, to simulate the comparison of phenotype groups, the experiment compared tissues from normal mice (wild type, labeled as group W) with tissues from four groups of mice (labeled as groups A, B, C and D) with four different mutations of the Mitf gene. Four mice were sampled from each of the five groups and four cRNA samples were prepared from each mouse. The NimbleGen microarrays with 12 mini-microarrays on each array are utilized. There are 99 genes in each mini-array. Each gene consists of 32 probes, and each probe is replicated twice. External reference samples are used to control for the platform change.

#### 3.5.1 Normalization Using an External Reference Sample

Gene expression measurements from these microarrays are normalized first to ensure observations from different sources are compatible before inferences are made.
There are two kinds of normalization techniques: the internal and external normalization. Internal normalization uses samples within a study as controls, while external normalization uses samples external to the study as controls. Some internal normalization techniques such as quantile normalization have been shown to be reliable within a study. However, how well inferences (such as prognostic algorithms) based on internal normalization carry across different studies is less clear. External normalization uses reference samples that are homogeneous and independent of platforms. Such external normalization can control for platform changes, provided it is as reliable within each platform as proven by internal normalization techniques.

We demonstrate the viability of external normalization by showing that internal and external normalizations produce almost identical results in our study. After correcting probe level measurements for backgrounds as described in Irizarry et al. (2003), we applied both normalization techniques, as described below.

With microarrays that allow only one biological sample to be placed on each array, it is unclear whether arrays from different groups should be normalized together or separately, due to confounding of array and group effects.

In our proof of concept experiment, however, the number of samples from each of the six groups is the same across all eight arrays. Specifically, every group of mice (groups A, B, C, D and W) and the external reference sample, UMRS, appear exactly twice on each array. It is thus reasonable to expect the distribution of the probe intensities to be the same across the arrays. We applied quantile normalization to equalize the distributions of the vectors of intensities from these eight arrays.

We propose an array-by-array external normalization process:
1. For each array, first generate a “reference” mini-microarray by averaging, for each probe, the intensities for that probe measured from UMRSs.

2. Then subtract the probe intensities in the reference mini-microarray from the corresponding probe intensities in every other (non-reference) mini-microarray.

To make an analogy to clinical trials, external normalization uses UMRS as a control.

We compared estimated differences between mutated types and wild type (A vs W, B vs W, C vs W and D vs W), after fitting the data normalized by the two techniques to the marginal model (3.13), using PROC MIXED of the SAS System. (Array 5 data was excluded due to bad quality.) Figure 3.2 shows that these two normalization techniques produce practically the same results. As quantile normalization is considered reliable, our study shows that normalization via external reference samples is a viable technique for coping with platform change issues.

Figure 3.3 displays observed expression levels of 99 genes from five groups of mice (after background correction and normalization). Separation of the groups cannot be seen. The reason turns out to be that mouse and sample variabilities overwhelm group differences.

In order to estimate the effect of each subject, and its variability, replicate samples from each subject is needed. In order to estimate the effect of each sample, and its variability, each sample needs to be probed multiple times.

To discover group differences, if subject and sample effects can be estimated unbiasedly, then removing them may make group differences reveal themselves more readily. By treating subject and sample as fixed effects (as one would adjusting for covariate effects), one may more readily identify differentially expressed genes. Multiple tests conducted in
Figure 3.2: Difference estimates between mutated types and wild type under quantile normalization and UMRS-based normalization.

this fashion control error rates conditionally, conditional on the subjects and the samples. Therefore, they control error rates unconditionally as well.

In analyzing expression levels to discover genes differentially expressed between groups, expression level is the response variable while group, subject, sample, and probe all are predictor variables. However, in training a classification algorithm based on differentially expressed genes, their expression levels then become predictors for treatment outcome. Sensitivity and specificity of such an algorithm depends on how variable expressions are between subjects within each group, and between samples within each subject. With replicate samples from each subject, and replicate probes for each sample, the variability of
subject and sample can be estimated by considering them as random effects in modeling expression level data. Analysis of gene expressions should be cognizant of this distinction between the roles of expression levels. Figure 3.4 displays estimated expression levels of 99 genes, after estimated mouse and sample effects are removed by modeling them as fixed effects. Clustering of 80 samples with the estimated gene expression levels rediscovered
the five groups of mice, arranging the columns corresponding to the 80 samples perfectly into five distinct groups.

Figure 3.4: Estimated expression levels of genes from inbred mice, with Mouse and Sample effects removed. Rows, from top to bottom, represent gene 1 to gene 99. Arranging the columns by unsupervised machine learning (clustering) results in five distinct groups: from left to right, groups C, D, A, B, and W (wildtype).
3.5.2 Multiple Testing for Differential Expressions

To discover genes differentially expressed between each mutated group (A, B, C, D) and wildtype (W), consider testing the $4 \times 99$ hypotheses

$$\{\{H_{ig} : \tau_{ig} = \tau_{iW}, \ g = A, B, C, D\}, \ i = 1, 2, \ldots, 99\}. \quad (3.23)$$

in the model (3.13).

For a particular gene $i$, estimates of $\tau_{ig} - \tau_{iW}, g = A, B, C, D$, are correlated due to estimating a common $\tau_{iW}$. For a balanced design, if the errors are i.i.d. with normal, this correlation is 0.5, not negligible, and it should be taken into account in multiple testing.

For a particular group $g$, estimates of $\tau_{ig} - \tau_{iW}, i = 1, \ldots, 99$, may have some correlation. Since there are $99 \times 98/2$ correlations, the correlation structure is not easy to infer given the typical amount of data from a microarray experiment. Therefore, one can take either a conservative approach regarding these correlations or a resampling approach to adjust for the correlations.

However, for a pair of $\tau_{ig} - \tau_{iW}$ and $\tau_{jh} - \tau_{jW}$ involving different genes and different groups, $i \neq j$ and $g \neq h$, correlation between estimates is expected to be small because the only dependence comes from the correlation between the estimates of $\tau_{iW}$ and $\tau_{jW}$. Suppose this correlation equals $\rho$. Then, for a balanced design, the correlation between the estimates of $\tau_{ig} - \tau_{iW}$ and $\tau_{jh} - \tau_{jW}$ equals $\rho/2$.

Therefore, a practical approach to controlling the error rate in testing the $4 \times 99$ hypotheses (3.23) is to take dependence into account in adjusting for multiplicity of testing $\{H_{ig} : \tau_{ig} = \tau_{iW}, \ g = A, B, C, D\}$ for each gene $i$ (in computing raw $p$-values). That is,
apply the Bonferroni correction adjusting for multiplicity of the four sets of hypotheses

\[ \{ H_{ig} : \tau_{ig} = \tau_{iW}, \; i = 1, 2, \ldots, 99 \}, \; g = A, B, C, D, \]  

(3.24)

and finally apply an appropriate method to control in testing each set of 99 hypotheses in (3.24).

Table 3.2 gives the number of genes discovered to be differentially expressed between each mutated group and wildtype, controlling gFWER at 5%, based on the linear model (3.13) with both mouse and sample effects as fixed effects. Reported are the results of applications of the Holm method, partition method using Markov’s inequality, and the augmentation method for testing each set of 99 hypotheses in (3.24) at gFWER of 5%/4 under the assumption that the errors are i.i.d. normal. Alternatively, without the normal assumption, the residuals from least squares estimates were resampled for 10,000 times, and the results of the resampling method are also reported.

All methods find that groups C and W are most different, and groups B and W are closest in terms of the measured gene expression levels. It confirms the findings in Figure 3.4. When FWER was controlled, Holm’s procedure and resampling method gave very similar results. When the number of mistakes allowed increased to 5, resampling method came out to be much less conservative than other gFWER-controlled methods, especially for the comparison between group B and group W.

3.6 Software Implementation

This section illustrates software implementation with SAS (linear mixed effects modeling) and R (multiple testing) for the modeling-based analyses of microarray data described in previous sections.
<table>
<thead>
<tr>
<th>m</th>
<th>Method</th>
<th>A vs. W</th>
<th>B vs. W</th>
<th>C vs. W</th>
<th>D vs. W</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Holm</td>
<td>94</td>
<td>18</td>
<td>99</td>
<td>92</td>
</tr>
<tr>
<td>5</td>
<td>Markov</td>
<td>94</td>
<td>27</td>
<td>99</td>
<td>95</td>
</tr>
<tr>
<td>5</td>
<td>Augmentation</td>
<td>99</td>
<td>23</td>
<td>99</td>
<td>97</td>
</tr>
<tr>
<td>0</td>
<td>Resampling</td>
<td>94</td>
<td>17</td>
<td>99</td>
<td>92</td>
</tr>
<tr>
<td>5</td>
<td>Resampling</td>
<td>97</td>
<td>48</td>
<td>99</td>
<td>96</td>
</tr>
</tbody>
</table>

Table 3.2: Number of rejections by the step-down method controlling the gFWER at 5%, Holm method, partition method using Markov’s inequality, augmentation method, and resampling method.

3.6.1 SAS Procedures

Sensitivity and specificity of a classification algorithm depend on how variable expressions are between subjects within each group and between samples within each subject, the variability of subject and sample can be estimated by considering them as random effects in modeling expression level data. In this case, the background corrected, log transformed and normalized probe level expressions for each gene are fitted by a linear mixed effect model with subject and sample effects as random:

```sas
proc mixed data=bcnorm;
  class group probe mice sample;
  model response= group probe/DFM=SATERTH solution;
  random mice(group) sample(group mice)/solution;
  lsmeans group/diff=control('W') adjust=dunnett;
run;
```

Group effects estimates, and variance components estimates for subject effect, sample effect and measurement error for each gene are saved for sample sizes calculation in validation trial.
On the other hand, to discover group differences, if subject and sample effects can be estimated unbiasedly, then removing them may make group differences reveal themselves more readily. By treating subject and sample as fixed, one may more readily identify differentially expressed genes. In this case, the background corrected and normalized probe level expressions for each gene are fitted by a linear fixed effect model. The residuals for each gene are saved for multiple testing procedure by resampling, without assuming the errors are normally distributed.

### 3.6.2 R Functions

To discover differentially expressed genes between two groups, multiple testing procedure by resampling the residual is applied to control the gFWER at level $\frac{5\%}{4}$. It is implemented by the following steps:

1. Resample independently with replacement the residual vectors after linear fixed effect modeling. To account for potential dependence among the measurement errors across genes, the residuals are resampled vector at a time, with each vector consisting of residuals from within each sample. For each re-sampled data set, we compute the test statistic for each gene, which is the difference of the average intensities, averaging within each mouse and then averaging over the mice within each group. After repeated resampling $B$ times, we have an estimated null distribution for the test statistic of each gene.

2. Calculate $p$-values for each gene by comparing the observed test statistics with the estimated null distribution `T.mat`.

\[
p < -(\text{abs(T.mat)} - \text{abs(t)} > 0) \times \text{rep(1, B)}/B
\]
3. Build null distribution for the $p$-values. Independently generate another matrix as we did in step 1. The null distribution for $p$-values is then estimated by comparing this new matrix `a.T.mat` with the matrix `T.mat`.

$$nullP<- 1- \text{apply}(T.mat,2,\text{fn.rawp.T}, a.T.mat)$$

$$\text{fn.rawp.T}<-\text{function}(nullT, obsT)$$

$$\{$$

$$\text{return}((\text{abs}(nullT)- \text{abs}(obsT)>0)\%*\%\text{rep}(1,B)/B)$$

$$\}$$

4. Calculate the adjusted p-value based on the generalized partitioning principal in Xu and Hsu (2007), to control the gFWER with $m = 5$.

```r
# sort the matrix nullP so that the corresponding
# raw p-values are increasing
ind<-\text{sort}(p,\text{index.return=T})$ix
P.mat.sort<-nullP[ind, ]

# step-down gFWER control
m<-5
adjp<-\text{rep}(\text{NA},99)
# step 0
adjp[1:m]<-0
P.mat1<-P.mat.sort
# step 1
# minp here is actually p(m+1)
minp<-\text{apply}(P.mat1,2,\text{function}(x) x[\text{(sort}(x,
+ \text{index.return=T})$ix)[m+1]])
adjp[m+1]<-\text{sum}(\text{minp}<p.H.sort[m+1,1])/B
# step 2 and etc.
for (i in (m+2):99)
{
P.mat1<-P.mat1[-1,]
minp<-\text{apply}(P.mat1,2, \text{function}(x) x[\text{(sort}(x,
+ \text{index.return=T})$ix)[m+1]])
adjp[i]<-\text{sum}(\text{minp}<p.H.sort[i,1])/B
}
```
# enforce adjp to be increasing
adjp[i]<-max(adjp[i-1], adjp[i])
}

# present the raw p-values, adj p-values
# and the corresponding gene No.
adjp.H<-cbind(adjp, p.H.sort)
colnames(adjp.H)<-c("adjp","rawp","gene")
adjp.sd.gfwer<-adjp
CHAPTER 4

DETERMINATION OF SAMPLE SIZE FOR VALIDATION STUDY IN PHARMACOGENOMICS

4.1 Introduction

Once a biomarker has been found in the training stage by multiple testing procedure with appropriately controlled error rate, we consider sample size calculation for a validation study in pharmacogenomics.

Generally speaking, sample size calculation is formulated as a problem of determining the number of subjects, \( n \), in a prospective experiment, where a random variable of interest is measured for each of the subjects, and its distribution is modeled with some unknown parameter. The sample size \( n \) is then calculated so that inferences and decisions about the parameter can be correctly made. It is customary to calculate sample size based on power (Adcock, 1997), that is, some hypotheses of interest are specified in terms of the parameter prior to the experiment, and then the sample size is determined to achieve a desired power at a fixed type I error rate. Chow, Shao and Wang (2003) elucidate this statistical approach to sample size calculation and provide its justification for different objectives in various clinical trials settings.
For multiple hypothesis testing with several parameters, appropriate definitions of power, type I error rate and the corresponding statistical test for the hypotheses of interest are necessary. Lee (2004) calculates sample size $n$ required for microarray experiments in which finding differentially expressed genes between a treatment condition and a control condition is of interest. In the approach, the number of subjects $n$ is determined to achieve a desired individual power level for a given mean number of false positives of multiple hypotheses (per-familywise type I error rate), when the ratio of mean difference to standard deviation for individual hypotheses and the anticipated number of undifferentially expressed genes are specified in advance. Depending on the purpose of experiments, sample size can also be specified to obtain optimal confidence regions for multiple hypothesis testing. In particular, for multiple comparisons using Tukey’s Multiple Pairwise Comparisons (MCA) method, the constrained Multiple Comparisons with the Best (MCB) method, and Dunnett’s two-sided Multiple Comparisons with a Control (MCC) method, Hsu(1988) suggests to calculate sample size so that with a pre-specified probability the confidence intervals for mean differences cover the true parameter values and be sufficiently narrow.

For development of prognostic or diagnostic devices, it is sensible to determine sample size to achieve specified levels of sensitivity and specificity, as they are common measures of prediction accuracy for binary diagnostic rules. Pepe (2003) uses this strategy to calculate sample size in the context of developing a medical device to differentiate a diseased group from non-diseased group. In her approach, a device is considered having minimally acceptable performance when the true positive fraction (TPF) is at least some value, say $\text{TPF}_0$ and the false positive fraction (FPF) is at most, say $\text{FPF}_0$. Note that TPF is sensitivity and FPF is one minus specificity. By setting a hypothesis to statistically prove that the device is minimally acceptable, that is, setting $H_0 : \text{TPF} \leq \text{TPF}_0$ or $\text{FPF} \geq \text{FPF}_0$, the sample
sizes for the two groups are chosen so that a positive conclusion would be drawn with a desired power at a specified type I error rate when the true TPF and FPF of the device are at some levels $TPF_1$ and $FPF_1$. Here, $TPF_1(>TPF_0)$ and $FPF_1(<FPF_0)$ are the values in the alternative hypothesis specified in advance by researchers.

Determination of sample size for validation study in pharmacogenomics is more complicated than the traditional sample size problem since it involves multiple layers of sample sizes and device-specific parameters. To design a microarray experiment for a validation study, in addition to the number of subjects per group, the number of replicate samples for each subject, the number of replicates of each probe on the chip, the number of genes, and the number of probes for each gene need to be decided. The sample sizes and device-specific parameters are determined so that the probability of sensitivity and specificity estimators being greater than some minimally acceptable values is sufficiently high.

To outline our proposed procedure, the number of genes as part of device-specific parameters is determined first by controlling a familywise type I error rate in multiple hypothesis testing for gene selection. Alternative gene selection procedures can be used if desired. A diagnostic rule is then built on the basis of the selected genes. Other device-specific parameters and the sample sizes are then calculated to meet pre-specified minimal sensitivity and specificity requirements for the diagnostic rule. Since the number of probes for each gene and the number of replicates of each probe are generally subject to spatial limitations of microarray platforms, there are typically upper bounds on these parameters, depending on the number of selected genes. A possible range of the number of replicated samples for each subject may be limited as well because the volume of a biological sample drawn from each subject is finite.
The aforementioned change of platforms in the pharmacogenomic setting brings another complication that the distributions of training data and validation data are not the same. Only when the training and validation experiments are properly designed, as in the proof-of-concept mice experiment to be shown later, the change of the distributions between two stages can be handled through statistical modeling. Otherwise, it does not seem feasible to deal with the change in a principled way.

Our formulation of sample size determination, in fact, requires sensitivity and specificity estimates. Compared to Pepe (2003)’s approach where $\text{TPF}_1$ and $\text{FPF}_1$ are pre-specified, the proposed approach using the confidence lower bound is more data-adaptive and controls directly the probability of meeting the minimal requirement. One may use either a model-based plug-in estimator or a sensible confidence lower bound of sensitivity and specificity based on the model for training and validation data. It will be shown that sample sizes calculated on the basis of a plug-in estimator may be too optimistic, but the bias can be remedied by a confidence lower bound.

Section 4.2 describes the mathematical formulation and general steps for sample sizes calculation in pharmacogenomics. For simplicity, how to determine sample sizes for sensitivity requirement is discussed only. Specificity requirement can be similarly dealt with. Section 4.3 illustrates an application of the procedure in the proof-of-concept mice experiment. Section 4.4 presents a simulation study for validation of the proposed sample sizes determination procedures followed by conclusions in Section 4.5.
4.2 Formulation of Sample Size Calculation

4.2.1 Basic Setting for Training and Validation Data

Suppose that there are \( K \) potential predictors of drug response in a training stage of a pharmacogenomic study. Let \( X \in \mathbb{R}^K \) be a random vector of the measurements of the predictors from a subject and \( Y \in \{0, 1\} \) be an indicator of whether the subject is a responder \((Y = 1)\) or a non-responder \((Y = 0)\). Let \( \mathcal{D}_n = \{(x_i, y_i)|i = 1, 2, ..., n\} \) denote the data with \( n \) subjects in the training stage, where \( x_i \), \( i = 1, 2, ..., n \) are the \( K \)-covariates, and \( y_i \) are the subgroup labels. \((x_i, y_i), i = 1, 2, ..., n \) are assumed to be independent and identically distributed with some unknown distribution \( P_{(X,Y)} \).

For validation data, let \((X^*, Y^*)\) denote a new case in the second stage, where \( X^* \in \mathbb{R}^k \) is a \( k \)-dimensional random vector, and \( Y^* \in \{0, 1\} \) is a subgroup label. Since only a subset of potential predictors of drug response are identified as a biomarker in the first stage and used in the second stage, \( k \) is typically much smaller than \( K \). Let \( Q_{(X^*, Y^*)} \) be the distribution of \((X^*, Y^*)\), which is different from \( P_{(X,Y)} \) because of the change of the platform. If microarrays are used to develop a diagnostic device, more replication of probes and more replicated samples from a subject are allowed in the validation stage. This change yields smaller variance parameters in \( Q_{(X^*, Y^*)} \) than \( P_{(X,Y)} \). A concrete example of such a change from \( P_{(X,Y)} \) to \( Q_{(X^*, Y^*)} \) is given in Section 4.3.

To take into account the change of platforms in a statistically principled way, we consider a model for the training data and derive a theoretically optimal diagnostic rule from the model for validation data (yet to be observed), reflecting the corresponding change.

Define \( \phi_v(x^*; \mathcal{D}_n) \) as such a theoretically optimal rule for the validation data, which depends on some unknown model parameters. \( \hat{\phi}_v(x^*; \mathcal{D}_n) \) is used to denote the plug-in diagnostic rule with the parameters replaced with estimates from the training data \( \mathcal{D}_n \). In
the validation stage, these parameter estimates are held fixed, and \( \hat{\phi}_V(x^*; D_n) \) is considered fixed. The sensitivity of \( \hat{\phi}_V(x^*; D_n) \) is then defined as

\[
\text{Sen}(\hat{\phi}_V) := P(\hat{\phi}_V(X^*; D_n) = 1|Y^* = 1).
\]

For brevity, \( \text{Sen}_V \) is used to refer to \( \text{Sen}(\hat{\phi}_V) \) in this paper.

Assume that given a validation sample of \( m \) i.i.d pairs of \((X_j^*, Y_j^*)\), \( j = 1, 2, ..., m \) with \( Y_j^* = 1 \), \( \text{Sen}_V \) is to be estimated by the sample proportion, \( \hat{\text{Sen}}_V \), of correctly predicting true positives by the diagnostic rule \( \hat{\phi}_V \) over validation data. That is,

\[
\hat{\text{Sen}}_V = \frac{1}{m} \sum_{j=1}^{m} I(\hat{\phi}_V(X_j^*, D_n) = 1|Y_j^* = 1).
\] (4.1)

\( \hat{\text{Sen}}_V \) is a simple unbiased estimator of the true sensitivity of \( \hat{\phi}_V \), and \( m \cdot \hat{\text{Sen}}_V \) follows a binomial distribution \( B(m, \text{Sen}_V) \).

### 4.2.2 Minimal Sensitivity Requirement

Consider the problem of finding the number of subjects \( m \) in each subgroup for a validation study such that the probability of \( \hat{\text{Sen}}_V \) exceeding the pre-specified level \( \gamma \) is at least \( 1 - \beta \), i.e.

\[
P(\hat{\text{Sen}}_V \geq \gamma) \geq 1 - \beta.
\] (4.2)

Figure 4.1 shows that given \( \beta \) and \( \text{Sen}_V(> \gamma) \), \( P(\hat{\text{Sen}}_V \geq \gamma) \) oscillates as a function of \( m \) and eventually surpasses \( 1 - \beta \) as \( m \) increases. As shown in the figure, there can be more than one crossing points of \( P(\hat{\text{Sen}}_V \geq \gamma) \) and \( 1 - \beta \), but for large enough \( m \), the inequality (4.2) holds. Hence, it is sensible to define the desired sample size \( m^* \) as the smallest number of subjects in each subgroup such that for any \( m \geq m^* \), the sensitivity requirement (4.2) is satisfied, given the minimal sensitivity \( \gamma \) and the minimal probability
of successful validation $1 - \beta$. The following proposition shows that as long as the true sensitivity is greater than the minimal level $\gamma$, $m^*$ is well-defined.

Figure 4.1: The probability $P(\hat{\text{Sen}}_{\gamma} \geq \gamma)$ as a function of $m$ and its lower bound $1 - I_{1-\text{Sen}_\gamma}(m - m\gamma, m\gamma + 1)$ when the true sensitivity $\text{Sen}_\gamma$ is 0.90 and the minimal level $\gamma$ is 0.80. The horizontal dashed line indicates the minimal probability of successful validation $1 - \beta$, with $\beta$ fixed at 0.05. $m_H$ is the upper bound given by Hoeffding’s inequality in (4.3), $m_0$ is the root of $I_{1-\text{Sen}_\gamma}(m - m\gamma, m\gamma + 1) = \beta$, and $m^*$ is the desired sample size. $m_H = 74.9$, $m_0 = 36.8$ and $m^* = 30$ in this example.

**Proposition 5.** If $\text{Sen}_\gamma > \gamma$, then the sample size $m^*$ is finite.

**Proof.** Since $m \cdot \hat{\text{Sen}}_\gamma$ follows $B(m, \text{Sen}_\gamma)$, by Hoeffding’s inequality (Hoeffding, 1963),

$$P(m \cdot \hat{\text{Sen}}_\gamma \leq m\gamma) \leq \exp\left(-2\frac{(m \cdot \text{Sen}_\gamma - m\gamma)^2}{m}\right) \text{ for } m\gamma < m \cdot \text{Sen}_\gamma.$$
So, if the probability upper bound is at most $\beta$, then (4.2) is satisfied, and
\[
\exp(-2 \frac{(m \cdot \text{Sen}_V - m\gamma)^2}{m}) \leq \beta
\]
implies
\[
m \geq -\frac{\log \beta}{2(S_{\text{env}} - \gamma)^2}.
\] (4.3)
Thus $m^*$ must be finite. $\square$

Let $m_H$ denote the upper bound given by Hoeffding’s inequality in (4.3). We can find $m^*$ easily by backward search; starting from $\lceil m_H \rceil$, where $\lceil x \rceil$ is the ceiling of $x$, i.e. the smallest integer greater than or equal to $x$, and decreasing the integer by one each time until we reach the first integer for which the inequality (4.2) is not satisfied. As the upper bound $m_H$ tends to be very large, this algorithm may not be efficient.

To sharpen the upper bound of $m^*$, we consider a continuous lower envelope of $P(\widehat{\text{Sen}}_V \geq \gamma)$ by using the relationship between the cdf of binomial distribution and the regularized incomplete Beta function. First observe that
\[
P(m \cdot \widehat{\text{Sen}}_V \geq m\gamma) \geq P(m \cdot \widehat{\text{Sen}}_V \geq \lceil m\gamma \rceil)
\]
\[
= 1 - P(m \cdot \widehat{\text{Sen}}_V \leq \lceil m\gamma \rceil - 1)
\]
\[
= 1 - I_{1-\text{Sen}_V}(m - \lceil m\gamma \rceil, \lceil m\gamma \rceil + 1),
\]
where $\lfloor x \rfloor$ is the floor of $x$. The last equality comes from the fact that for a binomial random variable $X$ with $B(n, p)$, $P(X \leq k) = I_{1-p}(n - k, k + 1)$ by integration by parts, where $I_x(a, b)$ is the cdf of $Beta(a, b)$, also known as the regularized incomplete beta function. Since $a$ and $b$ in $I_x(a, b)$ can be real values, and
\[
1 - I_{1-\text{Sen}_V}(m - \lfloor m\gamma \rfloor, \lfloor m\gamma \rfloor + 1) \geq 1 - I_{1-\text{Sen}_V}(m - m\gamma, m\gamma + 1),
\] (4.4)
by allowing \( m \) to be a real value, we obtain the right hand side of (4.4) as a continuous lower envelope of \( P(\hat{Sen}_V \geq \gamma) \). The dashed line in Figure 4.1 depicts such a lower envelope.

For \( Sen_V \geq \gamma \), the function \( 1 - I_{1-Sen_V}(m - m\gamma, m\gamma + 1) \) is shown to be strictly increasing in \( m \). By equating the function to \( 1 - \beta \) and solving for \( m \), we get a unique solution \( m_0 \), which serves as an upper bound of \( m^* \). That is, for any integer \( m \geq \lceil m_0 \rceil \), the minimal sensitivity requirement (4.2) is satisfied. \( m_0 \) is usually much smaller than \( m_H \) as illustrated in Figure 4.1. With this smaller initial point, \( \lceil m_0 \rceil \), the aforementioned backward search algorithm can be made more efficient. In summary, we start from \( \lceil m_0 \rceil \) and decrease \( m \) by one until the inequality (4.2) does not hold. Then the \( m \) value for the second last step is \( m^* \), the number of subjects per subgroup needed to meet the minimal sensitivity requirement.

Figure 4.2 illustrates how the sample size determined by the algorithm varies depending on the underlying true sensitivity at 95% of the probability of successful validation (\( \beta = 0.05 \)). Expectedly, the necessary sample size \( m^* \) decreases as the true sensitivity increases, and a larger sample size is required as the minimal level \( \gamma \) increases.

### 4.2.3 Estimation of True Sensitivity

The true sensitivity of the diagnostic rule \( \hat{\phi}_V \), i.e. \( Sen_V \) for the inequality (4.2), is usually unknown in practice, and it has to be estimated. As the diagnostic rule is derived from a model adjusting for change of platforms, its true sensitivity can be estimated by plugging in estimated model parameters. However, it is well known that such a plug-in estimator exhibits an upward bias in estimating the true sensitivity, and hence the sample size calculated based on the estimator could be smaller than necessary. Taking a conservative approach, we propose to determine sample sizes by replacing \( Sen_V \) by its confidence lower bound.
Figure 4.2: Sample sizes $m^*$ necessary to meet the minimal sensitivity requirement of $\gamma = 75\%, 80\%$ and $85\%$ with 95% of confidence as the true sensitivity varies.

For fixed $m$, the binomial distribution $B(m, Sen_v)$ is stochastically increasing in $Sen_v$. So, if $P(X \geq m\gamma) \geq 1 - \beta$ for a random variable $X$ of $B(m, p)$ with $p < Sen_v$, (4.2) also holds for $m \cdot \hat{Sen}_v$, which is stochastically greater than $X$. This is the rationale for replacing the true sensitivity with its confidence lower bound. Note that as in Proposition 5, for a proper sample size $m^*$, the confidence lower bound has to be greater than $\gamma$, and this condition provides a statistical criterion for futility.
4.3 Data Analysis

The purpose of the mice experiment described in Chapter 1 is to discriminate five mouse strains, A, B, C, D, and W (Wild Type) with different mutations in Mitf gene. In the experiment, the five strains are differentiated on the basis of the expression levels of 99 genes regulated by Mitf gene in spleen tissues. To mimic the situation in a pharmacogenomic study, consider classifying mice from two strains, say group A and group W. Regard group A as the responder group \( Y = 1 \) and group W as the non-responder group \( Y = 0 \). The 99 genes are taken as the potential genetic markers, from which a subset of genes predictive of responder or non-responder will be selected. This subset of genes are considered a biomarker. A diagnostic algorithm \( \hat{\phi}_V \) is then built based on these selected genes. It classifies mice into two groups: the biomarker positive subgroup \( \hat{\phi}_V = 1 \) and the biomarker negative subgroup \( \hat{\phi}_V = 0 \). The sensitivity of the algorithm is then the probability of predicting a mouse biomarker positive given the mouse is in group A. The specificity is the probability of predicting a mouse biomarker negative given the mouse is in group W.

Taking the training data from this experiment, we demonstrate the proposed procedure for determining sample sizes to achieve desired precision of sensitivity in a validation trial.

4.3.1 Construction of Diagnostic Rule

For statistical modeling of the data, let \( x_{igmspr} \) denote the background-corrected, log transformed and normalized probe intensity of mutated or wild type mouse for the \( i \)th gene \((i = 1, 2, \ldots, 99)\), the \( s \)th sample \((s = 1, 2, 3, 4)\) for the \( m \)th mouse \((m = 1, 2, 3, 4)\) in group \( g \) \((g = A, B, C, D, W)\), the \( p \)th probe \((p = 1, 2, \ldots, 32)\), and the \( r \)th replicate \((r = 1, 2)\). The probe intensities can be modeled separately for each gene. Assume \( x_{igmspr} \) to follow a
linear mixed effect model:

\[ x_{i\text{gmspr}} = \mu_i + \tau_{ig} + M_{im(g)} + S_{is(m(g))} + \pi_{p(i)} + \epsilon_{i\text{gmspr}}, \]  

(4.5)

where \( \mu_i \) is the mean gene expression for the \( i \)th gene, \( \tau_{ig} \) the \( g \)th group effect on the \( i \)th gene, \( M_{im(g)} \) is the \( m \)th subject effect in the \( g \)th group on the \( i \)th gene, \( S_{is(m(g))} \) is the \( s \)th sample effect from the \( m \)th subject in the \( g \)th group on the \( i \)th gene, and \( \pi_{p(i)} \) is the \( p \)th probe effect in the \( i \)th gene. We assume that \( \epsilon_{i\text{gmspr}} \) are independently and identically distributed with mean zero and variance \( \sigma_{i\epsilon}^2 \) within each gene. For the subject effects, we assume that \( M_{im(g)} \) are independent and identically distributed with mean zero and constant variance \( \sigma_{iM}^2 \) regardless of the group. Similarly for the sample effects, \( S_{is(m(g))} \) are assumed to be independent and identically distributed with mean zero and constant variance \( \sigma_{iS}^2 \) regardless of the subject and group. The \( \epsilon_{i\text{gmspr}} \)'s, \( M_{im(g)} \)'s and \( S_{is(m(g))} \)'s are also assumed to be independent.

Treating the problem in a general setting, suppose that there are \( n_i \) genes, \( n_m \) subjects in each group, \( n_p \) probes for each gene, \( n_s \) replicated samples for each subject, and \( n_r \) replicates of each probe on the chip. The sample sizes \( n_m \) and \( n_s \), and the device-specific parameters \( n_i, n_p \) and \( n_r \) can be different from the training stage to a validation stage. The linear mixed models in (4.5) now with these unspecified sample sizes and parameters \( n_m, n_s, n_p \) and \( n_r \) can characterize clearly the transition from \( P_{(X,Y)} \) to \( Q_{(X^*,Y^*)} \) due to the change of the platform. Hereafter, the superscripts \( t \) and \( v \) are used to indicate the change; \( t \) for training and \( v \) for validation. In the mice experiment, \( n_{t_m}^t = 4, n_{s}^t = 4, n_{i}^t = 99, n_{p}^t = 32 \), and \( n_{r}^t = 2 \).

Consider the normalized gene expression data from mutant group A and the wild type group W. Genes that seem reasonably good in separating A from W were found by the
average mean differences and multiplicity adjustment by the resampling-based partitioning testing procedure described in Hsu et al. (2009). The selected genes are RB1, USF1, Pu.1, Oa1, TPA1 and Bim, and their indices are 42, 45, 31, 29, 67, and 22, respectively. These six genes are used to build a diagnostic algorithm for discriminating the group A from the wild type W.

For simplicity, taking each mouse as a sampling unit, we consider prediction rules that use the average of probe intensities across the biological samples from the same mouse, and replicates as a summary measure of expression for each gene. The linear mixed effect model for the training experiment implies that the distribution of the average $\bar{X}_{igm...}$ is

$$N(\mu_i + \tau_{ig}, \sigma_{iM}^2 + \sigma_{iS}^2/n_s^i + \sigma_{i\epsilon}^2/(n_s^i n_p^i n_r^i)).$$  \hfill (4.6)

Suppose that a validation experiment has $n_v^m$ subjects (mice) in each group, $n_p^v$ probes for each gene, $n_s^v$ replicated samples for each subject, and $n_r^v$ replicates of each probe. Other than the sample sizes and device-specific parameters, validation data are assumed to have the same distribution as the training data. Then the distribution of the (unobserved) validation data $X_{igm...}$ is

$$N(\mu_i + \tau_{ig}, \sigma_{iM}^2 + \sigma_{iS}^2/n_s^v + \sigma_{i\epsilon}^2/(n_s^v n_p^v n_r^v)).$$  \hfill (4.7)

For convenience, the selected genes are relabeled so that their indices are from 1 to $n_v^i$, the number of genes used in the array for validation. $n_v^i = 6$ in the mice experiment.

As in Section 4.2.1, let $(X^*, Y^*)$ be a random vector for validation data. The conditional distribution of $X^* := (X_1^*, X_2^*, \ldots, X_{n_v^i})^\top$ given $Y^* = g$ is the same as the distribution of $(\bar{X}_{1gm...}^v, \bar{X}_{2gm...}^v, \ldots, \bar{X}_{n_v^i gm...}^v)^\top$. 87
If we assume independence across genes, the optimal classification rule under the normal distribution assumption is given by

$$
\phi_v(x^*) = \arg \min_g \sum_{i=1}^{n_v^r} \frac{(x^*_i - (\mu_i + \tau_{ig}))^2}{\sigma_{iM}^2 + \sigma_{iS}^2/n_v^s + \sigma_{ie}^2/n_p^vn_v^r},
$$

which is known as the diagonal linear discriminant analysis (DLDA). The decision rule $$\phi_v$$ can be estimated by a plug-in rule,

$$
\hat{\phi}_v(x^*) = \arg \min_g \sum_{i=1}^{n_v^r} \frac{(x^*_i - (\hat{\mu}_i + \hat{\tau}_{ig}))^2}{\hat{\sigma}_{iM}^2 + \hat{\sigma}_{iS}^2/n_v^s + \hat{\sigma}_{ie}^2/n_p^vn_v^r},
$$

where $$\hat{\mu}_i$$, $$\hat{\tau}_{ig}$$, $$\hat{\sigma}_{iM}$$, $$\hat{\sigma}_{iS}$$, and $$\hat{\sigma}_{ie}$$ are estimates of $$\mu_i$$, $$\tau_{ig}$$, $$\sigma_{iM}$$, $$\sigma_{iS}$$, and $$\sigma_{ie}$$ from the linear mixed effect model (4.5) for training data. We note that the estimates of the variance components are based on the observations from all five strains.

### 4.3.2 Sample Sizes Determination

In designing a validation experiment with microarrays, possible values of the devicespecific parameters $$n_p^v$$, $$n_r^v$$, and sample size $$n_s^v$$ are restricted due to the spatial and biological limitations mentioned in Section 4.1. In this mice experiment, $$n_p^v$$ is supposed to be the same as $$n_p^t = 32$$, as we believe that the 32 probes chosen by a biologist are sufficiently sensitive in measuring the expression levels of each gene. Since fewer genes are probed in each microarray for the validation stage than the training stage (from $$n_t^i = 99$$ to $$n_v^i = 6$$), this change allows more space for replication of each probe. With 200 spots in each mini-array, $$n_r^v$$ can be increased from $$n_r^t = 2$$ up to 33 now. For each mouse, its spleen tissue sample can only be divided into 4 to 6 pieces of replicated samples due to the fact that enough amount of a biological sample is required for efficient hybridization. This gives a range of 4 to 6 for $$n_s^v$$. Given $$n_p^v$$, $$n_r^v$$, and $$n_s^v$$, consider calculation of $$n_m^v$$ for validation of $$\hat{\phi}_v$$ such
that the sensitivity requirement (4.2) is satisfied. In other words,

\[ P(n_m^v, \widehat{Sen}_V \geq \gamma) \geq 1 - \beta, \] (4.8)

where \( n_m^v, \widehat{Sen}_V \sim B(n_m^v, Sen(\hat{\phi}_V)) \) with \( \hat{\phi}_V \) depending on \( n_s^v, n_p^v, n_r^v \), and the training data.

Under the assumption for the validation data in (4.7), direct calculation shows that the true sensitivity of \( \hat{\phi}_V \) is given by

\[ Sen(\hat{\phi}_V) = \Phi \left( \frac{(\hat{\theta}_1 - \hat{\theta}_0)^\top \hat{\Sigma}_v^{-1} \left[ \theta_1 - \frac{1}{2} (\hat{\theta}_0 + \hat{\theta}_1) \right]}{\left(\hat{\theta}_1 - \hat{\theta}_0\right)^\top \hat{\Sigma}_v^{-1} \hat{\Sigma}_v \hat{\Sigma}_v^{-1} (\hat{\theta}_1 - \hat{\theta}_0)}^{1/2} \right), \] (4.9)

where \( \theta_g = (\mu_1 + \tau_{1g}, \ldots, \mu_{n^v} + \tau_{n^v g})^\top \) for \( g = 0, 1 \), \( \Sigma_v = \text{diag}(\sigma_{1M}^2 + \sigma_{1S}^2/n_s^v + \sigma_{1E}^2/(n_p^v n_s^v n_r^v)) \), \( i = 1, 2, \ldots, i_v^v \), and \( \hat{\theta}_g \) and \( \hat{\Sigma}_v \) are plug-in estimates of \( \theta_g \) and \( \Sigma_v \). To determine the sample size \( n_m^v \), the true sensitivity of \( \hat{\phi}_V \) needs to be estimated. A simple plug-in estimator is

\[ \widehat{Sen}_V(\hat{\phi}_V) = \Phi \left( \frac{(\hat{\theta}_1 - \hat{\theta}_0)^\top \hat{\Sigma}_v^{-1} \left[ \theta_1 - \frac{1}{2} (\hat{\theta}_0 + \hat{\theta}_1) \right]}{\left(\hat{\theta}_1 - \hat{\theta}_0\right)^\top \hat{\Sigma}_v^{-1} \hat{\Sigma}_v \hat{\Sigma}_v^{-1} (\hat{\theta}_1 - \hat{\theta}_0)}^{1/2} \right) = \Phi \left( \frac{\hat{\delta}}{2} \right), \]

where \( \hat{\delta} \) is a sample version of the Mahalanobis distance between the two normal distributions \( \delta \) given as

\[ \{(\theta_1 - \theta_0)^\top \Sigma_v^{-1} (\theta_1 - \theta_0)\}^{1/2} = \left( \sum_{i=1}^{n_v^v} \frac{(\tau_{1i} - \tau_{0i})^2}{\sigma_{1M}^2 + \sigma_{1S}^2/n_s^v + \sigma_{1E}^2/(n_p^v n_s^v n_r^v)} \right)^{1/2}. \]

### 4.3.3 Confidence Lower Bound for Sensitivity

The simple plug-in estimator is generally observed to be biased upward, often yielding an optimistic estimate (P341, McLachlan, 1992, and Efron, 1983). Alternatively, one can use a confidence lower bound of \( Sen(\hat{\phi}_V) \) with appropriately chosen level \( 1 - \alpha \). The effect of \( \alpha \) on the sample sizes is investigated numerically in Section 4.4. Based on the numerical
results, $\alpha$ will be calibrated to attain the desired level $1 - \beta$ for the probability of meeting the sensitivity requirement in (4.2).

Letting $b = \hat{\Sigma}_v^{-1}(\hat{\theta}_1 - \hat{\theta}_0)$ and $a = \frac{1}{2} (\hat{\theta}_0 + \hat{\theta}_1)$, which are held fixed in the validation stage, we have the true sensitivity of $\hat{\phi}_v$ in (4.9) expressed as

$$Sen(\hat{\phi}_v) = \Phi \left( \frac{b^\top (\theta_1 - a)}{\sqrt{b^\top \Sigma_v b}} \right).$$

(4.10)

To construct a confidence lower bound for $Sen(\hat{\phi}_v)$, it is sufficient to construct a confidence lower bound of $\eta_{a,b} := \frac{b^\top (\theta_1 - a)}{\sqrt{b^\top \Sigma_v b}}$ since $\Phi(\cdot)$ is an increasing function.

To derive a confidence lower bound of $\eta_{a,b}$, we first consider an estimator of the form

$$\hat{\eta}_{a,b} = \frac{b^\top (\hat{\theta}_1 - a)}{\sqrt{b^\top \hat{\Sigma}_v b}},$$
given $a$ and $b$. $\hat{\theta}_1$ follows $N(\theta_1, \Sigma_t/n_m)$, where $\Sigma_t$ is a diagonal matrix with entries $\sigma_{iM}^2 + \sigma_{iS}^2/n_s + \sigma_{i\epsilon}^2/(n_p^l n_r^l n_u^l)$, $i = 1, 2, ..., n_v^v$. For the denominator,

$$b^\top \hat{\Sigma}_v b = \sum_{i=1}^{n_v} b_i^2 (\hat{\sigma}_{iM}^2 + \hat{\sigma}_{iS}^2/n_s^u + \hat{\sigma}_{i\epsilon}^2/(n_p^l n_r^l n_u^l)).$$

From the AVOVA table for the linear mixed effect model (4.5), we have the following expected mean squares:

$$E(MSE_i) = \sigma_{ie}^2,$$
$$E(MSS(GM)_i) = \sigma_{iS}^2 n_p^l n_r^l n_u^l + \sigma_{ie}^2,$$
$$E(MSM(G)_i) = \sigma_{iM}^2 n_p^l n_r^l n_u^l + \sigma_{iS}^2 n_p^l n_r^l n_u^l + \sigma_{ie}^2.$$
where \(MSE_i, MSS(GM)_i,\) and \(MSM(G)_i\) are the mean squares for \(\epsilon_i, S_{i(mq)},\) and \(M_{i(g)},\) respectively. The variance components \(\sigma^2_{IM}, \sigma^2_{IS},\) and \(\sigma^2_u\) are then estimated by the method of moment (P.398, Ravishanker and Dey, 2002). As a result, we have

\[
\begin{align*}
\hat{\sigma}^2_{ie} &= MSE_i, \\
\hat{\sigma}^2_{IS} &= \frac{MSS(GM)_i - MSE_i}{n_{b}^i n_{p}^t n_{r}^t}, \\
\hat{\sigma}^2_{IM} &= \frac{MSM(G)_i - MSS(GM)_i}{n_{b}^i n_{p}^t n_{r}^t}.
\end{align*}
\]

(4.11)

Hence

\[
b^\top \hat{\Sigma}_v b = \sum_{i=1}^{n_v} b_i^2 (c_MMSS(GM)_i + c_SMSS(GM)_i + c_EMSE_i),
\]

where \(c_M = \frac{1}{n^t_{b} n^t_{m} n^t_{p} n^t_{r}}, c_S = \frac{1}{n^t_{b} n^t_{p} n^t_{r}} - \frac{1}{n^t_{b} n^t_{r}}\) and \(c_E = \frac{1}{n^t_{b} n^t_{p} n^t_{r}} - \frac{1}{n^t_{b} n^t_{r}}\).

Since

\[
\begin{align*}
MSE_i &\sim \frac{\sigma^2_{ie}}{df_E} \chi^2_{df_E} \text{ with } df_E = n_{b}^t n_{m} n_{p}^t (n_{r}^t - 1), \\
MSS(GM)_i &\sim \frac{\sigma^2_{IS} n_{b}^t n_{p}^t + \sigma^2_{ie}}{df_S} \chi^2_{df_S} \text{ with } df_S = n_{b}^t n_{m} (n_{r}^t - 1), \\
MSM(G)_i &\sim \frac{\sigma^2_{IM} n_{b}^t n_{p}^t + \sigma^2_{IS} n_{b}^t n_{p}^t + \sigma^2_{ie}}{df_M} \chi^2_{df_M} \text{ with } df_M = n_{b}^t (n_{m}^t - 1),
\end{align*}
\]

(4.12)

and they are mutually independent, by Satterthwaite approximation (Satterthwaite, 1946), \(b^\top \hat{\Sigma}_v b\) follows approximately \((b^\top \hat{\Sigma}_v b / df) \chi^2_{df}\) with

\[
df = \left(\frac{\sum_{i=1}^{n_v} b_i^2 (c_MMSS(GM)_i + c_SMSS(GM)_i + c_EMSE_i)}{df_M}\right)^2 + \frac{(\sum_{i=1}^{n_v} b_i^2 (c_MMSS(GM)_i + c_SMSS(GM)_i + c_EMSE_i))^2}{df_M^2} + \frac{(\sum_{i=1}^{n_v} b_i^2 (c_MMSS(GM)_i + c_SMSS(GM)_i + c_EMSE_i))^2}{df_M^2}.
\]

By independence of \(\theta_1\) and \(b^\top \hat{\Sigma}_v b,\ \omega_b \eta_{\alpha,b}\) follows approximately \(t\)-distribution with degrees of freedom \(df\) and non-centrality parameter \(\omega_b \eta_{\alpha,b}\), where \(\omega_b = \sqrt{b^\top \hat{\Sigma}_v b / (b^\top n_m^t b)}\).
With $\omega_b$ estimated by $\hat{\omega}_b = \sqrt{b^\top \hat{\Sigma}_v b / (b^\top \hat{\Sigma}_t n)}$, the distribution of $\hat{\eta}_{a,b}$ is approximated by $t_{df}(\hat{\omega}_b \hat{\eta}_{a,b}) / \hat{\omega}_b$.

A $100(1 - \alpha)\%$ confidence lower bound of $\eta_{a,b}$ is then obtained by testing $H_0 : \eta_{a,b} = \eta_0$ versus $H_a : \eta_{a,b} > \eta_0$ with $\hat{\eta}_{a,b}$ as a test statistic and inverting the acceptance region for $\eta_{a,b}$. To find the expression of the confidence lower bound, let $\eta_{obs}$ be the observed value of $\hat{\eta}_{a,b}$. Denote the $p$-value for the one-sided test by $p(\eta_0) = P_{H_0}(\hat{\eta}_{a,b} \geq \eta_{obs})$. If $p(\eta_0) < \alpha$, $H_0$ is rejected. Under $H_0$, $\hat{\eta}_{a,b}$ follows $t_{df}(\hat{\omega}_b \eta_{a,b}) / \hat{\omega}_b$ approximately, and for fixed $\hat{\omega}_b$, $\hat{\eta}_{a,b}$ is stochastically increasing in $\eta_0$. Thus $p(\eta_0)$ is increasing in $\eta_0$, and the $100(1 - \alpha)\%$ confidence lower bound for $\eta_{a,b}$ is the smallest possible value $\eta_{LB}$ such that $p(\eta_{LB}) \geq \alpha$.

When sample sizes and device-specific parameters for training data are large, $df_E$, $df_S$, and $df_M$ tend to be large, which results in large degrees of freedom $df$. In this case, $\chi^2_{df} / df \approx 1$, and the random variable $\hat{\eta}_{a,b}$ can be further approximated by $N(\hat{\omega}_b \eta_{a,b}, 1) / \hat{\omega}_b$. Figure 4.3 shows the $p$-values of the hypothesis test, $p(\eta_0)$ for $\eta_{obs} = 1.94$, which is the observed value of $\hat{\eta}_{a,b}$ from the training data in the mice experiment. Note that the value of $\eta_{obs}$ depends on sample size $n_s^v$ and device-specified parameters $n_p^v$ and $n_r^v$ in the validation stage, and $n_s^v = 2$, $n_p^v = 32$, and $n_r^v = 2$ in this case. As shown in the figure, the 95% confidence lower bound found by non-central $t$ is less than that by normal approximation. Generally, normal approximation gives less conservative confidence lower bound.

Given $\eta_{LB}$, the confidence lower bound of sensitivity is given by $\Phi(\eta_{LB})$. Figure 4.4 shows sensitivity estimates by the simple plug-in estimator and $100(1 - \alpha)\%$ confidence lower bounds given by the proposed method. Clearly, the plug-in sensitivity estimates are bigger than the confidence lower bounds. As the number of replicated samples per subject $n_s^v$ increases, so does the true sensitivity of $\hat{\phi}_V$ in (4.10). Accordingly, the sensitivity estimates get higher as $n_s^v$ increases as shown in the figure.
Figure 4.3: p-value as a function of $\eta_0$ for testing $H_0: \eta_{a,b} = \eta_0$ versus $H_a: \eta_{a,b} > \eta_0$ when $n_s^v = 2$, $n_p^v = 32$, $n_r^v = 2$, and thus $df = 46.11$, $\hat{\omega}_b = 2.15$ and $\eta_{obs} = 1.94$ from the mice experiment. The solid line is for non-central $t$-distribution and the dashed line is for normal distribution. The 95% confidence lower bounds from non-central $t$ and normal approximation are $\eta_{LB}^t = 1.098$ and $\eta_{LB}^N = 1.177$ respectively, and the former is smaller than the latter.

Given a sensitivity estimate either by a simple plug-in estimator or 100$(1 - \alpha)$% confidence lower bound as in Figure 4.4, now consider determination of sample size $n_m^v$. Figure 4.5 gives the combination of $n_s^v$ and $n_m^v$ necessary to meet the minimal sensitivity level of $\gamma$ with at least 95% of probability when $\gamma$ was varied from 0.70 to 0.95. The left panel is for the simple plug-in estimator while the right panel is for 95% confidence lower bound. For example, to attain a minimal level of sensitivity of 90% with $n_s^v = 2$, $n_m^v$ has to be at least 20 from the left panel. Expectedly, as the minimal level of sensitivity increases, necessary
Figure 4.4: Sensitivity estimates by the simple plug-in estimator and $100(1 - \alpha)\%$ confidence lower bounds for varied number of replicated samples per subject with $\alpha = 0.10$ and $0.05$ respectively, given $n_{rv} = 2$ and $n_{pv} = 32$ in the mice experiment.

sample sizes increase as well, and relatively large sample sizes are needed when confidence lower bounds are used to estimate sensitivity.

4.4 Numerical Study

We investigate the validity of the proposed sample sizes determination procedure for a microarray experiment in the second stage of pharmacogenomics by simulation. Under the linear mixed effect model assumption in (4.5) for probe level data, training and validation data are generated from normal distributions as in (4.6) and (4.7), respectively.

For a numerical validation study, we take the following steps.
Figure 4.5: The required number of subjects per group and number of replicated samples per subject for various minimal sensitivity levels $\gamma$ when sensitivity is estimated either by the simple plug-in estimator (left panel) or by 95% confidence lower bounds (right panel). $n_v^p = 2$, $n_v^m = 32$ and $\beta = 0.05$.

Step 1: Specify the number of genes $n_v^i$ and true parameter values for means $\theta_g = (\mu_1 + \tau_{1g}, \mu_2 + \tau_{2g}, ..., \mu_{n_v^i} + \tau_{n_v^i g})^T$ and variance components $\sigma_{iM}^2$, $\sigma_{iS}^2$, and $\sigma_{i\epsilon}^2$, $i = 1, 2, ..., n_v^i$, in (4.6) and (4.7).

Step 2: After specifying sample sizes $n_g^t$, $n_m^t$, and $n_s^t$, and device-specific parameters $n_p^t$ and $n_r^t$ in the training stage, generate directly such summary statistics of training data as $\hat{\theta}_s$ from $N(\theta_g, \Sigma_t/n_m^t)$ and the values of mean squares $MSE_i$, $MSS(GM)_i$, and $MSM(G)_i$ from the scaled $\chi^2$ distributions in (4.12). Given the mean squares, the estimates of variance components $\hat{\sigma}_{i\epsilon}^2$, $\hat{\sigma}_{iS}^2$, and $\hat{\sigma}_{iM}^2$ are obtained from Equation (4.11).
Step 3: Regard $\hat{\theta}_g$, $\hat{\sigma}^2_{gE}$, $\hat{\sigma}^2_{gS}$, and $\hat{\sigma}^2_{gM}$ from the previous step as if they were calculated from raw probe-level training data. After specifying the sample size $n^v_s$ and the device-specific parameters $n^v_p$ and $n^v_r$, calculate sample size $n^v_m$ for a validation study to meet the minimal sensitivity requirement (4.2), as we did in the mice experiment in Section 4.3.

Step 4: Generate validation data from the model (4.7) with the sample size $n^v_m$ determined at the last step and the specified values of $n^v_s$, $n^v_p$, and $n^v_r$. Then calculate the empirical estimate of sensitivity $\hat{\text{Sen}}_v$ in (4.1).

Step 5: Repeat the previous step for a sufficient number of times to estimate $P(\hat{\text{Sen}}_v \geq \gamma)$.

If the estimated proportion at Step 5 is greater than $1 - \beta$, the procedure of sample size determination is deemed valid. Otherwise, it indicates that the sample sizes specified by the proposed approach are not large enough to meet the minimal sensitivity requirement.

Table 4.1 shows the estimates of $P(\hat{\text{Sen}}_v \geq \gamma)$ based on 5,000 repetitions from a simulation study, when $n^v_s$ varied from 2 up to 6, and the confidence level for a lower bound of $\text{Sen}_V$, $1 - \alpha$ varied from 98% down to 50%. The desired success $1 - \beta$ was 95% and the minimal level of sensitivity $\gamma$ was 0.8 for the study. To mimic the mice experiment, the parameters in the simulation were set to the actual estimates from the real data. The highlighted values are the smallest estimated probability of a successful validation experiment that exceeds the desired level 95% for each $n^v_s$, where the success is defined as $\hat{\text{Sen}}_v$ being at least 80%.

Results from the simulation study can be used to calibrate the confidence level $1 - \alpha$ for estimation of the true sensitivity to attain the desired success level $1 - \beta$. For example,
The level $100(1 - \alpha)\%$

<table>
<thead>
<tr>
<th>The number of replicated samples per subject ($n_v^v$) for confidence lower bound</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>98%</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>97%</td>
<td>0.996</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>96%</td>
<td><strong>0.962</strong></td>
<td>0.989</td>
<td>0.990</td>
<td>0.990</td>
<td>0.990</td>
</tr>
<tr>
<td>95%</td>
<td>0.938</td>
<td><strong>0.955</strong></td>
<td>0.969</td>
<td>0.974</td>
<td>0.978</td>
</tr>
<tr>
<td>94%</td>
<td>0.909</td>
<td>0.947</td>
<td><strong>0.953</strong></td>
<td><strong>0.964</strong></td>
<td>0.964</td>
</tr>
<tr>
<td>93%</td>
<td>0.891</td>
<td>0.922</td>
<td>0.945</td>
<td>0.947</td>
<td><strong>0.964</strong></td>
</tr>
<tr>
<td>92%</td>
<td>0.888</td>
<td>0.918</td>
<td>0.930</td>
<td>0.927</td>
<td>0.944</td>
</tr>
<tr>
<td>91%</td>
<td>0.875</td>
<td>0.912</td>
<td>0.917</td>
<td>0.928</td>
<td>0.942</td>
</tr>
<tr>
<td>90%</td>
<td>0.857</td>
<td>0.889</td>
<td>0.927</td>
<td>0.935</td>
<td>0.932</td>
</tr>
<tr>
<td>85%</td>
<td>0.847</td>
<td>0.894</td>
<td>0.887</td>
<td>0.893</td>
<td>0.905</td>
</tr>
<tr>
<td>80%</td>
<td>0.835</td>
<td>0.867</td>
<td>0.875</td>
<td>0.899</td>
<td>0.902</td>
</tr>
<tr>
<td>50%</td>
<td>0.718</td>
<td>0.735</td>
<td>0.749</td>
<td>0.763</td>
<td>0.776</td>
</tr>
</tbody>
</table>

Table 4.1: Proportions of $\hat{\text{Sen}}_v^v$ greater than pre-specified minimal level of sensitivity $\gamma = 0.8$ for different values of $n_v^v$ and confidence levels $1 - \alpha$ of lower bounds of the true sensitivity.

when $n_v^v = 6$, the table suggests that 93% confidence level is sufficient to obtain a proper sample size for $n_m^v$. In this study, the simple plug-in estimates for sensitivity are close to the 50% confidence lower bound, for which the estimates of $P(\hat{\text{Sen}}_v^v \geq \gamma)$ are all around 75%. The low success rate of the plug-in estimator is attributed to the fact that it gives a smaller sample size than necessary.

### 4.5 Conclusions

The main objective of this paper is to calculate sample size for a validation study to meet pre-specified sensitivity and specificity requirement, in order to avoid a futility of the pharmacogenomic development. Change of platforms is taken into account in the sample size calculation by statistical modeling. The proposed formulation for meeting minimal
sensitivity and specificity requirements calls for estimation of both measures. Their confidence lower bounds can substitute the true values in the sample size calculation procedure. However, the level $1 - \alpha$ has to be calibrated for appropriate sample sizes.

We can further investigate the robustness of the proposed procedure to see how sensitive the sample size calculation procedure is to the model assumptions. Our approach can also be extended to more general scenario where more complex diagnostic rules than DLDA are desired to take into account potential correlations among genes.
Pharmacogenomics is the co-development of a drug and a device, with the drug targeting a patient subgroup, and the device predicting whether a patient is in the subgroup of responders to the drug. Its development has two major stages: training stage and validation stage. To ensure a successful development in pharmacogenomics, statistically and practically sound methods are developed. Specially, we have formulated two procedures to solve two major statistical problems in the process of pharmacogenomic development. These two problems are the selection of biomarker in the training stage, to differentiate responders to the drug from non-responders, and determination of sample sizes for microarray experiments in validation study.

We illustrated our ideas of pharmacogenomic development in a proof-of-concept mice experiment, starting from design of microarray experiments, normalization of gene expressions to biomarker selection in training stage and sample sizes calculation for validation stage. In particular, a re-sampling based step down test, the RB-based step down test, is constructed by following the generalized partitioning principle and applied for biomarker selection. Dependence among hypotheses are estimated by a linear model and then re-sampling the residuals. The RB-based step down test is computationally feasible and it controls the generalized Familywise Error Rate asymptotically. After a biomarker is found
in the training stage by multiple testing procedure with appropriately controlled error rate, sample size calculation for validation study is formulated. Sample sizes for microarray experiments are determined to meet pre-specified sensitivity and specificity requirements, in order to avoid futility of pharmacogenomic development. Change of platforms brings a complication in sample size calculation, since the training data and validation data do not follow the same distribution. When the training and validation experiments are properly designed, the change of the distributions between two stages can be handled by statistical modeling. Our formulation of sample size determination procedure calls for estimation of the true sensitivity and specificity. Their confidence lower bounds can substitute the true values in the sample size calculation procedure. However, the confidence level for lower bounds has to be calibrated for appropriate sample sizes.

Although the RB-based step down test controls gFWER asymptotically, it does not control the error rate strongly with finite samples, when the exchangeability assumption of the test statistics does not hold. It will be of great interest to further investigate the finite sample performance of this re-sampling based multiple testing procedure, especially for microarray data. Current available gFWER-controlling procedures all use a single order statistic as the test statistic for each disjoint hypothesis. Alternative test statistics, such as a vector of order statistics, deserve further investigation if desired. FWER and gFWER are exceedance probability of the number of false discoveries, and FDR is an expectation of the number of false discoveries. They only partially convey the information of false-discovery distribution. Further study of the relevant distribution of false discovery can produce a global image of the corresponding multiple-inference errors. That will greatly improve our understandings on multiple-inference errors.
In the proposed sample size determination procedure, further study need to be done to give explicit formulation of the calibration of confidence level for lower bounds based on pre-specified sensitivity and specificity requirement. If it is infeasible, a general instruction of choosing appropriate confidence level for lower bounds is expected. We will also investigate the robustness of the proposed sample size procedure to see how sensitive our procedure is to the model assumptions. The sample size procedure can also be extended to more general scenarios, where more complicate diagnostic rules than diagonal discriminant linear analysis are required to take into account the potential correlations across different genes.


103


