ROBUST ESTIMATION AND HYPOTHESIS TESTING IN MICROARRAY ANALYSIS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

BURÇİN EMRE ÜLGEN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
STATISTICS

AUGUST 2010

Approval of the thesis:

ROBUST ESTIMATION AND HYPOTHESIS TESTING IN MICROARRAY ANALYSIS

submitted by **BURÇİN EMRE ÜLGEN** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Statistics, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. H. Öztaş Ayhan _____
Head of Department, **Statistics**

Prof. Dr. Ayşen Akkaya _____
Supervisor, **Statistics Dept., METU**

Examining Committee Members:

Prof. Dr. Zeki Kaya _____
Biology Dept., METU

Prof. Dr. Ayşen Akkaya _____
Statistics Dept., METU

Assoc. Prof. Dr. Barış Sürücü _____
Statistics Dept., METU

Assistant Prof. Dr. Tolga Can _____
Computer Engineering Dept., METU

Assistant Prof. Dr. Özlen Konu _____
Molecular Biology and Genetics Dept., Bilkent University

Date: 05.08.2010

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: **Burçin Emre ÜLGEN**

Signature:

ABSTRACT

ROBUST ESTIMATION AND HYPOTHESIS TESTING IN MICROARRAY ANALYSIS

Ülgen, Burçin Emre
Ph.D., Department of Statistics
Supervisor: Prof. Dr. Ayşen Akkaya

August 2010, 116 pages

Microarray technology allows the simultaneous measurement of thousands of gene expressions simultaneously. As a result of this, many statistical methods emerged for identifying differentially expressed genes. Kerr et al. (2001) proposed analysis of variance (ANOVA) procedure for the analysis of gene expression data. Their estimators are based on the assumption of normality, however the parameter estimates and residuals from this analysis are notably heavier-tailed than normal as they commented. Since non-normality complicates the data analysis and results in inefficient estimators, it is very important to develop statistical procedures which are efficient and robust. For this reason, in this work, we use Modified Maximum Likelihood (MML) and Adaptive Maximum Likelihood estimation method (Tiku and Suresh, 1992) and show that MML and AMML estimators are more efficient and robust. In our study we compared

MML and AMML method with widely used statistical analysis methods via simulations and real microarray data sets.

Keywords: gene expression, long-tailed symmetric family, modified maximum likelihood, robustness.

ÖZ

ROBUST ESTIMATION AND HYPOTHESIS TESTING IN MICROARRAY ANALYSIS

Ülgen, Burçin Emre
Doktora, İstatistik Bölümü
Tez Yöneticisi: Prof. Dr. Ayşen Akkaya

Ağustos 2010, 116 sayfa

Mikrodizin teknolojisi, binlerce gen ifadesinin eşzamanlı olarak ölçülmesine olanak sağlamaktadır. Bunun sonucu olarak, farklı ifade olan genlerin belirlenmesi için birçok istatistiksel yöntem ortaya çıkmıştır. Kerr ve diğerleri (2001), mikrodizin verisinin analizi için varyans analizi yöntemini önermişlerdir. Fakat çalışmalarında açıkladıkları gibi, bu analizden elde edilen parametre tahminleri ve artıkların normalden daha uzun kuyruklu olmalarına rağmen, analizleri ve tahminleyicileri normallik varsayımına dayanmaktadır. Normal olmama durumu, veri analizini zorlaştırdığı ve verimsiz tahminleyicilere yol açtığı için, etkin ve sağlam istatistiksel yöntemler geliştirmek çok önemlidir. Bu amaçla, bu çalışmada, varyans analizi için uyarlanmış en çok olabilirlik tahminleme yöntemi (Tiku ve Suresh, 1992) ile adaptif uyarlanmış en çok olabilirlik tahminleme yöntemi kullanılmış ve bu tahminleyicilerin daha etkin ve sağlam

oldukarı gösterilmiştir. Uyarlanmış ve adaptif uyarlanmış en çok olabilirlik tahminleyicileri, yaygın kullanılan yöntemlerle simulasyonlar ve gerçek mikrodizin verileri kullanılarak karşılaştırılmıştır.

Anahtar kelimeler: gen ifadesi, uzun kuyruklu simetrik dağılım, uyarlanmış en çok olabilirlik, sağlamlık.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

Microarray technology is an array-based technology that was developed for measuring the expression levels of large number of genes at once, thereby bringing about a tremendous improvement over the "one gene per experiment" paradigm (Amaratunga and Cabrera, 2004). Consequently, they have become common tools in biological research and triggered the need for effective statistical methods for data analysis.

Microarrays have already been excessively used in biological research to address a wide variety of questions. One of the most frequently used microarray applications is to compare gene expression levels under two or more different conditions. The main purpose of such a microarray experiment is to identify differences in gene expression among varieties (the categories of the factors under the study such as tissue types, drug treatments etc.). Since the data is noisy due to the variability arising throughout the measurement process, the problem is how to determine that the observed level of differential expression is statistically significant. A number of statistical methods have been suggested for the identification of differentially expressed genes.

Kerr et al. (2000) use a log-linear ANOVA model to make valid estimates of the relative expression for genes that are not biased by

ancillary sources of variation. They demonstrated that ANOVA methods can be used to normalize microarray data and provide estimates of changes in gene expression. To account for the multiple sources of the variation in a microarray experiment, they consider the following model

$$\log(y_{ijkg}) = \mu + A_i + V_k + G_g + (AG)_{ig} + (VG)_{kg} + \varepsilon_{ijkg} \qquad (1.1)$$

where μ is the average overall signal, A_i represents the effect of the i^{th} array, V_k represents the effect of the k^{th} variety, G_g represents the effect of the g^{th} gene and $(VG)_{kg}$ represents the interaction between the k^{th} variety and the g^{th} gene. The error terms ε_{ijkg} are assumed to be independently and identically distributed with mean zero.

Because of the fact that the distribution of residuals are notably different than normal as Kerr et al. (2000) commented, they used a bootstrap approach to obtain confidence intervals without relying on normality assumptions. However, the model fit and parameter estimates in their study were obtained by the method of least squares, which is most efficient for normal data. Since non-normality complicates the data analysis and results in inefficient estimators, it is very important to develop statistical procedures which are efficient and robust for non-normal distributions.

A number of studies have been carried out to investigate the effect of non-normality on the test statistics used in the analysis of variance. The effect of non-normality on Type I error was studied by Pearson (1931), Geary (1947), Gayen (1950), Box and Andersen (1955), Hack

(1958), Box and Watson (1962), Tiku (1964) and the effect of non-normality on Type II error was studied by David and Johnson (1951), Srivastava (1959), Donaldson (1968) and Tiku (1971). The effect of moderate non-normality on the level of significance is known to be not serious but the power is considerably lower. Since non-normal distributions occur so frequently in practice, it is very important to develop statistical procedures which are robust and efficient for non-normal distributions.

The aim of this thesis is to research the possibilities to create and deploy robust estimation techniques for the analysis of variance for microarray data. We study the unbalanced two-way ANOVA model with interaction where error terms have a distribution from long-tailed symmetric (LTS) family and suggest robust estimators and test statistics obtained by Modified Maximum Likelihood (MML) estimation technique introduced by Tiku (1967) and Tiku and Suresh (1992). We also facilitate Adaptive Modified Maximum Likelihood (AMML) estimation technique introduced by Tiku and Sürücü (2009) for robust estimators and test statistics. The results are compared with widely used parametric and non-parametric methods via simulated datasets and real microarray experiments.

The outline of this thesis is as follows: Chapter 1 presents a biological background of the gene, DNA and RNA molecules and provides a brief information about the usage of microarray technology in analyzing gene expression. It also gives issues about data analysis preparation, transformation and normalization methods, statistical techniques used for analysis of microarray data and multiple testing procedure. In Chapter 2, theoretical background of the unbalanced two-way

classification fixed-effect model with interaction is given in detail. Since the microarray data used in the model fit a LTS distribution, the LTS family is introduced and its properties are explained. Also the model parameters are estimated by using LS and MML methods and efficiency properties of the estimators are examined. Testing the main and interaction effects under the assumption of normality is given. Then test statistics for testing main and interaction effects are developed by using the MML estimators in the case that the errors have a distribution from the LTS family. Further, a test statistic is obtained to make pairwise multiple comparisons of the treatment means under the LTS distribution by using W24 and MML estimators of location and scale parameter and its properties are studied. Adaptive modified maximum likelihood (AMML) estimators are obtained for the unbalanced two-way classification model with interaction and the corresponding hypothesis tests are given in Chapter 3. In Chapter 4, using both the three real microarray experiments and the simulated datasets, parametric methods such as t-test, Bayes t-test, ANOVA, Huber estimation, MMLE and AMMLE and non-parametric method such as SAM are compared. Finally, the conclusions are presented in Chapter 5.

1.1 Biological Background

A cell is the minimal and fundamental unit of all living organisms, both structurally and functionally. A cell contains many macromolecules that organize and coordinate all of the events. Macromolecules control and govern most of the activities of life. Deoxyribonucleic acid (DNA) molecules store information about the

structure of macromolecules, allowing them to be made precisely according to cells' specifications and needs (Lee, 2004).

DNA is a very stable molecule that forms the blueprint of an organism. The DNA structure encodes information as a sequence of chemically linked molecules that can be read by cellular machinery and guides the construction of proteins which are essential parts of organisms. Its' structure consists of two long strands wound tightly around each other in a spiral structure known as double helix. These strands are chains of chemical building blocks called nucleotides. The four type of nucleotides in DNA are; adenine (A), guanine (G). thymine (T) and cytosine (C). Genetic information is encoded in DNA by the sequence of these nucleotides (Amaratunga and Cabrera., 2004)

A gene is a segment or region of DNA that encodes specific instructions, which allow a cell to produce a specific product. Genes act as tiny switches that direct the specific sequence of events that are necessary to create a human being. They affect every part of our physical and biochemical systems, acting in cascade of events turning on and off the expression, or production, of key proteins that are involved in different steps of development. A gene is active, or expressed, if the cell produces the protein encoded by the gene. If a lot of protein is produced, the gene is said to be highly expressed. If no protein is produced, the gene is not expressed (unexpressed). The expression of the gene will be determined by various internal (such as gender, hormones, metabolism etc.) and external factors (such as drugs, temperature, light etc.) The objective of researchers is to

detect and quantify gene expression levels under particular circumstances (Draghici, 2003).

The process of using the information encoded into a gene to produce the coded protein involves reading the DNA sequence of the gene. The first part of this process is called transcription and is performed by a specialized enzyme called ribonucleic acid (RNA) polymerase. The transformation process converts the information coded into DNA sequence of the gene into an RNA sequence. Then the RNA is transferred into a machinery that synthesizes protein molecules based on the information carried by the RNA. The process is called translation. This flow of genetic information from DNA to RNA to proteins mentioned above called the central dogma of molecular biology that formulates how information is stored and converted to all the components that build up a living organism (Lee, 2004).

During the transcription process, a high proportion of messenger RNA (mRNA) which is one of the three types of RNA, are produced for encoding in most molecules. Since mRNA is an exact copy of the DNA coding regions, in experiments usually mRNA levels are measured to explore the process in coding regions of DNA (Draghici, 2003). Consequently, the measure of gene expression under different conditions such as drug treatments, shocks, diseases can be determined from the analysis of the mRNA levels. For this reason, scientists study the amounts of mRNA produced by a cell to learn which genes are expressed, which in turn provides insights into how the cell responds to its changing needs.

1.2 Microarray Technology

A DNA microarray (microarray, for short) is a tool for analyzing gene expression that consists of a small membrane or glass slide containing samples of many genes arranged in a regular pattern. The surface of a microarray is spotted with oligonucleotides (small parts of DNA molecules up to 25 nucleotides), complementary DNA (cDNA) or small fragments of polymerase chain reaction (a technique in molecular biology to amplify a single or few copies of a piece of DNA) products that represent specific gene coding regions. A microarray contains thousands of microscopic spot, also known as probes, each of which is for one particular gene (Amaratunga and Cabrera, 2004).

The basic premise of the microarray is that mRNA samples prepared by the researcher from experimental organisms (such as tumor cells) are bind, or hybridize, to known probes on the arrays based on the central dogma biology mentioned in Section 1.1. In other words, a microarray works by exploiting the ability of a given mRNA molecule specifically hybridize to, the DNA template from which it originated. Since the mRNA samples used were labeled by a fluorescent dye, the mRNA that is hybridized to its complementary DNA on the microarray leaves its fluorescent tag. Total strength of the signal, from a probe, depends upon the amount of target sample binding to the probes present on that spot. Then a special scanner is used to measure the fluorescent areas on the microarray and convert the signals into raw data. By this way, the amount of mRNA bound to the spots on the microarray is precisely measured, generating a profile of gene expression in the cell. At this point, the data may then be entered into a database and analyzed by a number of statistical methods. A

typical microarray experiment can be summarized by the following five steps (Parmigiani et al., 2003):

1. Preparing the microarray
2. Preparing the sample
3. Hybridizing the labeled sample to the microarray
4. Scanning the microarray
5. Analyzing the data by statistical methods

There are two main types of microarray; single-channel (one-color) and two-channel (two-color) arrays. One-channel arrays allow the hybridization of only one biologic sample per array whereas two-channel arrays incorporates the hybridization of two samples per array since they use different dyes for two samples. However, two-channel arrays do have an additional variance (other source of variations will be mentioned in Section 1.3) due to the dye effects. In this thesis, analyses and estimators are derived for one-channel arrays for conciseness but the results can be generalized to two-channel arrays as well.

1.3 Data Analysis Preparation

Researchers are interested in two kinds of basic qualities about the microarray data; biological significance and statistical significance. The biological significance tells how much the expression of a gene is influenced by the condition under study. The statistical significance tells how trustworthy the biological significance is i.e. whether a result occurs by chance or not. The statistical analysis is crucial for successful interpretation of the biological phenomena under study

since there are many sources of variability in microarray experiments. Noise is introduced at each step of various procedures (Schuchhardt et. al, 2000). The main challenging issues about the analysis of microarrays can be listed as follows:

1.3.1 Transformation

Gene expression values do not have desired properties such as constant variance without being transformed. To transform the data into a scale suitable for analysis, different data transformation methods have been used. An overview of transformation techniques are given by Lee (2004).

In this thesis, we analyze the data on the log scale since the log transform is the natural method for analyzing data with an additive model where the effects in the data are believed to be multiplicative (Kerr et. al., 2000). There are other reasons why log-transformation is beneficial. First, microarray intensities are typically asymmetrically distributed. This makes it difficult to estimate certain characteristics of the data. Log-transforming the data makes the distribution more symmetric. Second, variation in intensities typically grows with average intensities. This is a violation of the general assumption of parametric models that all groups have similar variances. The variation of logged intensities tends to be less dependent on the magnitude of the values and as a result the power of statistical tests increases. Further, exploration of untransformed data and the examination of other transformations led us to conclude that the log transform is a good choice (Sapir and Churchill, 2000).

1.3.2 Background correction

The background correction step aims to remove non-biological contributions to a measured signal. Typical examples of non-biological contributions to a signal are caused from mRNA preparation (tissues, kits and procedures vary), transcription (inherent variation in the reaction, enzymes), surface chemistry, humidity, slide inhomogeneities, hybridization parameters (time, temperature, etc.), unspecific hybridization (labelled cDNA hybridized on areas which do not contain perfectly complementary sequences) and scanning.

The non-biological contributions complicates the analysis of microarray data when comparing different tissues or different experiments. Because it makes difficult to determine whether the variation of a particular gene is due to the noise or due to the difference between the different conditions tested. This kind of noise is an unescapable phenomenon for microarray data because there is inherent noise in the data even after systematic sources of variation are removed. In order to reduce the noise as possible, Lee et. al. (2000) noted replication is crucial to microarray studies. For this reason, we study the statistical analysis of microarray experiments with replication in this thesis.

1.3.3 Normalization

Normalization is a data pre-processing step by which one makes the different samples of an experiment comparable to one another. In other words, it aims to remove systematic differences across different

data sets and eliminate artifacts by minimizing extraneous variation in the measured gene expression levels of hybridized mRNA samples. By this way, biological differences can be more easily distinguished.

Normalization can be necessary for different reasons such as different quantities of mRNA, saturation toward the extremities of range, etc. For this reason, there are different normalization procedures which differ with respect to which kind of average is used and what sources of variability are taken into account. An overview of normalization methods is given by Quackenbush (2002). However, such correction procedures will most likely to remove some of the biological signal as well. The extent of how much biological signal is removed depends on the characteristics of both the biological experiment and the technical quality of the microarray experiment.

In this thesis, we do not facilitate any normalization methods prior to the data analysis since Kerr et al. (2000) stated that the effects in the ANOVA model that they used for analysis of microarray data, normalize the data without the need to introduce preliminary data manipulation. In this way, the normalization process can be combined with the data analysis. As they stated, this kind of normalization is based on a clearly stated set of assumptions that can be evaluated using information in the data and the ANOVA analysis systematically estimates the normalization parameters based on all of the relevant data.

1.4 Statistical Methods for Differential Gene Expression

A common objective in microarray studies is to identify the genes that are consistently differentially expressed under certain conditions. The null hypothesis being tested is that there is no difference in expression between the conditions. To this end, the difference between the expression levels is estimated and tested whether the observed differences are statistically significant.

Detecting differential expression between conditions depends on the choice of test statistic, which in turn depends on assumptions are believed to be distributed across the samples. The choice of test statistics can greatly affect the set of genes that are identified, particularly in small sample-sized studies (Draghici, 2009). A complete overview of statistical methods for microarray data can be found in Lee (2004).

In the following sections, we review several widely used statistical tests for determining differential expression in microarray data. It should also be noticed that fold change and clustering are not covered because of the following reasons. Fold change is not a statistical test, and there is no associated value that can indicate the level of confidence in the designation of genes as differentially expressed or not. Also cluster analysis is not a statistical test and it is not a sensitive method for this type of study because it focuses on group similarities, not differences between each individual gene.

1.4.1 The t-test

The t-test is a simple, statistically based method for detecting differentially expressed genes. The two sample t-test statistic with two independent normal samples without assuming the equal variances between two samples is as follows;

$$t_g = \frac{\bar{y}_{g1} - \bar{y}_{g2}}{\sqrt{\frac{s_{g1}^2}{n_1} + \frac{s_{g2}^2}{n_2}}} \qquad (1.4.1.1)$$

Suppose that the experimental data consist of measurements y_{gi} under i conditions, where g=1,2,...,G denotes the g^{th} gene, and n_i is the replication number of gene under the i^{th} condition. Let the sample mean and sample variance of y_{gi}'s for gene g be denoted as $\bar{y}_{gi}$ and s_{gi}^2 respectively.

A gene with very small variance due to its low expression level contributes to have large absolute t-value regardless of the mean difference under two conditions and this gene can be selected as the differentially expressed although they are not truly differentially expressed (Kim et. al, 2006). To overcome this problem of traditional t-test, various methods (two examples of these methods are ment,oned in Section 1.4.2 and 1.4.3) have been proposed.

It should be noted that, the t-test assumes normality and constant variance for every gene across all samples. These assumptions are

certainly inappropriate for a subset of genes despite any given transformation (Thomas et. al, 2001).

1.4.2 Significance Analysis of Microarrays

Tusher et al. (2001) developed significance analysis of microarrays (SAM), also known as penalized t-test, which assigns a score to each gene on the basis of change in gene expression relative to the standard deviation of repeated measurements. For genes with scores greater than an adjustable threshold, SAM uses permutations of the repeated measurements to estimate the false discovery rate (FDR) which is mentioned in Section 1.5 and to avoid the small variance problem of t-test. As mentioned in Section 1.4.1, the shortcoming of traditional t-test is that genes with small variances due to the low expression levels have high chance of being declared as the differentially expressed genes. SAM added a small positive constant, s_0 to alleviate this problem. The SAM statistic is

$$t_g = \frac{\bar{y}_{g1} - \bar{y}_{g2}}{s_g + s_0} \qquad (1.4.2.1)$$

where

$$s_g = \sqrt{a\left\{\sum_g \left[y_{g1} - \bar{y}_{g1}\right]^2 + \sum_g \left[y_{g2} - \bar{y}_{g2}\right]^2\right\}},$$

$$a = (1/n_1 + 1/n_2)/(n_1 + n_2 - 2).$$

The coefficient of variation of t_g is computed as a function of s_g in moving windows across the data. The value of s_0 is chosen is to minimize the coefficient of variation.

SAM makes use of permutations to simulate for every gene a situation when there is no difference between the two groups. First the samples are randomly shuffled between two groups a number of times (about 1000 times) and afterwards the t_g is calculated in each of these datasets. The average of all these t_g values is then used as an estimate of the expected t_g value of that gene if it would have not been differentially expressed. The observed t_g values are plotted versus the expected t_g values. Next, an arbitrary cut-off needs to be chosen. The choice is not straightforward as it reflects the compromise one needs to make between the number of significant genes and the number of false positive results. A gene that deviates more than one delta from the diagonal, and all genes that have more extreme t_g values than this gene are called significant.

SAM is based on a nice rationale, but it is computationally quite intensive and limited to comparisons between two groups (Göhlmann, 2009).

1.4.3 Bayes t-test

Baldi and Long (2001) developed a Bayesian probabilistic framework for microarray data analysis. At the simplest level, they modelled log-expression values by independent normal distributions,

parameterized by corresponding means and variances with hierarchical prior distributions. Their statistic is used to solve small sample variances and use the parametric Bayesian method to estimate the parameters for the t-test.

Bayes t-test uses the estimate of parameters such as population mean (μ) and variance (σ^2) by Bayesian method. The mean and variance of posterior estimate for the j^{th} group is given as

$$\mu_j = \mu_{nj} \ , \quad \sigma_j^2 = \frac{v_j \sigma_{nj}^2}{v_j - 2} \tag{1.4.3.1}$$

where the mean of the posterior estimate (μ_{nj}) is a convex weighted average of the prior mean (μ_{0j}) and the sample mean ($\bar{y}_j$) for the j^{th} group, that is

$$\mu_{nj} = \frac{\lambda_{0j}}{\lambda_{0j} + n_j} \mu_{0j} + \frac{n_j}{\lambda_{0j} + n_j} \bar{y}_j \tag{1.4.3.2}$$

The hyperparameters μ_{0j} and σ_j^2/λ_{0j} can be interpreted as the location and scale of μ_j, respectively, and n_j is the sample size for each group. σ_{nj}^2 is posterior variance component, and the posterior degree of freedom is $v_j = v_{0j} + n_j$. The hyperparameters for the prior v_{0j} and σ_{0j}^2 can be interpreted as the degree of freedom and scale of σ_j^2, respectively.

This statistic is well known for its effectiveness in analyzing the samples having small size, but it still heavily depends on the parametric assumption.

1.4.4 Analysis of Variance

Analysis of variance (ANOVA) methods play a major role in statistical analysis in many fields of scientific investigation and now have become an important methodology in microarray studies. An ANOVA analyzes the differences in expression levels between two or more groups. It is a linear model in which all explanatory variables are categorical. The response variable is a numerical continuous variable and is in microarray typically the expression levels of a single gene.

An ANOVA basically partitions the observed variation in gene expression between the samples into components due to different explanatory variables and unexplained variation (the residual noise). It determines the significance of each of the differences between groups by comparing the differences between the groups to the variation within the groups.

Kerr et al. (2000) proposed the following model to account for multiple sources of variation in a microarray experiment:

$$\log(y_{ijkg}) = \mu + A_i + V_k + G_g + (AG)_{ig} + (VG)_{kg} + \varepsilon_{ijkg} \qquad (1.4.4.1)$$

where μ is the average overall signal, A_i represents the effect of the i^{th} array, , V_k represents the effect of the k^{th} variety, G_g represents

the effect of the g^{th} gene, $(AG)_{ig}$ represents a combination of array i and gene g (a particular spot on a particular array), and $(VG)_{kg}$ represents the interaction between the k^{th} variety and the g^{th} gene. The error terms ε_{ijkg} are assumed to be independently and identically distributed with mean zero.

Because of the fact that the distribution of residuals are notably different than normal as Kerr et al. (2000) commented, they employed bootstrap approach to obtain confidence intervals for the estimated differences in expression without relying on normality assumptions. However, the model fit and parameter estimates in their study were obtained by the method of least squares, which is most efficient for normal data, non-normality complicates the data analysis and results in inefficient estimators.

In addition to Kerr et al. (2000), Churchill (2002) and Yang and Speed (2002) discussed the experimental design issues concerning ANOVA. Wolfinger et al. (2001) proposed a two-stage approach for fitting linear models, including mixed effects model. The detailed properties of ANOVA model will be discussed in Chapter 2.

1.5 Multiple Testing

Analyzing microarray data involves performing a very large number of statistical tests, as a test is being run on each and every gene. The problem of multiple testing, also referred to as multiplicity, is the problem of having an increased number of false positive result, i.e., genes that are found to be statistically different between conditions,

but are not in reality since the same hypothesis is tested multiple times. Multiple testing corrections adjust p-values to quantify and correct for this occurrence of false positives due to multiple testing.

Multiple testing correction adjusts the individual p-value for each gene to control family-wise error rate of the overall false discovery rate. The family-wise error rate (FWER) is the probability of making false discoveries, or type I errors, among the hypotheses when performing multiple tests whereas the FDR controls the probability of having false tests among all the significant genes.

There are many different procedures to correct for multiple testing. The most important variation in these methods is how stringently they correct for the number of applied tests. The stringency is a double-edged sword because of the existing trade-off between the proportion of successfully identifying a real effect, sensitivity and the proportion of successfully rejecting a false effect, specificity.

The comparison of multiple testing procedures is out of the scope of this thesis (for a complete review, see Amaratunga and Cabrera, 2004) but it should be noted that Benjamini and Hochberg method (Benjamini and Hochberg, 1995) is facilitated for multiple testing correction in this study whenever needed. We prefer this method since simulations suggest that it is unlikely to fail (Reiner et al., 2003) for realistic scenarios and is therefore widely used as it is not too conservative (Göhlmann, 2009).

The Benjamini and Hochberg FDR is calculated as shown below:

$$p_g^{BH} = p_g \frac{G}{\text{order}(p_g)}, \; g=1,2,\ldots,G \qquad (1.5.1)$$

where p_g is the p-value corresponding to the test statistic of the g-th gene.

CHAPTER 2

UNBALANCED TWO-WAY CLASSIFICATION WITH INTERACTION

In this study, we are interested in identifying differences among levels of variety for every gene expression levels. Therefore, we consider an unbalanced two-way classification fixed-effect model with interaction for the microarray experiment. Every measurement in the experiment is associated with a combination of a variety and a gene. Let y^*_{kgl} denote the l^{th} measurement from the k^{th} variety and the g^{th} gene. Thus the used model on the log scale is

$$\log(y^*_{kgl}) = y_{kgl} = \mu + V_k + G_g + (VG)_{kg} + \varepsilon_{kgl}, \quad 1 \le k \le K, 1 \le g \le G, 1 \le l \le n_k \tag{2.1}$$

where μ is overall average signal, V_k is the effect of the k^{th} variety, G_g is the effect of the g^{th} gene, $(VG)_{kg}$ is the interaction between the k^{th} variety and the g^{th} gene, ε_{kgl} are error terms and n_k is the number of observations in the k^{th} variety for every gene.

The terms V_k account for overall differences in the varieties. Such differences could arise if some varieties have more transcription activity in general, or simply because of differential concentration of

mRNA in the labeled sample. The terms G_g account for average effects of individual genes spotted on the arrays in the experiment. The terms $(VG)_{kg}$ capture departures from the overall averages that are attributable to the specific combinations of a variety k and a gene g. Non-zero differences in variety×gene interactions across varieties for a given gene indicate differential expression.

Without loss of generality, we assume that it is a fixed effects model where

$$\sum_{k=1}^{K} V_k = \sum_{g=1}^{G} G_g = \sum_{k=1}^{K} (VG)_{kg} = \sum_{g=1}^{G} (VG)_{kg} = 0. \qquad (2.2)$$

The data are analyzed on the log scale since log transform is the natural method for analyzing data with an additive model where the effects in the data are believed to be multiplicative. The common use of ratios to analyze microarray data illustrates that this is a prevalent assumption. In fact, some tools for clustering genes based on microarray data suggest using the log transform on ratios (Eisen, 1999). Furthermore, the explanation of untransformed data and the examination of other transformations such as square-root, reciprocal, etc. conclude that the log transform is a good choice (Sapir and Churchill, 2000).

To determine the distribution of errors, we examined the normal Q-Q plots of residuals obtained by using least square estimation for real life data sets and observed that the plots generally have an 'S' shape as in the example below:

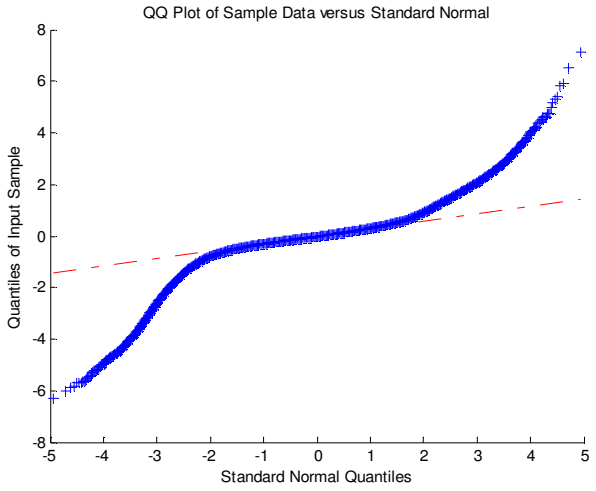

Figure 2.1 An example of Q-Q plot for a heavy tailed distribution

'S' shaped Q-Q plot indicates that the distribution of errors has heavier tails than normal distribution (Hamilton, 1992). Thus we assume that ε_{kgl} are iid and have one of the distributions in the family of long-tailed symmetric distribution.

In this chapter, parameters of the unbalanced two-way ANOVA model with interaction are estimated by using the modified maximum likelihood estimation method when the errors have a distribution from long-tailed symmetric family. The test statistics for testing the variety, gene and interaction effects are developed. Also pairwise comparisons for the interaction terms for every gene are performed. Lastly the robustness properties of the test statistics are examined.

2.1 Long-Tailed Symmetric Family

A rich family of unimodal long-tailed symmetric (LTS) distributions is given by

$$f(z) = \frac{1}{\sqrt{q}B\left(\frac{1}{2}, p-\frac{1}{2}\right)\sigma}\left(1+\frac{1}{q}z^2\right)^{-p}, \quad -\infty < z < \infty \qquad (2.1.1)$$

where $z = (x-\mu)/\sigma$, $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$, $q = 2p-3$ and $p \geq 2$. It can be easily shown that $E(Z) = 0$ and $Var(Z) = 1$. For $1 \leq p < 2$, $Var(Z)$ does not exist in which case σ is simply a location parameter. The kurtosis of the distribution is $3(p-3/2)/(p-5/2)$. It is always greater than 3 and becomes 3 since (2.1.1) reduces to N(0, 1) for $p = \infty$. Note that the distribution of $t = z\sqrt{v/q}$ has Student's t with $v = 2p-1$ degrees of freedom.

2.2 Least Squares Estimation

Consider the unbalanced two-way ANOVA model with interaction given in (2.1). To find the least squares estimators (LSEs) of μ, V_k, G_g and $(VG)_{kg}$, we form the sum squares of the errors

$$RSS = \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\varepsilon_{kgl}^2 = \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\left(y_{kgl} - \mu - V_k - G_g - (VG)_{kg}\right)^2 \qquad (2.2.1)$$

and choose the values μ, V_k, G_g and $(VG)_{kg}$, say $\tilde{\mu}, \tilde{V}_k, \tilde{G}_g$ and $(\widetilde{VG})_{kg}$ which minimize RSS. Thus the LSEs of μ, V_k, G_g and $(VG)_{kg}$ are obtained as follows:

$$\tilde{\mu} = \tilde{\mu}_{...}, \tag{2.2.2}$$

$$\tilde{V}_k = \tilde{\mu}_{k..} - \tilde{\mu}, \tag{2.2.3}$$

$$\tilde{G}_g = \tilde{\mu}_{.g.} - \tilde{\mu} \tag{2.2.4}$$

and

$$(\widetilde{VG})_{kg} = \tilde{\mu}_{kg.} - \tilde{\mu} - \tilde{V}_k - \tilde{G}_g \tag{2.2.5}$$

where

$$\tilde{\mu}_{...} = \frac{\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k} y_{kgl}}{N}, \quad \tilde{\mu}_{k..} = \frac{\sum_{g=1}^{G}\sum_{l=1}^{n_k} y_{kgl}}{Gn_k}, \quad \tilde{\mu}_{.g.} = \frac{\sum_{k=1}^{K}\sum_{l=1}^{n_k} y_{kgl}}{n_T}, \quad \tilde{\mu}_{kg.} = \frac{\sum_{l=1}^{n_k} y_{kgl}}{n_k}$$

and

$$N = Gn_T \text{ and } n_T = \sum_{k=1}^{K} n_k .$$

The terms $\tilde{\mu}_{k..}$ and $\tilde{\mu}_{.g.}$ indicate the LSEs of the factor level means and the term $\tilde{\mu}_{kg.}$ indicates the LSEs of the treatment means.

The least squares estimator of σ^2 is equal to mean square error (MSE) where

$$\tilde{\sigma}^2 = \text{MSE} = \frac{\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}(y_{kgl} - \tilde{\mu}_{kg.})^2}{N - GK} \qquad (2.2.6)$$

The variances of the estimators $\tilde{\mu}, \tilde{V}_k, \tilde{G}_g$ and $(\widetilde{VG})_{kg}$ are also obtained as follows:

$$\text{Var}(\tilde{\mu}) = \frac{\sigma^2}{N}, \qquad (2.2.7)$$

$$\text{Var}(\tilde{V}_k) = \frac{(n_T - n_k)\sigma^2}{Nn_k}, \qquad (2.2.8)$$

$$\text{Var}(\tilde{G}_g) = \frac{(G-1)\sigma^2}{N} \qquad (2.2.9)$$

and

$$\text{Var}\left((\widetilde{VG})_{kg}\right) = \frac{\sigma^2(N - n_T - Gn_k)}{Nn_k} \qquad (2.2.10)$$

Unbiased estimators of these variances are obtained by replacing σ^2 with MSE.

2.3 Maximum Likelihood Estimation

As mentioned at the beginning of this chapter, we assume that ε_{kgl} are iid and have one of the distributions in the family of long-tailed symmetric distribution for the model given in (2.1).

The Fisher likelihood function is

$$L = C\frac{1}{\sigma^N}\prod_{g=1}^{G}\prod_{k=1}^{K}\prod_{l=1}^{n_k}\left(1+\frac{1}{q}z_{kgl}^2\right)^{-p} \tag{2.3.1}$$

where

$$C = \left(\sqrt{q}B\left(\frac{1}{2}, p-\frac{1}{2}\right)\right)^{-N} \text{ and } z_{kgl} = \frac{\varepsilon_{kgl}}{\sigma} = \frac{y_{kgl}-\mu-V_k-G_g-(VG)_{kg}}{\sigma}.$$

Thus, the likelihood equations for estimating μ, V_k, G_g, $(VG)_{kg}$ $(1 \le k \le K, 1 \le g \le G)$ and σ^2 are

$$\frac{\partial \ln L}{\partial \mu} = \frac{2p}{q\sigma}\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k} g(z_{kgl}) = 0, \tag{2.3.2}$$

$$\frac{\partial \ln L}{\partial V_k} = \frac{2p}{q\sigma}\sum_{g=1}^{G}\sum_{l=1}^{n_k} g(z_{kgl}) = 0, \tag{2.3.3}$$

$$\frac{\partial \ln L}{\partial G_g} = \frac{2p}{q\sigma}\sum_{k=1}^{K}\sum_{l=1}^{n_k} g(z_{kgl}) = 0, \tag{2.3.4}$$

$$\frac{\partial \ln L}{\partial (VG)_{kg}} = \frac{2p}{q\sigma}\sum_{l=1}^{n_k} g(z_{kgl}) = 0 \qquad (2.3.5)$$

and

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{N}{\sigma} + \frac{2p}{q\sigma}\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k} z_{kgl}\, g(z_{kgl}) = 0 \qquad (2.3.6)$$

where

$$g(z_{kgl}) = \frac{z_{kgl}}{1 + \frac{1}{q} z_{kgl}^2}\,.$$

Likelihood equations given above have no explicit solutions since they include non-linear function $g(z_{kgl})$. Solving the non-linear equations in (2.3.2)- (2.3.6) by iteration is enormously problematic (Puthenpura and Sinha, 1986; Akkaya and Tiku, 2008a; Islam and Tiku, 2004). For example, the iterations may never converge or converge to wrong values (e.g., they correspond to local rather than global maximum of L). Moreover, there are too many equations to iterate simultaneously which is formidable task. Also, it is difficult to make any analytical study of the resulting maximum likelihood estimators (MLEs), especially for small samples. Therefore, method of modified maximum likelihood (MML) developed by Tiku (1967) is used to find the explicit solutions.

2.4 Modified Maximum Likelihood Estimation

Tiku and Suresh (1992) introduced modified maximum likelihood estimation for location-scale models with the following properties:

1. The estimates are explicit functions of sample observations and are easier to compute than the maximum likelihood estimates. Also their properties are simple to determine (Vaughan and Tiku, 2000).

2. It is asymptotically equivalent to maximum likelihood when regularity conditions hold. Thus, asymptotically the MML estimators are fully efficient, i.e., they are unbiased and their variances are equal to the minimum variance bounds (Tiku et al., 1986, Vaughan and Tiku, 2000 and Bhattacharyya, 1985).

3. The estimates are almost fully efficient, that is, they have no or negligible bias and their variances are only marginally bigger than the Minimum Variance Bounds (MVBs) even for small samples (Lee et al., 1980; Vaughan, 1992a, Tiku et al., 1986; Smith et al., 1973; Tan, 1985).

4. The method is essentially self-censoring since it assigns small weights to extremes.

Tiku's modified maximum likelihood methodology proceeds in three steps as follows:

1. Express the likelihood equations in terms of ordered variates,

2. linerarize the intractable terms in the likelihood equations by using the first two terms of the Taylor series expansion and

3. solve the resulting equations to get the modified maximum likelihood estimators.

For the model given in (2.1), let $y_{kg(1)} \le y_{kg(2)} \le \ldots \le y_{kg(n_k)}$ be the order statistics for the n_k observations y_{kgl} $(1 \le l \le n_k)$ in the $(g, k)^{th}$ cell.

Then

$$z_{kg(l)} = \frac{y_{kg(l)} - \mu - V_k - G_g - (VG)_{kg}}{\sigma} \quad (1 \le k \le K, 1 \le g \le G) \qquad (2.4.1)$$

are the ordered z_{kgl} $(1 \le l \le n_k)$ variates.

In this method, the likelihood equations given in (2.3.2)-(2.3.6) are expressed in terms of the ordered variates $z_{kg(l)}$. Since summations are invariant to ordering, the resulting likelihood equations are written as follows:

$$\frac{\partial \ln L}{\partial \mu} = \frac{2p}{q\sigma} \sum_{g=1}^{G} \sum_{k=1}^{K} \sum_{l=1}^{n_k} g(z_{kg(l)}) = 0 , \qquad (2.4.2)$$

$$\frac{\partial \ln L}{\partial V_k} = \frac{2p}{q\sigma} \sum_{g=1}^{G} \sum_{l=1}^{n_k} g(z_{kg(l)}) = 0 , \qquad (2.4.3)$$

$$\frac{\partial \ln L}{\partial G_g} = \frac{2p}{q\sigma}\sum_{k=1}^{K}\sum_{l=1}^{n_k} g(z_{kg(l)}) = 0\,, \tag{2.4.4}$$

$$\frac{\partial \ln L}{\partial (VG)_{kg}} = \frac{2p}{q\sigma}\sum_{l=1}^{n_k} g(z_{kg(l)}) = 0 \tag{2.4.5}$$

and

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{N}{\sigma} + \frac{2p}{q\sigma}\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k} z_{kg(l)} g(z_{kg(l)}) = 0\,. \tag{2.4.6}$$

Since the function $g(z)$ is almost linear in small intervals $a < z < b$ (Tiku, 1967, 1968) and $z_{kg(l)}$ is located in the vicinity of $t_{kg(l)} = E(z_{kg(l)})$ at any rate for large n_k, an appropriate linear approximation for $g(z_{kg(l)})$ is obtained by using the first two terms of a Taylor series expansion, namely

$$\begin{aligned} g(z_{kg(l)}) &\cong g(t_{kg(l)}) + g'(t_{kg(l)})(z_{kg(l)} - t_{kg(l)}) \\ &\cong \alpha_{kg(l)} + \delta_{kg(l)} z_{kg(l)} \quad (1 \le k \le K, 1 \le g \le G, 1 \le l \le n_k) \end{aligned} \tag{2.4.7}$$

where $t_{kg(l)} = E(z_{kg(l)})$ is the expected value of the l^{th} order statistic $z_{kg(l)}$ in the g^{th} gene and k^{th} variety, $\alpha_{kg(l)} = g(t_{kg(l)}) - \delta_{kg(l)} t_{kg(l)}$ and $\delta_{kg(l)} = g'(t_{kg(l)})$. Here,

$$\alpha_{kg(l)} = \frac{(2/q)t_{kg(l)}^3}{\left(1 + (1/q)t_{kg(l)}^2\right)^2} \quad \text{and} \quad \delta_{kg(l)} = \frac{1 - (1/q)t_{kg(l)}^2}{\left(1 + (1/q)t_{kg(l)}^2\right)^2}\,. \tag{2.4.8}$$

Tables of $t_{kg(l)}$, the variances of $z_{kg(l)}$ and the covariances of $(z_{kg(l)}, z_{kg(j)})$ are given in Tiku and Kumra (1981) for $p = 2(0.5)10$ and $n \le 20$. For $n \ge 10$, $t_{kg(l)}$ may be approximated from the following equality:

$$\frac{1}{\sqrt{k}B(1/2, p-1/2)} \int_{-\infty}^{t_{kg(l)}} \left(1 + \frac{z^2}{k}\right)^{-p} dz = \frac{i}{(n+1)} \quad (1 \le i \le n) \tag{2.4.9}$$

A MATLAB subroutine is available to evaluate (2.4.9).

Incorporating (2.4.7) into (2.4.2)-(2.4.6) the following modified likelihood equations are obtained:

$$\frac{\partial \ln L}{\partial \mu} \cong \frac{\partial \ln L^*}{\partial \mu} = \frac{2p}{q\sigma} \sum_{g=1}^{G} \sum_{k=1}^{K} \sum_{l=1}^{n_k} (\alpha_{kg(l)} + \delta_{kg(l)} z_{kg(l)}) = 0, \tag{2.4.10}$$

$$\frac{\partial \ln L}{\partial V_k} \cong \frac{\partial \ln L^*}{\partial V_k} = \frac{2p}{q\sigma} \sum_{g=1}^{G} \sum_{l=1}^{n_k} (\alpha_{kg(l)} + \delta_{kg(l)} z_{kg(l)}) = 0, \tag{2.4.11}$$

$$\frac{\partial \ln L}{\partial G_g} \cong \frac{\partial \ln L^*}{\partial G_g} = \frac{2p}{q\sigma} \sum_{k=1}^{K} \sum_{l=1}^{n_k} (\alpha_{kg(l)} + \delta_{kg(l)} z_{kg(l)}) = 0, \tag{2.4.12}$$

$$\frac{\partial \ln L}{\partial (VG)_{kg}} \cong \frac{\partial \ln L^*}{\partial (VG)_{kg}} = \frac{2p}{q\sigma} \sum_{l=1}^{n_k} (\alpha_{kg(l)} + \delta_{kg(l)} z_{kg(l)}) = 0 \tag{2.4.13}$$

and

$$\frac{\partial \ln L}{\partial \sigma} \cong \frac{\partial \ln L^*}{\partial \sigma} = -\frac{N}{\sigma} + \frac{2p}{q\sigma} \sum_{g=1}^{G} \sum_{k=1}^{K} \sum_{l=1}^{n_k} z_{kg(l)} (\alpha_{kg(l)} + \delta_{kg(l)} z_{kg(l)}) = 0. \tag{2.4.14}$$

These equations are asymptotically equivalent to the corresponding likelihood equations (2.4.2)-(2.4.6) and their solutions yield the following MML estimators:

$$\hat{\mu} = \hat{\mu}_{...}, \tag{2.4.15}$$

$$\hat{V}_k = \hat{\mu}_{k..} - \hat{\mu}, \tag{2.4.16}$$

$$\hat{G}_g = \hat{\mu}_{.g.} - \hat{\mu}, \tag{2.4.17}$$

$$(\widehat{VG})_{kg} = \hat{\mu}_{kg.} - \hat{\mu} - \hat{V}_k - \hat{G}_g \tag{2.4.18}$$

and

$$\hat{\sigma} = \frac{-B + \sqrt{B^2 + 4NC}}{2\sqrt{N(N - KG)}} \tag{2.4.19}$$

where

$$\hat{\mu}_{...} = \frac{\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\delta_{kg(l)}y_{kg(l)}}{\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\delta_{kg(l)}}, \quad \hat{\mu}_{k..} = \frac{\sum_{g=1}^{G}\sum_{l=1}^{n_k}\delta_{kg(l)}y_{kg(l)}}{\sum_{g=1}^{G}\sum_{l=1}^{n_k}\delta_{kg(l)}}, \quad \tilde{\mu}_{.g.} = \frac{\sum_{k=1}^{K}\sum_{l=1}^{n_k}\delta_{kg(l)}y_{kg(l)}}{\sum_{k=1}^{K}\sum_{k=1}^{n_k}\delta_{kg(l)}},$$

$$\hat{\mu}_{kg.} = \frac{\sum_{l=1}^{n_k}\delta_{kg(l)}y_{kg(l)}}{\sum_{l=1}^{n_k}\delta_{kg(l)}}, \quad B = \frac{2p}{q}\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\alpha_{kg(l)}(y_{kg(l)} - \hat{\mu}_{kg.}) \quad \text{and}$$

$$C = \frac{2p}{q}\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\delta_{kg(l)}(y_{kg(l)} - \hat{\mu}_{kg.})^2 .$$

$\hat{\sigma}$ is the bias-corrected estimator of σ. The estimators are explicit functions of sample observations and, therefore easy to compute.

It may be noted that the coefficients $\delta_{kg(l)}$ increase until the middle value and then decrease in a symmetric fashion. If $\delta_{kg(l)} > 0$, then all the remaining $\delta_{kg(l)}$ coefficients are positive. As a consequence, $\hat{\sigma}$ is real and positive. For small p and large sample sizes, however, $\delta_{kg(l)}$ can be negative as a result of which $\hat{\sigma}$ can cease to be real. Thus, if C in (2.4.18) assumes a negative value, we calculate the MML estimators from the sample by replacing $\alpha_{kg(l)}$ and $\delta_{kg(l)}$ by $\alpha^{*}_{kg(l)}$ and $\delta^{*}_{kg(l)}$, respectively (Islam and Tiku, 2004):

$$\alpha^{*}_{kg(l)} = \frac{(1/q)t^3_{kg(l)}}{\left(1+(1/q)t^2_{kg(l)}\right)^2} \qquad (2.4.20)$$

and

$$\delta^{*}_{kg(l)} = \frac{1}{\left(1+(1/q)t^2_{kg(l)}\right)^2} . \qquad (2.4.21)$$

Corollary 2.4.1: Asymptotically, the estimator $\hat{\mu}$ is the MVB estimator μ and is normally distributed with variance

$$\mathrm{Var}(\tilde{\mu}) \cong \frac{q\sigma^2}{2p\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\delta_{kg(l)}}. \qquad (2.4.22)$$

Corollary 2.4.2: Asymptotically, the estimator $\hat{V}_k$ is the MVB estimator of V_k and is normally distributed with variance

$$\mathrm{Var}(\hat{V}_k) = \frac{q\sigma^2}{2p\sum_{g=1}^{G}\sum_{l=1}^{n_k}\delta_{kg(l)}}. \qquad (2.4.23)$$

Corollary 2.4.3: Asymptotically, the estimator $\hat{G}_g$ is the MVB estimator of G_g and is normally distributed with variance

$$\mathrm{Var}(\hat{G}_g) = \frac{q\sigma^2}{2p\sum_{k=1}^{K}\sum_{l=1}^{n_k}\delta_{kg(l)}}. \qquad (2.4.24)$$

Corollary 2.4.4: Asymptotically, the estimator $(\widehat{VG})_{kg}$ is the MVB estimator of VG_{kg} and is normally distributed with variance

$$\mathrm{Var}\left((\widehat{VG})_{kg}\right) = \frac{q\sigma^2}{2p\sum_{l=1}^{n_k}\delta_{kg(l)}}. \qquad (2.4.25)$$

Lemma 2.4.1: Asymptotically, $\frac{N\hat{\sigma}^2}{\sigma^2}$ is distributed as chi-square with N-GK degrees of freedom.

2.4.1 Efficiency Properties

The estimator $\hat{\mu}$ is unbiased, in fact, it is asymptotically MVB estimator of μ, and is normally distributed. Therefore, $\hat{\mu}$ is best asymptotically normal (BAN) estimator. The MVB for estimating μ is as follows:

$$\text{MVB}(\mu) = \frac{\sigma^2 q(p+1)}{2Np(p-1/2)} \qquad (2.4.1.1)$$

The estimator $\hat{V}_k$ is unbiased, in fact, it is asymptotically MVB estimator of V_k, and is normally distributed. Therefore, $\hat{V}_k$ is the BAN estimator. The MVB for estimating V_k is as follows:

$$\text{MVB}(V_k) = \frac{\sigma^2 q(p+1)}{2Gn_k p(p-1/2)} \qquad (2.4.1.2)$$

The estimator $\hat{G}_g$ is unbiased, in fact, it is asymptotically MVB estimator of G_g, and is normally distributed. Therefore, $\hat{G}_g$ is the BAN estimator. The MVB for estimating G_g is as follows:

$$\text{MVB}(G_g) = \frac{\sigma^2 q(p+1)}{2n_T p(p-1/2)} \qquad (2.4.1.3)$$

The estimator $\widehat{(VG)}_{kg}$ is unbiased, in fact, it is asymptotically MVB estimator of $(VG)_{kg}$, and is normally distributed. Therefore, $\widehat{(VG)}_{kg}$ is the BAN estimator. The MVB for estimating $(VG)_{kg}$ is as follows:

$$\text{MVB}((VG)_{kg}) = \frac{\sigma^2 q(p+1)}{2n_k p(p-1/2)} \qquad (2.4.1.4)$$

The estimator of $\hat{\sigma}^2$ is asymptotically MVB estimator of σ^2 and is distributed as a multiple of chi-square; see Lemma 2.4.1. The MVB for estimating σ is as follows:

$$\text{MVB}(\sigma) = \frac{\sigma^2(p+1)}{N(2p-1)} \qquad (2.4.1.5)$$

To examine the properties of the estimators used in ANOVA under long-tailed symmetric distribution, the means and the variances of LS and MML estimators of μ, V_k, G_g and $(VK)_{kg}$ are simulated based on 100,000/n Monte Carlo runs where $k = 2$, $n_1 = n_2 = n$ and $G = 2000$.

Given in Table 2.1, are the simulated means of LS and MML estimators of μ, V_1, G_1, $(VK)_{11}$ and σ. To decide whether MML estimators are unbiased or not, the simulated means of MML estimates $\hat{\mu}$, $\hat{V}_1$, $\hat{G}_1$, $(\widehat{VG})_{11}$ and $\hat{\sigma}$ are compared with the simulated means of $\tilde{\mu}$, $\tilde{V}_1$, $\tilde{G}_1$, $(\widetilde{VG})_{11}$ and $\tilde{\sigma}$ obtained by using LS method which are known as unbiased estimators. Table 2.1 shows that the simulated means of $\hat{\mu}$, $\hat{V}_1$, $\hat{G}_1$, $(\widehat{VG})_{11}$ and $\hat{\sigma}$ obtained by using MML estimators are almost same with those of $\tilde{\mu}$, $\tilde{V}_1$, $\tilde{G}_1$, $(\widetilde{VG})_{11}$ and $\tilde{\sigma}$ obtained by using LS estimators and, therefore, MML estimators are unbiased.

Table 2.1 Means of the LS and MML estimators of;

(1) μ (2) V_1 (3) G_1 (4) $(VG)_{11}$ (5) σ

	p:	2	2.5	3.5	5	10
n_k=5	(1) MML	5.0923	5.3332	5.3863	5.3100	5.4521
	LS	5.0923	5.3332	5.3863	5.3100	5.4521
	(2) MML	0.5727	0.3466	0.1418	0.4505	0.1781
	LS	0.5727	0.3466	0.1418	0.4505	0.1781
	(3) MML	6.3507	2.9148	3.9250	-1.8241	5.3914
	LS	6.3518	2.9149	3.9251	-1.8238	5.3914
	(4) MML	0.2530	-0.5959	-0.4419	0.2447	0.6379
	LS	0.2525	-0.5953	-0.4418	0.2447	0.6379
	(5) MML	0.0102	0.0058	0.0526	0.2015	0.3063
	LS	0.0101	0.0058	0.0524	0.2015	0.3064
n_k=10	(1) MML	5.1495	5.4801	5.2969	5.3657	5.2911
	LS	5.1495	5.4801	5.2969	5.3657	5.2911
	(2) MML	0.8032	0.8296	0.7725	0.7635	0.1856
	LS	0.8032	0.8296	0.7725	0.7635	0.1856
	(3) MML	-0.6993	4.3040	-0.5803	3.2440	-2.8542
	LS	-0.6991	4.3044	-0.5805	3.2435	-2.8541
	(4) MML	-0.8662	-0.0621	0.7457	0.2363	0.5101
	LS	-0.8669	-0.0624	0.7456	0.2360	0.5101
	(5) MML	0.1053	0.2058	0.0122	0.6212	0.5898
	LS	0.1053	0.2055	0.0120	0.6211	0.5898
n_k=50	(1) MML	5.2231	5.0672	5.1407	5.0661	5.4154
	LS	5.2231	5.0672	5.1407	5.0661	5.4154
	(2) MML	0.8769	0.4096	0.2652	0.1195	0.6996
	LS	0.8769	0.4096	0.2652	0.1195	0.6996
	(3) MML	1.3944	1.6534	0.5215	-0.8554	-1.7499
	LS	1.3945	1.6536	0.5213	-0.8554	-1.7498
	(4) MML	-0.0906	-0.5839	-0.0874	0.7543	0.3350
	LS	-0.0906	-0.5845	-0.0876	0.7541	0.3351
	(5) MML	0.1026	0.5265	0.8900	0.9578	0.3205
	LS	0.1026	0.5264	0.8900	0.9576	0.3205

Table 2.2 gives the relative efficiencies of LS estimators, $\tilde{\mu}$, $\tilde{V}_1$, $\tilde{G}_1$, $(\widetilde{VG})_{11}$ and $\tilde{\sigma}$ for 100,000/n Monte Carlo runs where $k = 2$, $n_1 = n_2 = n$ and $G = 2000$. The table indicates that the MML estimators $\hat{\mu}$, $\hat{V}_1$, $\hat{G}_1$, $(\widehat{VG})_{11}$ and $\hat{\sigma}$ are considerably more efficient even for small sample sizes. Note that for $p = 10$, the LS estimators are almost as efficient as MML estimators. This is an expected result since the long-tailed symmetric distribution reduces to a normal for $p = \infty$. For small p values which are more appropriate for heavy-tailed microarray data, MML estimators are enormously more efficient than LS estimators. The relative efficiencies of LS estimators, $\tilde{\mu}$, $\tilde{V}_1$, $\tilde{G}_1$, $(\widetilde{VG})_{11}$ and $\tilde{\sigma}$ decreases as sample size increases, called as disconcerting feature.

Table 2.2 Relative efficiencies of the LS estimators of;

(1) μ (2) V_1 (3) G_1 (4) $(VG)_{11}$ (5) σ

	p:	2	2.5	3.5	5	10
n_k=5	(1)	68.654	86.065	94.932	97.377	99.800
	(2)	70.489	86.220	94.880	98.200	99.596
	(3)	72.667	84.522	94.711	98.702	99.455
	(4)	72.055	84.906	94.321	97.182	99.701
	(5)	70.660	75.890	84.659	95.486	98.566
n_k=10	(1)	57.156	76.405	90.527	96.467	99.620
	(2)	54.463	76.747	91.423	95.876	98.838
	(3)	45.729	73.862	90.339	96.806	99.076
	(4)	45.548	76.758	92.544	96.145	98.977
	(5)	40.587	65.875	80.569	93.890	96.400
n_k=20	(1)	48.884	72.980	85.903	94.104	98.455
	(2)	52.903	70.651	87.167	95.387	98.740
	(3)	45.220	72.084	86.424	95.415	99.012
	(4)	44.628	70.259	88.202	95.509	98.565
	(5)	38.748	63.524	76.480	91.475	95.488
n_k=50	(1)	47.003	70.380	84.069	94.052	98.405
	(2)	52.261	70.108	86.855	94.743	98.689
	(3)	45.057	69.767	85.914	94.270	98.668
	(4)	44.073	68.831	86.174	92.850	98.428
	(5)	34.586	60.934	75.086	90.001	94.701

2.4.2 Testing Main and Interaction Effects

Consider the model given in (2.1) and assume that errors are normally and independently distributed with mean zero and variance σ^2. Then, the observations y_{kgl}'s are also normally and independently distributed with mean $\mu + V_k + G_g + (VG)_{kg}$ and variance σ^2.

To test the equality of the main and interaction effects,

$$H_{01} : V_1 = V_2 = \ldots = V_k = 0 ,$$

$$H_{02} : G_1 = G_2 = \ldots = G_g = 0$$

and

$$H_{03} : (VG)_{kg} = 0 \text{ for all } k = 1, 2, .., K \text{ and } g = 1, 2, .., G , \quad (2.4.2.1)$$

the analysis of variance procedure is used. This procedure partitions the total sum of squares which is the measure of total variability of the y_{kgl} observations. The total sum of squares denoted by SS_T is

$$SS_T = \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}(y_{kgl} - \tilde{\mu}_{...})^2 . \quad (2.4.2.2)$$

Total sum of squares can be decomposed as follows:

$$\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}(y_{kgl} - \tilde{\mu}_{...})^2 = \sum_{g=1}^{G}\sum_{k=1}^{K} n_k(\tilde{\mu}_{kg.} - \tilde{\mu}_{...})^2 + \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}(y_{kgl} - \tilde{\mu}_{kg.})^2 \quad (2.4.2.3)$$

The first term on the right is the treatment sum of squares denoted by SS_{Tr} and the second term is the error sum of squares denoted by SS_E. SS_{Tr} reflects the variability between the KG treatment means and SS_E reflects the variability within treatments (Neter et al., 1985).

The breakdown of treatment sum of squares is given by

$$SS_{Tr} = SS_V + SS_G + SS_{VG} \quad (2.4.2.4)$$

where

$$SS_V = G\sum_{k=1}^{K} n_k(\tilde{\mu}_{k..} - \tilde{\mu}_{...})^2 ,$$

$$SS_G = n_T\sum_{g=1}^{G}(\tilde{\mu}_{.g.} - \tilde{\mu}_{...})^2$$

and

$$SS_{VG} = \sum_{g=1}^{G}\sum_{k=1}^{K} n_k(\tilde{\mu}_{kg.} - \tilde{\mu}_{k..} - \tilde{\mu}_{.g.} + \tilde{\mu}_{...})^2 .$$

SS_V, called the factor V sum of squares, measures the variability of the estimated V factor level means $\tilde{\mu}_{k..}$. Similarly SS_G, called the factor G sum of squares, measures the variability of the estimated G factor level means $\tilde{\mu}_{.g.}$. Finally, SS_{VG}, called the VG interaction sum of squares, measures the variability of the estimated interactions $\tilde{\mu}_{kg.} - \tilde{\mu}_{k..} - \tilde{\mu}_{.g.} + \tilde{\mu}_{...}$ for KG treatments.

According to Cochran's theorem, $\frac{SS_V}{\sigma^2}$, $\frac{SS_G}{\sigma^2}$ and $\frac{SS_{VG}}{\sigma^2}$ are independent and have chi-square distributions with (K-1), (G-1) and (K-1)(G-1) degrees of freedom, respectively.

The F statistics based on the LSEs of the parameters in (2.2.2)-(2.2.5) for testing H_{01}, H_{02} and H_{03}, respectively are given by

$$F_1 = \frac{SS_V/(K-1)}{SS_E/(N-KG)} = \frac{G\sum_{k=1}^{K} n_k \tilde{V}_k^2}{(K-1)\tilde{\sigma}^2}, \qquad (2.4.2.5)$$

$$F_2 = \frac{SS_G/(G-1)}{SS_E/(N-KG)} = \frac{n_T \sum_{g=1}^{G} \tilde{G}_g^2}{(G-1)\tilde{\sigma}^2} \qquad (2.4.2.6)$$

and

$$F_3 = \frac{SS_{VG}/\big((K-1)(G-1)\big)}{SS_E/(N-KG)} = \frac{\sum_{g=1}^{G}\sum_{k=1}^{K} n_k (\tilde{VG})_{kg}^2}{(K-1)(G-1)\tilde{\sigma}^2}. \qquad (2.4.2.7)$$

Under the null hypotheses, the distributions of F_1, F_2 and F_3 are central F with degrees of freedom (K-1, N-KG), (G-1, N-KG) and $\big((K-1)(G-1), N-KG\big)$, respectively. Large values of F_1, F_2 and F_3 lead to the rejection of H_{01}, H_{02} and H_{03}, respectively.

If the null hypotheses are not true, the distributions of F_1, F_2 and F_3 are non-central F with the same degrees of freedom and non-centrality parameters

$$\lambda_{F_1}^2 = \frac{G\sum_{k=1}^{K} n_k V_k^2}{\sigma^2}, \qquad (2.4.2.8)$$

$$\lambda_{F_2}^2 = \frac{n_T \sum_{g=1}^{G} G_g^2}{\sigma^2} \qquad (2.4.2.9)$$

and

$$\lambda^2_{F_3} = \frac{\sum_{g=1}^{G}\sum_{k=1}^{K}(VG)^2_{kg}}{\sigma^2}, \qquad (2.4.2.10)$$

respectively (Akkaya and Tiku, 2004).

Under the normality assumption, the F statistics provides the most powerful test of the null hypotheses. Under non-normality, their Type I errors are generally not much different than those under normality but their powers are adversely affected.

Since the error terms have a distribution from the LTS family in this study, we extend the hypothesis testing technique to non-normal distributions by adopting the methodology of modified likelihood.

By using the MML estimators of the model parameters in (2.4.14)-(2.4.18), we obtain the decomposition of the total sum of squares such that

$$SS_T = SS_V + SS_G + SS_{VG} + SS_E \qquad (2.4.2.11)$$

where

$$SS_T = \frac{2p}{q}\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\delta_{kg(l)}(y_{kg(l)} - \hat{\mu}_{...})^2 ,$$

$$SS_V = \frac{2p}{q}\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\delta_{kg(l)}(\hat{\mu}_{k..} - \hat{\mu}_{...})^2 ,$$

$$SS_G = \frac{2p}{q}\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\delta_{kg(l)}(\hat{\mu}_{.g.} - \hat{\mu}_{...})^2 ,$$

$$SS_{VG} = \frac{2p}{q}\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\delta_{kg(l)}(\hat{\mu}_{kg.} - \hat{\mu}_{k..} - \hat{\mu}_{.g.} + \hat{\mu}_{...})^2$$

and

$$SS_E = \frac{2p}{q}\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\delta_{kg(l)}(y_{kg(l)} - \hat{\mu}_{kg.})^2 .$$

For large sample sizes, we have $SS_E \cong N\hat{\sigma}^2$.

To test the null hypotheses in (2.4.2.1), the test statistics are given by

$$W_1 = \frac{SS_V/(K-1)}{SS_E/(N-KG)} \cong \frac{\frac{2p}{q}\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\delta_{kg(l)}(\hat{V}_k)^2}{(K-1)\hat{\sigma}^2}, \qquad (2.4.2.12)$$

$$W_2 = \frac{SS_G/(G-1)}{SS_E/(N-KG)} \cong \frac{\frac{2p}{q}\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\delta_{kg(l)}(\hat{G}_g)^2}{(G-1)\hat{\sigma}^2}, \qquad (2.4.2.13)$$

and

$$W_3 = \frac{SS_{VG}/((K-1)(G-1))}{SS_E/(N-KG)} \cong \frac{\frac{2p}{q}\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\delta_{kg(l)}(\hat{V}G_{kg})^2}{(K-1)(G-1)\hat{\sigma}^2}, \qquad (2.4.2.14)$$

respectively. For large sample sizes, their null distributions are central F with degrees of freedom (K-1, N-KG), (G-1, N-KG) and $((K-1)(G-1), N-KG)$, respectively.

If the null hypotheses are not true, the distributions of W_1, W_2 and W_3 have non-central F with the same degrees of freedom and non-centrality parameters

$$\lambda^2_{W_1} = \frac{\frac{2p}{q}\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\delta_{kg(l)}(V_k)^2}{\sigma^2}, \qquad (2.4.2.15)$$

$$\lambda^2_{W_2} = \frac{\frac{2p}{q}\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\delta_{kg(l)}(G_g)^2}{\sigma^2} \qquad (2.4.2.16)$$

and

$$\lambda^2_{W_3} = \frac{\frac{2p}{q}\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\delta_{kg(l)}(VG_{kg})^2}{\sigma^2}, \qquad (2.4.2.17)$$

respectively for large sample sizes.

Since $\lambda^2_{W_i} > \lambda^2_{F_i}$ $(i = 1, 2, 3)$, the W-test is more powerful than F-test. This is expected since more efficient estimators are used in W-test.

The simulated values of the power values of the W and F-tests are given in Table 2.3 for 100,000/n Monte Carlo runs where $k = 2$, $n_1 = n_2 = n$ and $G = 2000$. For detectable difference $d = 0$, the power reduces to Type I error. The presumed value of Type I error is 0.05

Table 2.3 Values of the power of F and W-tests for;

(1) V (2) G (3) VG

p	d:	0.00	0.25	0.50	0.75	1.00
2	(1) W	0.047	0.752	0.951	0.975	0.999
	F	0.062	0.595	0.701	0.903	0.985
	(2) W	0.047	0.751	0.952	0.974	0.998
	F	0.061	0.704	0.702	0.905	0.986
	(3) W	0.044	0.776	0.960	0.982	0.999
	F	0.049	0.590	0.682	0.852	0.945
2.5	(1) W	0.047	0.750	0.949	0.962	0.997
	F	0.060	0.599	0.712	0.913	0.987
	(2) W	0.046	0.740	0.948	0.961	0.997
	F	0.060	0.609	0.713	0.915	0.987
	(3) W	0.045	0.770	0.957	0.978	0.998
	F	0.047	0.599	0.695	0.884	0.949
3.5	(1) W	0.048	0.732	0.935	0.951	0.996
	F	0.059	0.600	0.825	0.925	0.989
	(2) W	0.048	0.738	0.934	0.950	0.995
	F	0.059	0.605	0.824	0.924	0.988
	(3) W	0.045	0.768	0.949	0.965	0.997
	F	0.046	0.604	0.809	0.907	0.956
5.0	(1) W	0.049	0.701	0.906	0.948	0.993
	F	0.053	0.651	0.853	0.928	0.990
	(2) W	0.049	0.702	0.905	0.948	0.994
	F	0.054	0.650	0.855	0.928	0.990
	(3) W	0.051	0.742	0.910	0.952	0.995
	F	0.052	0.641	0.828	0.910	0.982
10.0	(1) W	0.051	0.695	0.896	0.932	0.992
	F	0.052	0.660	0.878	0.930	0.992
	(2) W	0.051	0.699	0.896	0.935	0.991
	F	0.053	0.682	0.875	0.933	0.990
	(3) W	0.050	0.710	0.900	0.942	0.990
	F	0.053	0.697	0.854	0.935	0.989

Table 2.3 indicates that W-test has a double advantage. Both it has smaller Type I error and it is clearly more powerful than the traditional F-test (even for approximately normal distribution when p=10).

2.4.3 Comparisons of Treatment Effects

In microarray studies, the variety and gene effects are generally not of interest, but account for sources of variation in microarray data. The effects of interest in model (2.1) are the interactions between varieties and genes, $(VG)_{kg}$ since these reflect the differential expression of genes across varieties.

If the ANOVA test obtained for gene×variety interaction effects lead to rejection of the null hypothesis, comparison of the treatments means across variety levels for a given gene are be of interest (Neter et al., 1985).

Thus, we deal with comparing the treatment means μ_{kg} and construct the hypotheses as follows:

$$H_0 : \mu_{ig.} - \mu_{jg.} = 0 \quad (i, j = 1, 2, \ldots, K, i \neq j, g = 1, 2, \ldots, G) \quad (2.4.3.1)$$
$$H_1 = \mu_{ig.} - \mu_{ig.} \neq 0$$

Since we compare the treatment means across all varieties for every gene separately, we can fit a one-way ANOVA model for every gene.

Under the normality assumption, the Tukey multiple comparison procedure may be used to test the pairwise equality of the treatment means for every gene. According to Tukey method, the test statistic used for the hypothesis in (2.4.3.1) is given by

$$t = \frac{(\tilde{\mu}_{ig.} - \tilde{\mu}_{jg.}) - (\mu_{ig.} - \mu_{jg.})}{\sqrt{Var(\tilde{\mu}_{ig.}) + Var(\tilde{\mu}_{jg.})}} \qquad (2.4.3.2)$$

where

$$Var(\tilde{\mu}_{ig.}) = \frac{MSE}{n_i}.$$

If $|t| \geq \frac{1}{\sqrt{2}} q\left(1-\alpha; K, \sum_{k=1}^{K} n_k - K\right)$, we reject the null hypothesis in (2.4.3.1) with the α level of significance. Here, $q\left(1-\alpha; K, \sum_{k=1}^{K} n_k - K\right)$ is the upper α percentage point of the studentized range distribution.

Under non-normality, Dunnett (1982) gives $\max\left|\tilde{\tilde{t}}_{ijg}\right|$ as a test statistic for pairwise multiple comparisons where

$$\tilde{\tilde{t}}_{ijg} = \frac{(\tilde{\tilde{\mu}}_{ig.} - \tilde{\tilde{\mu}}_{jg.}) - (\mu_{ig.} - \mu_{jg.})}{\sqrt{\frac{\tilde{\tilde{\sigma}}^2_{ig.}}{\tilde{n}_i} + \frac{\tilde{\tilde{\sigma}}^2_{jg.}}{\tilde{n}_j}}}. \qquad (2.4.3.3)$$

Here, $\tilde{\tilde{\mu}}_{ig.}$ is the robust estimate of location for the i^{th} sample taken for the g^{th} gene, $\tilde{\tilde{\sigma}}^2_{ig.}$ is the corresponding robust estimate of variance, $\tilde{n}_i$ is the effective sample size.

If $\max\left|\tilde{\tilde{t}}_{ijg}\right| \geq A^*_{ijg,\alpha,K}$, the null hypothesis in (2.4.3.1) is rejected at the level of significance α. To determine the value of $A^*_{ijg,\alpha,K}$, the α-point of the distribution of $\max\left|\tilde{\tilde{t}}_{ijg}\right|$ is required, however, its distribution is different from the distribution of $\tilde{\tilde{t}}_{ijg}$ since $\max\left|\tilde{\tilde{t}}_{ijg}\right|$ is the largest order statistic. Therefore, the value of $A^*_{ijg,\alpha,K}$ is chosen so that the true experimentwise error rate α is achieved.

To provide robustness under a distribution from the LTS family, we use the MML estimators of the location and scale parameters to obtain the following test statistic:

$$T_{ijg} = \frac{(\hat{\mu}_{ig.} - \hat{\mu}_{jg.}) - (\mu_{ig.} - \mu_{jg.})}{\sqrt{\frac{\hat{\sigma}^2_{ig.}}{n_i} + \frac{\hat{\sigma}^2_{jg.}}{n_j}}}. \qquad (2.4.3.4)$$

For illustration, the simulated values of the power of the t and T-tests are given in Table 2.4 for 100,000/n Monte Carlo runs where $k = 2$, $n_1 = n_2 = n$ and $G = 2000$. The presumed value of Type I error is 0.050.

Table 2.4 Values of the power for the T and t-tests

p	d:	0.00	0.25	0.50	0.75	1.00
2	T	0.043	0.685	0.936	0.995	0.999
	t	0.065	0.501	0.775	0.920	0.997
2.5	T	0.045	0.677	0.931	0.992	0.999
	t	0.069	0.513	0.786	0.932	0.994
3.5	T	0.048	0.673	0.926	0.989	0.999
	t	0.066	0.516	0.789	0.956	0.997
5.0	T	0.050	0.665	0.915	0.963	0.998
	t	0.068	0.601	0.875	0.960	0.998
10.0	T	0.052	0.633	0.901	0.955	0.998
	t	0.065	0.620	0.890	0.954	0.998

Table 2.4 indicates that the T-test maintains higher power compared to t-test.

2.4.4 Robustness of Estimators and Tests

In experimental design it is very important to obtain estimators and hypothesis testing procedures which have certain optimal properties with respect to an assumed error distribution. In spite of our best efforts to identify the underlying distribution through graphical techniques or goodness-of-fit tests, in practice, the shape parameters might be misspecified or the data might contain outliers, inliers or be contaminated. Thus deviations from an assumed distribution occur. That brings the issue of robustness in focus. An estimator is called robust if it is fully efficient (or nearly so) for an assumed distribution but maintains high efficiency for plausible alternatives. Also, a test is said to have criterion robustness if its Type I error is not substantially higher than a pre-specified level and is said to have

efficiency robustness if its power is high, at any rate for plausible alternatives to an assumed distribution (Tiku et al., 1986).

To show the robustness of both MML estimators and the test procedures based on MMLE, we consider the following plausible alternatives (1)-(4) to the assumed long-tailed symmetric distribution in (2.1.1) with p=3:

(1) Misspecification of the distribution: LTS (p=2.0, σ)
(2) Dixon's outlier model: (n-r) observations come from LTS (p=3.0, σ) but r observations (we do not know which ones) come from LTS (p=3.0, 2σ)
(3) Mixture model: 0.90 LTS (p=3.0, σ)+0.10 LTS (p=3.0, 2σ)
(4) Contamination model:
0.90 LTS (p=3.0, σ)+ 0.10 Uniform (-1/2,1/2)

Table 2.5 are the values of relative efficiency of the LS estimators of μ, V_1, G_1, $(VG)_{11}$ and σ. Simulations are based on 100000/n Monte Carlo runs where $k = 2$, $n_1 = n_2 = n = 30$ and $G = 2000$.

Table 2.5 Relative efficiencies of LS estimators of

μ, V_1, G_1, $(VG)_{11}$ and σ

Model	$\tilde{\mu}$	$\tilde{V}_1$	$\tilde{G}_1$	$(\widetilde{VG})_{11}$	σ
(1)	68.15	55.14	54.01	50.54	45.36
(2)	46.17	35.48	34.50	31.14	30.01
(3)	39.56	44.89	44.78	40.85	35.99
(4)	80.65	72.48	72.30	68.20	56.68

Table 2.5 indicates that the MML estimators $\hat{\mu}$, $\hat{V}_1$, $\hat{G}_1$ and $(\widehat{VG})_{11}$ are remarkably efficient and robust than LS estimators.

To show the robustness property of W-test, the simulated values of Type I error and the power of W and F-tests are given in Table 2.6 for 100,000/n Monte Carlo runs where $k = 2$, $n_1 = n_2 = n$ and $G = 2000$.

Table 2.6 Values of the power for the W and F-tests

		Model (1)		Model (2)		Model (3)		Model (4)	
d		W	F	W	F	W	F	W	F
	V	0.027	0.044	0.041	0.047	0.039	0.048	0.045	0.051
0.00	G	0.029	0.046	0.038	0.044	0.037	0.051	0.044	0.048
	VG	0.028	0.046	0.041	0.051	0.039	0.048	0.046	0.054
0.25	V	0.582	0.521	0.526	0.482	0.601	0.529	0.659	0.595
	G	0.584	0.502	0.460	0.432	0.621	0.551	0.654	0.602
	VG	0.577	0.503	0.472	0.425	0.603	0.539	0.661	0.625
0.50	V	0.885	0.795	0.865	0.769	0.920	0.845	0.952	0.925
	G	0.901	0.795	0.832	0.731	0.915	0.842	0.952	0.965
	VG	0.890	0.774	0.830	0.735	0.926	0.830	0.950	0.930
0.75	V	0.995	0.972	0.956	0.945	0.964	0.925	0.993	0.975
	G	0.994	0.975	0.958	0.953	0.968	0.934	0.991	0.978
	VG	0.995	0.972	0.953	0.945	0.972	0.929	0.993	0.975
1.00	V	0.998	0.997	0.997	0.997	0.999	0.995	0.998	0.998
	G	0.999	0.999	0.996	0.994	0.998	0.995	0.999	0.997
	VG	0.999	0.998	0.995	0.990	0.998	0.991	0.999	0.996

Table 2.6 shows that W-test has smaller Type I error and it has also higher power than the F-test.

The simulated values of Type I error and the power of the T and t tests for n=20 and d=0.50 for 100,000/n Monte Carlo and $G = 2000$ are given in Table 2.7.

Table 2.7 Values of Type I error and power for the T and t-tests

	Model (1)		Model (2)		Model (3)		Model (4)	
d	T	t	T	t	T	t	T	t
0.00	0.021	0.058	0.030	0.052	0.034	0.048	0.039	0.052
0.25	0.687	0.496	0.628	0.445	0.606	0.593	0.619	0.654
0.50	0.880	0.763	0.856	0.778	0.925	0.842	0.949	0.928
0.75	0.995	0.964	0.953	0.952	0.978	0.914	0.995	0.960
1.00	0.998	0.998	0.995	0.990	0.998	0.991	0.998	0.997

Table 2.7 indicates that T-test has smaller Type I error and it has also higher power than the t-test.

CHAPTER 3

ADAPTIVE MODIFIED MAXIMUM LIKELIHOOD ESTIMATION

The MML estimators developed by Tiku and Suresh (1992) are based on the assumption of a particular distribution. However, in some cases like machine data processing, the nature of the underlying distribution cannot be determined and it is just assumed that it is a member of a broad class of distributions (Hampel et al., 1986). It can be also assumed that the sample contains mild to strong outliers or other data anomalies. For such common situations when a statistician has no opportunity to investigate the nature of the underlying distribution, Huber (1964) and his collaborators developed M-estimators which are efficient and robust when the underlying distribution is one of the broad family of long-tailed symmetric distributions (Huber, 1981; Hampel et al., 1986; Staudte and Sheather, 1990). In this chapter, following Tiku and Sürücü (2009) and Dönmez (2010), we use a new form of the MML estimators which only assume that the distribution is unspecified long-tailed symmetric distribution as M-estimators do. Tiku and Sürücü (2009) called these estimators MML30 whereas Dönmez (2010) called them Revised Modified Maximum Likelihood estimators. Prof. Moti Lal Tiku who suggested the revised version of MML estimators decided to call these estimators as Adaptive Modified Maximum Likelihood (AMML) estimators (Akkaya and Tiku, 2011). In this chapter, along with the AMML estimators for the unbalanced two-way classification model,

the efficiency and robustness properties of AMML estimators and hypothesis tests based on AMML estimators are investigated and compared with LS and Huber's M-estimators.

3.1 Huber's M-Estimators

Consider a random sample from a distribution of type

$$\frac{1}{\sigma} f\left(\frac{y-\mu}{\sigma}\right) \tag{3.1.1}$$

where μ and σ are the location and scale parameters, respectively.

Huber (1964) proposed a new method to estimate μ assuming in particular that f is symmetric and long-tailed distribution.

The log-likelihood function is

$$\ln L = -n\ln\sigma + \sum_{i=1}^{n} \ln f\left(\frac{y_i-\mu}{\sigma}\right). \tag{3.1.2}$$

If the functional form of f is known, the maximum likelihood estimator of μ is obtained from the equation

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma}\sum_{i=1}^{n} \psi(z_i) = 0 \tag{3.1.3}$$

where

$$\psi(z_i) = -\frac{f'(z_i)}{f(z_i)} \quad \text{and} \quad z_i = \frac{y_i - \mu}{\sigma}.$$

By letting $w_i = w_i(z) = \frac{\psi(z_i)}{z_i}$, equation (3.1.3) reduces to $\sum_{i=1}^{n} w_i(y_i - \mu) = 0$. The solution of the equation gives μ as follows:

$$\mu = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}. \tag{3.1.4}$$

Given σ and $\psi(z_i)$, equation (3.1.2) may be solved by iteration (Low, 1959). It may also be solved by applying Newton-Raphson's procedure to equation (3.1.3) (Gross, 1976).

In practice, however, σ and $\psi(z_i)$ are not known. Therefore, Huber (1964) proposed a function $\psi(z_i)$ as

$$\psi(z) = \begin{cases} z & \text{if} \quad |z| \le c \\ c\,\text{sgn}(z) & \text{if} \quad |z| > c \end{cases} \tag{3.1.5}$$

which is the combination of the normal distribution in the middle with the double-exponential distribution in the tails. Birch and Myers (1982) give 1.345, 1.5, and 2.0 as the popular choice of c values since these choices correspond roughly to 10, 5, and 2.5 percent censoring on either side of a normal sample.

The solution of (3.1.3) is referred to as Huber's M-estimator and denoted by $\hat{\mu}_H$.

For unknown σ, $\tilde{\sigma}_0$ = mad = median $|y_i - \text{median}\,(y_i)|$ is used by Huber (1964, 1977), Gross (1976, 1977) to estimate σ. However, Huber (1981) and Birch and Myers (1982) suggest to use $\tilde{\sigma}_0 / 0.6745$ instead of $\tilde{\sigma}_0$ to obtain an asymptotically unbiased estimator of σ in the case of normal distribution.

By using the asymptotic variance of the M-estimator $\hat{\mu}_H$

$$\frac{(1/n)\sigma^2 E\left(\psi^2(z)\right)}{\left[E\left(\psi'(z)\right)\right]^2}, \tag{3.1.6}$$

Huber (1977, 1981) calculated the estimator of scale, $\hat{\sigma}_H$, as

$$\hat{\sigma}_H = \left\{ \frac{n\hat{\sigma}_0^2 \left[\sum_{i=1}^{n} \psi^2\left(\frac{y_i - \hat{\mu}_H}{\tilde{\sigma}_0}\right)\right]}{\left[\sum_{i=1}^{n} \psi'\left(\frac{y_i - \hat{\mu}_H}{\tilde{\sigma}_0}\right)\right]^2} \right\}^{1/2} . \tag{3.1.7}$$

When the functional form of f is not known, $\psi(z)$ may be approximated by descending functions. The function which decreases as $|z|$ increases is called as descending function. There are three important descending functions:

1. The wave function (Andrews et al., 1972; Andrews, 1974)

$$\psi(z) = \begin{cases} \sin(z) & \text{if} \quad |z| \le \pi \\ 0 & \text{if} \quad |z| > \pi \end{cases} \tag{3.1.8}$$

2. The bisquare function (Beaton and Tukey, 1974)

$$\psi(z) = \begin{cases} z(1 - z^2)^2 & \text{if} \quad |z| \le 1 \\ 0 & \text{if} \quad |z| > 1 \end{cases} \tag{3.1.9}$$

3. The Hampel piecewise linear function (Hampel, 1974)

$$\psi(z) = \text{sgn}(z) \begin{cases} |z| & 0 \le |z| < a \\ a & a \le |z| < b \\ \dfrac{c - |z|}{c - b} & b \le |z| < c \\ 0 & c \le |z| \end{cases} \tag{3.1.10}$$

For different values of a, b, and c, different estimators are obtained.

Gross (1976) showed that the wave, bisquare, and Hampel piecewise linear functions were the most efficient descending functions when the adjusting constant h was equal to 2.4, 8.2, and 2.2, respectively. The estimators of location and scale obtained by using these three functions are called as the wave estimator (W24), bisquare estimator (BS82), and Hampel estimator (H22). These estimators are as follows where $T_0 = \text{median}\{y_i\}$ and $S_0 = \text{median}\{|y_i - T_0|\}$:

3.1.1 W24 Estimator

$$\hat{\mu}^{W24} = T_0 + (hS_0)\tan^{-1}\left[\frac{\sum \sin(z_i)}{\sum \cos(z_i)}\right] \qquad (3.1.1.1)$$

and

$$\hat{\sigma}^{W24} = (hS_0)\left[n\frac{\sum \sin^2(z_i)}{\left(\sum \cos(z_i)\right)^2}\right]^{1/2} \qquad (3.1.1.2)$$

where

$$z_i = \frac{y_i - T_0}{hS_0} \quad \text{and} \quad h = 2.4\,.$$

Here, summations include only those i such that $|z_i| < \pi$.

3.1.2 BS82 Estimator

$$\hat{\mu}^{B82} = T_0 + (hS_0)\frac{\sum \psi(z_i)}{\sum \psi'(z_i)} \qquad (3.1.2.1)$$

and

$$\hat{\sigma}^{B82} = (hS_0)\left[n\frac{\sum \psi^2(z_i)}{\left(\sum \psi'(z_i)\right)^2}\right]^{1/2} \qquad (3.1.2.2)$$

where

$$z_i = \frac{y_i - T_0}{hS_0} \quad \text{and} \quad h = 8.2 .$$

Here, $\psi(z)$ is the Beaton and Tukey's (1974) bisquare function given in (3.1.9) and $\psi'(z)$ is the derivative of $\psi(z)$.

3.1.3 H22 estimator

$$\hat{\mu}^{H22} = T_0 + \frac{S_0 \sum \psi(z_i)}{\sum \psi'(z_i)} \tag{3.1.3.1}$$

and

$$\hat{\sigma}^{H22} = \left[\frac{n(S_0)^2 \sum \psi^2(z_i)}{\left(\sum \psi'(z_i)\right)^2} \right]^{1/2} \tag{3.1.3.2}$$

where

$$z_i = \frac{y_i - T_0}{S_0} .$$

Here, $\psi(z)$ is the Hampel piecewise linear function given in (3.1.10) for $a = 2.25$, $b = 3.75$, and $c = 15.0$ and $\psi'(z)$ is the derivative of $\psi(z)$.

For symmetric distributions, $\hat{\mu}^{W24}$, $\hat{\mu}^{B82}$ and $\hat{\mu}^{H22}$ are unbiased and have very good efficiency. They are also uncorrelated with $\hat{\sigma}^{W24}$, $\hat{\sigma}^{B82}$ and $\hat{\sigma}^{H22}$, respectively (Tiku, Tan, Balakrishnan, 1986). For long-tailed symmetric distributions, however, the M-estimators of σ can

have substantial downward bias, even asymptotically (Tiku, 1980; Dunnett, 1982).

3.1.4 Influence Function

The concept of influence function which is also known as breakdown was introduced by Hampel (1974) with the aim of verifying the robustness of an estimator. According to this concept, if an estimator assumes infinity values when the observations in a sample are shifted in either direction to infinity, the estimator is non-robust. In this respect, the M-estimators described in Section 3.1 are robust and also have bounded influence functions since their empirical influence functions are bounded (Hampel et al., 1986).

3.2 Adaptive Modified Maximum Likelihood (AMMLE) Estimator

Assume that the underlying distribution is one of the long-tailed symmetric family given in (2.1.1). MML estimators of μ and σ are given as follows (Tiku and Suresh, 1992):

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \beta_i x_{(i)}}{m} \qquad (3.2.1)$$

and

$$\hat{\sigma} = \frac{B + \sqrt{B^2 + 4nC}}{2\sqrt{n(n-1)}} \qquad (3.2.2)$$

where

$$m = \sum_{i=1}^{n} \beta_i ,$$

$$B = \frac{2p}{q} \sum_{i=1}^{n} \alpha_i (x_{(i)} - \hat{\mu})$$

and

$$C = \frac{2p}{q} \sum_{i=1}^{n} \beta_i (x_{(i)} - \hat{\mu})^2 .$$

The coefficients α_i and β_i are given by (Tiku and Suresh, 1992)

$$\alpha_i = \frac{(2/q)t_{(i)}^3}{\left(1 + (1/q)t_{(i)}^2\right)^2} \quad \text{and} \quad \beta_i = \frac{1 - (1/q)t_{(i)}^2}{\left(1 + (1/q)t_{(i)}^2\right)^2} \tag{3.2.3}$$

where $t_{(i)} = E(z_{(i)})$ and $z_{(i)} = (x_{(i)} - \mu)/\sigma$. These coefficients are obtained from Taylor series expansions.

As mentioned in Chapter 2, if C in (3.2.2) assumes a negative value, we replace α_i and β_i by α_i^* and β_i^*, respectively to obtain real valued $\hat{\sigma}$ (Islam and Tiku, 2004):

$$\alpha_i^* = \frac{(1/q)t_{(i)}^3}{\left(1 + (1/q)t_{(i)}^2\right)^2} \quad \text{and} \quad \beta_i^* = \frac{1}{\left(1 + (1/q)t_{(i)}^2\right)^2} . \tag{3.2.4}$$

Tiku and Sürücü (2009) showed that when β_i in (3.2.3) are estimated from a random sample and used in (3.2.1) and (3.2.2), the resulting MML estimators $\hat{\mu}^a$ and $\hat{\sigma}^a$ have high breakdown and they are

overall more efficient and robust than the M-estimators. Besides, they are easier to compute and utilize all the observations in a sample while the M-estimators implicitly censor a number of observations.

To estimate the parameters α_i and β_i, let

$$T_0 = \text{median}\{x_i\} \quad \text{and} \quad S_0 = 1.483 \text{ median}\{|x_i - T_0|\} \tag{3.2.5}$$

as in M-estimators (Huber,1981; Hampel et al., 1986). Here, T_0 is an unbiased estimator of μ for symmetric distributions and S_0 is asymptotically an unbiased estimator of σ for a normal distribution. Then we can estimate $t_{(i)}$ in (3.2.3) by $(x_{(i)} - T_0)/S_0$. Therefore, β_i are estimated by

$$w_{(i)} = \frac{1}{\left(1 + \frac{1}{q}\left(\frac{x_{(i)} - T_0}{S_0}\right)^2\right)^2} \quad (1 \le i \le n). \tag{3.2.6}$$

Since complete sums are invariant to ordering,

$$\hat{\mu}_x = \frac{\sum_{i=1}^{n} w_i x_i}{w} \qquad \left(w = \sum_{i=1}^{n} w_i\right) \tag{3.2.7}$$

and

$$\hat{\sigma}_x = \frac{B + \sqrt{B^2 + 4nC}}{2\sqrt{n(n-1)}} \tag{3.2.8}$$

where

$$B \cong \frac{2p}{q}\sum_{i=1}^{n} v_i(x_i - \hat{\mu}_x) \qquad \left(v_i = (w_i/q)\frac{x_i - T_0}{S_0}\right),$$

$$C = \frac{2p}{q}\sum_{i=1}^{n} w_i(x_i - \hat{\mu}_x)^2$$

and

$$w_i = \frac{1}{\left(1 + \frac{1}{q}\left(\frac{x_i - T_0}{S_0}\right)^2\right)^2}.$$

The coefficient v_i in the expression given for B are obtained from a_i by equating $((x_i - T_0)/S_0)^2$ to its expected value which is almost 1 for $p = 16.5 \quad (q = 30)$. This is necessary to have a bounded influence function (Dönmez, 2010). If we choose q very large, $w_i \; (1 \le i \le n)$ reduces to the sample mean $\bar{x}$ which, although fully efficient for a normal distribution, has zero breakdown and is not efficient and robust for long-tailed symmetric distributions even to moderate outliers in a sample. On the other hand, if we choose q small, $\hat{\mu}_x$ and $\hat{\sigma}_x$ are enormously inefficient for normal and near-normal distributions. Thus the choice $q = 30$ turns out to be a good compromise. The corresponding MML estimators are called as MML30 by Tiku and Sürücü (2009). They also examined the efficiency and robustness properties of MML30 estimators and

showed that MML30 estimators are more efficient than Huber's W24 estimators and have high breakdown.

3.3 Unbalanced Two-Way Classification with Interaction via AMML

Consider the model given in (2.1) when error terms have a distribution from LTS symmetric family. The MML estimation procedure for the model parameters is given in Chapter 2. The method of obtaining AMML estimators are similar to the ones used for MML. The estimates of $\alpha_{kg(l)}$ and $\delta_{kg(l)}$ used in linear approximations given in (2.4.7) are obtained by replacing $t_{kg(l)}$ in (2.4.8) by $\tilde{t}_{kg(l)} = (y_{kg(l)} - T_{0kg})/S_{0kg}$ where

$$T_{0kg} = \text{median}\{y_{kgl}\} \text{ and } S_{0kg} = 1.483 \text{ median}\{|y_{kgl} - T_{0kg}|\} \quad (3.3.1)$$

for the $(g,k)^{th}$ cell $(1 \le g \le G, 1 \le k \le K)$.

We disregard the ordering of z_{kgl} since complete sums are invariant to ordering and take $\tilde{t}_{kgl} = (y_{kgl} - T_{0kg})/S_{0kg}$ as the initial estimate of t_{kgl}. Thus the initial estimates of α_{kgl} and δ_{kgl} are obtained by replacing t_{kgl} by $\tilde{t}_{kgl}$ and resulting coefficients are denoted by $\tilde{\alpha}_{kgl}$ and $\tilde{\delta}_{kgl}$, respectively. The resulting AMML estimators of the parameters which have the properties described in Section 2.4 are given as follows:

$$\hat{\mu}^{a} = \hat{\mu}^{a}_{\ldots}, \tag{3.3.2}$$

$$\hat{V}^{a}_{k} = \hat{\mu}^{a}_{k..} - \hat{\mu}^{a}, \tag{3.3.3}$$

$$\hat{G}^{a}_{g} = \hat{\mu}^{a}_{.g.} - \hat{\mu}^{a}, \tag{3.3.4}$$

$$(\widehat{VG})^{a}_{kg} = \hat{\mu}^{a}_{kg.} - \hat{\mu}^{a} - \hat{V}^{a}_{k} - \hat{G}^{a}_{g} \tag{3.3.5}$$

and

$$\hat{\sigma}^{a} = \frac{-B + \sqrt{B^2 + 4NC}}{2\sqrt{N(N - KG)}} \tag{3.3.6}$$

where

$$\hat{\mu}^{a}_{\ldots} = \frac{\sum_{g=1}^{G}\sum_{k=1}^{K} m_{kg}\hat{\mu}^{a}_{kg.}}{\sum_{g=1}^{G}\sum_{k=1}^{K} m_{kg}}, \quad \hat{\mu}^{a}_{k..} = \frac{\sum_{g=1}^{G} m_{kg}\hat{\mu}^{a}_{kg.}}{\sum_{g=1}^{G} m_{kg}}, \quad \hat{\mu}^{a}_{.g.} = \frac{\sum_{k=1}^{K} m_{kg}\hat{\mu}^{a}_{kg.}}{\sum_{k=1}^{K} m_{kg}},$$

$$\hat{\mu}^{a}_{kg.} = \frac{\sum_{l=1}^{n_k} \tilde{\delta}_{kgl} y_{kgl}}{m_{kg}}, \quad m_{kg} = \sum_{l=1}^{n_k} \tilde{\delta}_{kgl}, \quad B = \frac{2p}{q}\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k} \tilde{a}_{kgl}(y_{kgl} - \hat{\mu}^{a}_{kg.})$$

and

$$C = \frac{2p}{q}\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{l=1}^{n_k} \tilde{\delta}_{kgl}(y_{kgl} - \hat{\mu}^{a}_{kg.})^2 .$$

3.3.1 Efficiency Properties

Table 3.1 gives the relative efficiencies of LS estimators, $\tilde{\mu}$, $\tilde{V}_1$, $\tilde{G}_1$ and $(\widetilde{VG})_{11}$ with respect to MML and AMML estimators for 100,000/n Monte Carlo runs where $k = 2$, $n_1 = n_2 = n$ and $G = 2000$. Their biases are not reported since the biases in them were found to be negligible.

As it is seen from Table 3.1, The table indicates that the AMML estimators $\hat{\mu}^a$, $\hat{V}_1^a$, $\hat{G}_1^a$ and $(\widehat{VG})_{11}^a$ are considerably more efficient than LS estimators even for small sample sizes. Their efficiencies are also slightly higher than MML estimators. Note that for $p = 10$, the LS estimators are almost as efficient as AMML estimators. This is expected result since the long-tailed symmetric distribution reduces to a normal for $p = \infty$. For small p values which are more appropriate for heavy-tailed microarray data, AMML estimators are enormously more efficient than LS estimators. The relative efficiencies of LS estimators, $\tilde{\mu}$, $\tilde{V}_1$, $\tilde{G}_1$ and $(\widetilde{VG})_{11}$ decreases as sample size increases.

Table 3.1 Relative efficiencies of the LS estimators with respect to AMML and MML estimators

(1) $V(\hat{\mu}^a)/V(\tilde{\mu}).100$ (2) $V(\hat{\mu})/V(\tilde{\mu}).100$ (3) $V(\hat{V}_k^a)/V(\tilde{V}_k).100$

(4) $V(\hat{V}_k)/V(\tilde{V}_k).100$ (5) $V(\hat{G}_g^a)/V(\tilde{G}_g).100$ (6) $V(\hat{G}_g)/V(\tilde{G}_g).100$

(7) $V(\widehat{VG}_{kg}^a)/V(\widetilde{VG}_{kg}).100$ (8) $V(\widehat{VG}_{kg})/V(\widetilde{VG}_{kg}).100$

	p:	2	2.5	3.5	5	10
$n_k=5$	(1)	66.895	85.154	93.540	95.600	98.992
	(2)	68.654	86.065	94.932	97.377	99.800
	(3)	69.255	84.996	92.005	95.458	99.001
	(4)	70.489	86.220	94.880	98.200	99.596
	(5)	70.588	82.695	91.962	95.485	99.010
	(6)	72.667	84.522	94.711	98.702	99.455
	(7)	71.801	81.477	92.648	96.870	98.870
	(8)	72.055	84.906	94.321	97.182	99.701
$n_k=10$	(1)	53.147	73.335	89.455	93.066	97.056
	(2)	57.156	76.405	90.527	96.467	99.620
	(3)	52.565	75.691	90.015	94.560	97.322
	(4)	54.463	76.747	91.423	95.876	98.838
	(5)	44.898	73.025	87.025	94.362	97.890
	(6)	45.729	73.862	90.339	96.806	99.076
	(7)	43.005	73.960	90.652	94.055	97.010
	(8)	45.548	76.758	92.544	96.145	98.977
$n_k=20$	(1)	45.802	71.050	82.458	91.036	98.050
	(2)	48.884	72.980	85.903	94.104	98.455
	(3)	49.740	69.056	86.744	93.001	96.050
	(4)	52.903	70.651	87.167	95.387	98.740
	(5)	44.010	70.256	85.940	93.658	98.554
	(6)	45.220	72.084	86.424	95.415	99.012
	(7)	41.500	70.156	87.596	92.885	96.589
	(8)	44.628	70.259	88.202	95.509	98.565
$n_k=50$	(1)	45.895	66.965	81.960	92.020	96.660
	(2)	47.003	70.380	84.069	94.052	98.405
	(3)	51.465	68.956	85.655	91.658	96.581
	(4)	52.261	70.108	86.855	94.743	98.689
	(5)	44.585	68.555	85.032	91.056	97.010
	(6)	45.057	69.767	85.914	94.270	98.668
	(7)	44.000	67.569	83.458	90.008	96.050
	(8)	44.073	68.831	86.174	92.850	98.428

3.3.2 Robustness Properties

To illustrate the robustness of AMML estimators, we consider the following models as plausible alternatives:

(1) $N(0, \sigma^2)$

(2) LTS(p = 5) (3) LTS(p = 3.5) (4) LTS(p = 2.5) (5) LTS(p = 2)

Outlier models: (n-r) observations come from $N(0, \sigma^2)$ and r observations (we do not know which ones) come from

(6) $N(0, 4\sigma^2)$ (7) $N(0, 16\sigma^2)$

where $r = [0.5 + 0.1n]$

Mixture models:

(8) $0.90\ N(0, \sigma^2) + 0.10\ N(0, 4\sigma^2)$ (9) $0.90\ N(0, \sigma^2) + 0.10\ N(0, 16\sigma^2)$

(10) Student's t distribution with 2 degrees of freedom

(11) Cauchy distribution

(12) Slash (Normal/Uniform) distribution

Models (1)-(9) have finite mean and variance, (10) has finite mean but non-existent variance, and (11)-(12) have non-existent mean and variance.

We generated 100,000/n samples of size n from each of the models (1)-(12) where G=2000. The observations generated from the models (6)-(9) were divided by suitable constants to make their variances equal to σ^2. Table 3.2 are the values of variances of AMML and W24 estimators for the location parameter. Here we take only W24 estimators from M-estimator since the results for W24, BS82 and

H22 estimators have almost same results. We do not give the simulated means of these estimators since both are unbiased.

It is seen from the table that $\hat{\mu}_x$ is a little less efficient than $\hat{\mu}^{W24}$ for normal distribution. For models (2)-(9), $\hat{\mu}_x$ is more efficient. Also, $\hat{\mu}_x$ is considerably more efficient for models (10)-(12) which have non-existent variances.

Table 3.2 Simulated values of $(n/\sigma^2)\mathrm{Var}(\hat{\mu}_x)$ and $(n/\sigma^2)\mathrm{Var}(\hat{\mu}^{W24})$

	n = 10		n = 20		n = 50	
Model	$\hat{\mu}_x$	$\hat{\mu}^{W24}$	$\hat{\mu}_x$	$\hat{\mu}^{W24}$	$\hat{\mu}_x$	$\hat{\mu}^{W24}$
(1)	1.095	1.061	1.066	1.037	1.019	1.001
(2)	0.954	0.958	0.925	0.939	0.936	0.945
(3)	0.902	0.918	0.866	0.889	0.884	0.912
(4)	0.755	0.787	0.751	0.789	0.720	0.755
(5)	0.569	0.623	0.541	0.610	0.535	0.580
(6)	0.954	0.957	0.945	0.943	0.941	0.949
(7)	0.555	0.590	0.548	0.577	0.558	0.581
(8)	0.935	0.934	0.935	0.940	0.933	0.951
(9)	0.579	0.620	0.558	0.599	0.560	0.612
(10)	2.270	2.624	2.012	2.618	1.964	2.301
(11)	4.710	6.458	3.896	5.201	3.288	4.459
(12)	7.826	10.325	7.489	9.305	6.687	8.250

The simulated means and variances of the AMML and W24 estimators for the scale parameter are given in Table 3.3 and Table 3.4, respectively (100,000/n Monte Carlo runs where $k = 2$,

$n_1 = n_2 = n$ and $G = 2000$). They indicate that $\hat{\sigma}_x$ has a little larger bias than $\hat{\sigma}^{W24}$, however, it has smaller mean square errors. Therefore, AMML estimators are as good as M-estimators or better. These results are also in a good agreement with the results for the balanced two-way model with interaction given in Dönmez (2010).

Table 3.3 Simulated values of $(1/\sigma)$mean of $\hat{\sigma}_x$ and $\hat{\sigma}^{W24}$

	n = 10		n = 20		n = 50	
Model	$\hat{\sigma}_x$	$\hat{\sigma}^{W24}$	$\hat{\sigma}_x$	$\hat{\sigma}^{W24}$	$\hat{\sigma}_x$	$\hat{\sigma}^{W24}$
(1)	0.918	0.920	0.969	0.988	0.989	1.010
(2)	0.906	0.910	0.934	0.959	0.936	0.966
(3)	0.865	0.872	0.909	0.945	0.911	0.935
(4)	0.809	0.815	0.845	0.868	0.836	0.870
(5)	0.718	0.721	0.725	0.756	0.754	0.775
(6)	0.889	0.892	0.928	0.960	0.929	0.956
(7)	0.721	0.719	0.751	0.759	0.745	0.760
(8)	0.905	0.906	0.935	0.969	0.936	0.955
(9)	0.731	0.733	0.754	0.765	0.751	0.778
(10)	1.418	1.429	1.435	1.485	1.430	1.605
(11)	2.071	2.085	1.936	2.029	1.918	2.045
(12)	2.844	2.848	2.789	2.862	2.605	2.930

Table 3.4 Simulated values of (n/σ^2)variance of $\hat{\sigma}_x$ and $\hat{\sigma}^{W24}$

	n = 10		n = 20		n = 50	
Model	$\hat{\sigma}_x$	$\hat{\sigma}^{W24}$	$\hat{\sigma}_x$	$\hat{\sigma}^{W24}$	$\hat{\sigma}_x$	$\hat{\sigma}^{W24}$
(1)	0.565	0.539	0.528	0.519	0.529	0.518
(2)	0.641	0.630	0.633	0.665	0.589	0.612
(3)	0.681	0.690	0.660	0.672	0.629	0.695
(4)	0.378	0.703	0.665	0.726	0.654	0.711
(5)	0.655	0.691	0.580	0.654	0.578	0.640
(6)	0.584	0.588	0.542	0.559	0.539	0.561
(7)	0.455	0.457	0.431	0.478	0.425	0.452
(8)	0.645	0.646	0.618	0.656	0.590	0.634
(9)	0.700	0.788	0.632	0.755	0.618	0.695
(10)	3.266	3.620	2.969	3.256	2.875	3.275
(11)	14.010	16.901	9.152	10.896	8.922	10.758
(12)	25.560	28.905	14.045	19.569	12.001	18.648

3.3.3 Comparisons of Treatment Effects

In Chapter 2, we suggest using T_{ijg} given in (2.4.3.4) as a test statistic to make comparisons of treatment means under a distribution from LTS family. Here we will use AMML estimators of mean and variance in testing procedure since they are more efficient and robust than M-estimators (Tiku and Sürücü, 2009).

To provide robustness under a distribution from the LTS family, we replace location and scale parameters in T_{ijg} with the corresponding AMML estimators and obtain the following test statistic:

$$T_{ijg}^{a} = \frac{(\hat{\mu}_{ig.}^{a} - \hat{\mu}_{jg.}^{a}) - (\mu_{ig.} - \mu_{jg.})}{\sqrt{\frac{(\hat{\sigma}_{ig.}^{a})^2}{n_i} + \frac{(\hat{\sigma}_{jg.}^{a})^2}{n_j}}}. \qquad (3.3.3.1)$$

where for $(i,g)^{th}$ cell, $\hat{\mu}_{ig.}^{a}$ and $\hat{\sigma}_{ig.}^{a}$ are computed from (3.2.1) and (3.2.2), respectively. The simulated power values of the tests T_{ijg}^{a} and t_{ijg}^{W24} (100,000/n Monte Carlo runs where $k = 2$, $n_1 = n_2 = n$ and $G = 2000$) obtained by incorporating W24 estimators into (2.4.3.3), respectively are given in Table 3.5 for various values of $\mu_{ig} - \mu_{jg}$ $(i, j = 1, 2, ..., K, i \neq j, g = 1, 2, ..., G)$. For $d = 0$, the power reduces to Type I error which is assumed as 0.05 in this study.

Table 3.5 Values of Type I error and power for the T^{a} and t^{W24} tests

p	d:	0.00	0.25	0.50	0.75	1.00
2	T^{a}	0.039	0.705	0.960	0.998	0.999
	t^{W24}	0.054	0.559	0.784	0.971	0.998
2.5	T^{a}	0.041	0.698	0.956	0.995	0.999
	t^{W24}	0.058	0.562	0.789	0.975	0.999
3.5	T^{a}	0.044	0.688	0.944	0.991	0.999
	t^{W24}	0.065	0.570	0.805	0.980	0.999
5.0	T^{a}	0.048	0.670	0.931	0.988	0.999
	t^{W24}	0.067	0.654	0.859	0.983	0.999
10.0	T^{a}	0.051	0.662	0.910	0.986	0.999
	t^{W24}	0.058	0.641	0.903	0.985	0.999

Table 3.5 indicates that T^a test has smaller Type I error and it has also higher power than the t^{W24}-test.

3.3.4 Robustness Comparisons of the Tests

Since our aim is to obtain robust estimators for the comparisons of the treatment means under a distribution from LTS family when the nature of the underlying distribution cannot be determined, LTS distributions need to be inclusive of extreme distributions like Cauchy and also situations when a sample contains strong outliers and other strong data anomalies. Therefore as the plausible alternatives, we consider again the distributions given in section 3.3.2.

To show the robustness properties of T^a and t^{W24} obtained by using AMML and W24 estimators, respectively, the simulated values of the power of T^a and t^{W24} tests for detectable difference d=0.5 and 100,000/n Monte Carlo runs where $k = 2$, $n_1 = n_2 = n$ and $G = 2000$ are given in Table 3.6.

Table 3.6 Values of the power for the T^a and t^{W24} tests

	n = 10		n = 20		n = 50	
Model	T^a	t^{W24}	T^a	t^{W24}	T^a	t^{W24}
(1)	0.755	0.751	0.783	0.785	0.795	0.792
(2)	0.771	0.756	0.789	0.763	0.793	0.789
(3)	0.785	0.769	0.796	0.771	0.803	0.795
(4)	0.805	0.765	0.812	0.768	0.839	0.796
(5)	0.864	0.790	0.875	0.801	0.880	0.865
(6)	0.763	0.735	0.772	0.742	0.781	0.766
(7)	0.699	0.638	0.701	0.640	0.709	0.638
(8)	0.765	0.735	0.777	0.741	0.780	0.765
(9)	0.690	0.625	0.696	0.624	0.703	0.688
(10)	0.455	0.502	0.462	0.509	0.496	0.516
(11)	0.601	0.775	0.612	0.781	0.635	0.790
(12)	0.405	0.520	0.439	0.538	0.449	0.540

Table 3.6 indicates that T^a and t^{W24} tests give almost the same power values for normal distribution denoted by Model (1). For models (2)-(9) including LTS distributions with different shape parameters, outlier models and mixture models, T^a test is apparently superior to t^{W24} test. However, for model (10) with finite mean and non-existent variance and for models (11)-(12) which have non-existent mean and variance, t^{W24} test is more powerful than T^a test. Overall, T^a test based on AMML estimators performs much better than t^{W24} test based on W24 estimators for most of the cases.

CHAPTER 4

COMPARISON OF STATISTICAL METHODS FOR IDENTIFYING DIFFERENTIAL EXPRESSION

Although various statistical methods have been suggested to test the differential gene expression, there have been a few studies which compare the different statistical approaches. It is due to the fact that there are no golden standards to assess accuracy of microarray analysis (Gyorffy et al., 2009). Some parametric methods were compared by Smyth et al. (2003) whereas the performances of some nonparametric methods were evaluated by Troyanskaya et al. (2002). In addition to these, comparative studies including both parametric and nonparametric methods were conducted by Broberg (2002), Jeffery et al. (2006) and Kim et al. (2006).

In this chapter, we extensively compare six types of parametric methods (t-test, Bayes t-test, ANOVA, W24, MMLE and AMMLE) and one non-parametric method (SAM) using both the three real microarray experiments and the simulated datasets. t-test, Bayes t-test and ANOVA are as described in Section 1.4 whereas W24 test is discussed in Chapter 3. Throughout this chapter, the abbreviation "MMLE" stands for the whole procedure described in Chapter 2, consisting of analysis of variance using MML estimators followed by the pairwise multiple comparisons. The abbreviation "AMMLE"

denotes the complete estimation and testing method using AMML estimators introduced in Chapter 3.

4.1 Comparisons via Real Datasets

Each of the three real data sets are normalized by subtracting the median and dividing its interquartile range (IQR) as in Broberg (2002). This preprocessing method is used t-test, Bayes t-test and SAM except for ANOVA, W24, MMLE, and AMMLE techniques. For ANOVA, W24, MMLE and AMMLE methods, raw data were used because of the reasons described in Chapter 1. It should also be noted that all of the computations for the statistical methods other than W24, MMLE and AMMLE are carried out by using FlexArray (Blazejczyk, 2007) which is a Microsoft Windows software package for statistical analysis of microarray expression.

4.1.1 Leukemia Data

The leukemia dataset of Golub et al. (1999) consists of 38 bone marrow samples on the microarray chips containing $G = 7129$ human genes. The samples either belong to the acute lymphoblastic leukemia (ALL) or the acute myeloid leukemia (AML) patients, with 27 categories of the first category and 11 of the second. The goal of this experiment is to identify differentially expressed genes in 27 acute ALL patients and 11 acute myeloid leukemia AML patients.

4.1.2 Melanoma Data

The melanoma dataset of Bittner et al. (2000) was gathered from a study of gene expression profiles for 38 samples, including 31 melanomas and 7 controls. The samples were hybridized to microarray chips containing $G = 8067$ genes. The goal of this experiment is to find differentially expressed genes in the melanomas compared to healthy cells.

4.1.3 Apolipoprotein AI Mouse Data

Apolipoprotein AI dataset of Callow et al. (2000) was obtained from a study which consists of treatment group of 8 mice with the apoliprotein AI gene knocked out and control group of 8 normal mice. The samples were hybridized to microarray chips containing $G = 6384$ genes. The goal of this experiment is to find differentially expressed genes in the livers of treatment mice compared to healthy mice.

4.1.4 Real Dataset Results

t-test, Bayes t-test, ANOVA, SAM, W24, MMLE and AMMLE methods are compared by three real microarray datasets mentioned in Section 4.1. Average ranks of reference genes which are believed to be differentially expressed are used in the comparison process since there are no golden standards to assess accuracy of microarray analysis (Gyorffy et al., 2009). Therefore, the choices of reference genes become very important in this comparison study.

Broberg (2002) used 50 reference genes that were selected by Mixture Model Method (MMM) of Pan et al. (2003) in the leukemia data and ranked all genes in order of absolute values of each test statistic. Then comparisons were made by evaluating the average ranks of these testing methods. Kim et al. (2006) pointed out a problem in this study of Broberg (2002). They stated that this study practically failed to select fair reference genes because of the fact that the use of MMM to select reference genes for comparing six testing methods gives the best performance of the testing method which is most similar to MMM method. For this reason, we adopted the approach of Kim et al. (2006) in our study. According to this approach, we used these reference genes which show significant difference between two samples by all the tests such as t-test, Bayes t-test, SAM, ANOVA, W24, MMLE and AMMLE methods. We initially selected top 5% significant genes by each of seven testing methods and finally selected a small number of reference genes (65 in leukemia, 58 in melanoma and 18 in mouse dataset) that were commonly found to be significant by all the seven methods. Table 4.1 shows the average ranks of the reference genes in both large and small sample cases. It should be noted that lower average rank means higher performance since it implies that the method identifies the differentially expressed genes more precisely.

Table 4.1 Table of average ranks of the reference genes

		Leukemia	Melanoma	AI
AMMLE	Large	58.60	120.41	45.16
	Small	454.20	735.43	659.61
MMLE	Large	61.40	122.50	49.88
	Small	456.80	737.81	662.05
W24	Large	61.60	122.46	47.27
	Small	457.80	738.79	674.00
ANOVA	Large	84.60	127.93	71.38
	Small	495.80	903.72	745.44
SAM	Large	71.05	125.10	64.22
	Small	479.55	786.37	716.38
t-test	Large	135.00	128.58	58.94
	Small	534.80	1206.81	702.83
Bayes t	Large	126.60	246.05	85.88
	Small	701.20	1394.32	677.72

For the leukemia dataset, we used large sample (26 replications of ALL and 10 replications of AML) and small sample (5 replications of ALL and 5 replications of AML) for two groups. We initially selected 356 significant genes (% 5 of 7129 genes) from each method, and finally selected 65 reference genes that were commonly found to be significant by all the seven methods. As shown in Table 4.1, AMMLE gives the smallest average rank in both large and small sample cases. MMLE and W24 values are almost the same and they give the second smallest rank for both small and large samples whereas t-test and Bayes t-test seems to be poor in both cases.

For the melanoma dataset, both large (31 replications of melanomas and 7 replications of control group) and small samples (4 replications of melanomas and 4 replications of control group) were used. We initially selected 407 significant genes (5% of a total of 8067 genes) obtained from each method and finally selected 58 reference genes that were commonly found to be significant by all the seven methods. In large sample case, ANOVA, SAM and t-test give almost the same average ranks. MMLE and W24 are slightly better than ANOVA, SAM and t-test but much better than Bayes t-test. AMMLE performs better than the other tests in for both large and small samples.

For apolipoprotein AI dataset, both large sample (8 replications of the apolipoprotein AI gene knocked out group and 8 replications of control group) and small samples (4 replications of the apolipoprotein AI gene knocked out group and 4 replications of control group). We initially selected 319 significant genes (5% of a total of 6384 genes) from each method, and finally selected 18 reference genes that were commonly found to be significant by all the five methods. In large and small sample cases, AMMLE performs the best overall, whereas W24 is the second best. In small sample case, AMMLE and MMLE gives the smallest and the second smallest average rank, respectively.

Through the analysis of three real datasets, we are able to recognize that the rankings of the all methods except AMMLE which gives the best results for all cases, differ depending on the microarray data. Kim et al. (2006) explained the reasons of this situation by the fact that the performance of testing methods depends on the normal distribution assumption or equal variance assumption. They noted that the percentages of genes which satisfy the normality assumption

by Kolmogorov-Smirnov test are 31.5%, 36.3% and 78.5% whereas the percentages of genes which satisfy the equal variance assumption by F-test are 23.7%, 24.2% and 85.0% for the leukemia, melanoma and apolipoprotein AI mouse data, respectively. For illustrative purpose, we constructed the Q-Q plots of residuals of the ANOVA model to check the distributional assumptions. Figure 4.1-4.3 show that the residuals are considerably heavy-tailed than normal which supports our assumption of long-tailed symmetric distribution. Even for Apolipoprotein AU data of which %78.5 of genes are satisfying the normality assumption, the Q-Q plot indicates that it is apparently not a normal distribution. Moreover, the skewness values of these three datasets are 0.042, 0.067, 0.093 whereas the kurtosis values are 8.629, 8.743 and 7.506; the shape parameters, p are 3.13, 3.28 and 3.20, respectively. These values satisfy the equality concerning the kurtosis value for long-tailed symmetric family given in Section 2.1.

By comparing the performances of seven different methods by using the reference genes from each dataset, we have seen that AMMLE and MMLE gives consistently good performance regardless of the sample size and the distributional assumptions. Also it performs much better than the other methods for small sample cases which are more common than large samples in the microarray experiments.

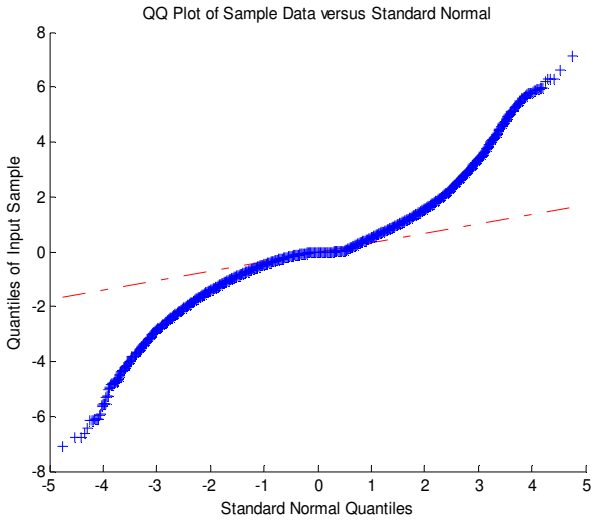

Figure 4.1 The Q-Q plot of leukemia data

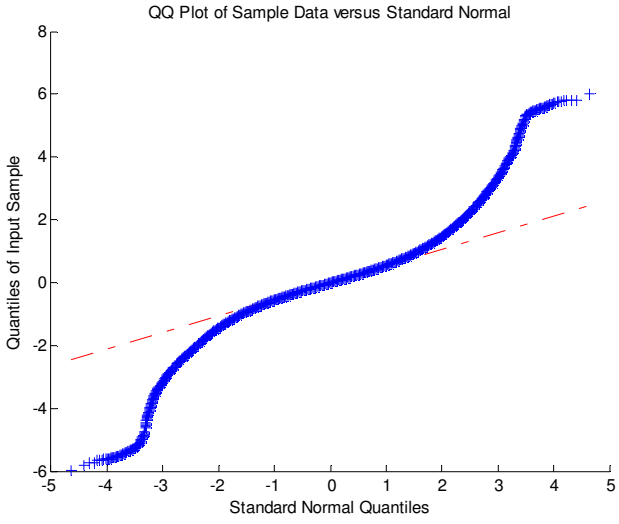

Figure 4.2 The Q-Q plot of melanoma data

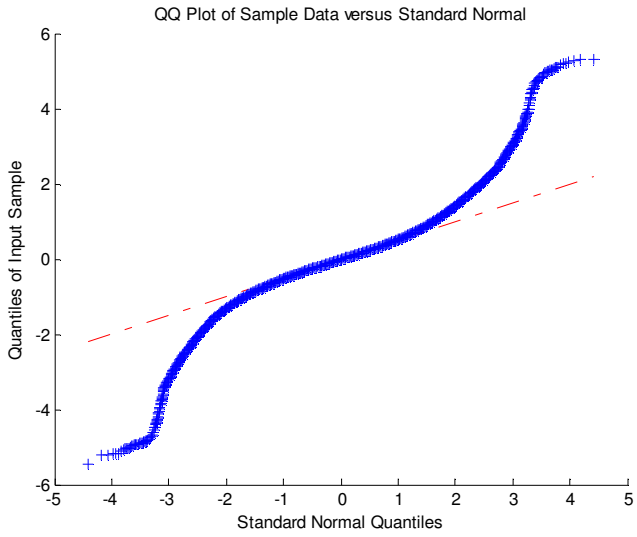

Figure 4.3 The Q-Q plot of apolipoprotein AI mouse data

4.2 Comparisons via Simulated Datasets

We carried out an extensive simulation study to evaluate each of the seven methods discussed in previous methods. It should be noted that the simulations in this section are different in the aspect of data generation from the ones we discussed in Chapter 2. In this section, SIMAGE (Albers, 2006), a software for simulation of microarray gene expression data is facilitated in order to mimic real nature of microarray data as close as possible.

4.2.1 Simulations

We generated 10000 genes from selected large (20 & 15 arrays) and small size samples (5 & 5 arrays). Simulated data contained 5% changed genes out of these 10000 genes. Since ANOVA and SAM methods require equal variance assumption under the null hypothesis, to check their robustness to the assumption violation, we also considered the case where the two distributions have different variances.

4.2.2 Simulation Results

The number of true positive genes and the average ranks for various methods among the top 500 (%5 of 10000 genes) ranked genes were compared using simulation study. Table 4.2 and Table 4.3 show the main results of simulation study. It should be noted that a higher number of true positives and a lower average rank implies a better estimation method since a true positive gene is a statistically significant gene which is truly differentially expressed.

Table 4.2 The number of true positives and the average when the variances are same under the null hypothesis.

Method	(20, 15) arrays		(5,5) arrays	
	True Positives	Average Rank	True Positives	Average Rank
AMMLE	498	252.98	369	851.25
MMLE	497	253.45	363	851.48
W24	495	252.80	362	853.56
ANOVA	484	253.96	282	767.64
SAM	479	255.84	230	1123.80
t-test	463	265.48	271	984.04
Bayes t	460	286.68	353	848.22

Table 4.2 shows the simulation results when the two groups have equal variances. It indicates that AMMLE, W24 and MMLE perform well when there are 20 and 15 samples in each group. For the dataset containing 5 samples per each group, AMMLE appears to perform well whereas the ANOVA, SAM and t-test seems to be poor compared to their performances for large samples. AMMLE appears to be the best in both large (20 & 15 arrays) and small sample (5 & 5 arrays) cases.

Table 4.3 shows the simulation results when the two groups have unequal variances. As shown in Table 4.3, the violation of the assumption of equal variance makes non-ignorable effects on the performance of testing methods. AMMLE appears to be the best for both large and small sample cases. ANOVA, SAM and t-test seems to

be poor compared to their performances under the assumption of equal variances for both large and small sample cases.

Table 4.3 The number of true positives and the average when the variances are different under the null hypothesis.

Method	(20, 15) arrays		(5,5) arrays	
	True Positives	Average Rank	True Positives	Average Rank
AMMLE	481	268.53	293	1182.88
MMLE	475	266.89	288	1181.56
W24	476	267.02	288	1182.05
ANOVA	456	284.18	259	1134.08
SAM	441	295.11	186	1276.30
t-test	429	319.57	231	1391.99
Bayes t	464	271.32	276	1177.03

Through our comparison study, we can see that the performance of testing methods is affected by sample size, distributional assumption and variance structure. Therefore applying the most appropriate testing method under the given situation is very important for the analysis of microarray data. As the results of our study imply, estimation and hypothesis testing methods based on AMML and MML estimators seem appropriate choices for microarray data analysis since they perform better than the other five methods in finding the significant genes and are also robust to the deviations from the assumed situations.

CHAPTER 5

SUMMARY AND CONCLUSIONS

In the framework of the differential gene expression analysis, the biological background of genes, DNA and RNA molecules is given and issues about data analysis preparation, statistical techniques used for analysis of microarray data and multiple testing procedures are explored.

The distribution of the microarray data is determined as a distribution from the LTS family and theoretical background for LTS family is presented in detail. In the framework of unbalanced two-way classification model with interaction for the microarray data under the assumption of LTS distributed error terms, the model parameters are estimated by using the MML estimation method. MML method is theoretically and computationally straightforward besides being flexible in the sense that it can be used for location-scale distributions, symmetric or skew. It also provides explicit solutions for the likelihood equations when Fisher method of maximum likelihood becomes intractable.

The W statistics for testing main and interaction effects are developed and a simulation study is carried out to analyze the efficiency and robustness of the estimators as well as the test statistics.

By using robust estimators of location and scale parameters such as MML and Huber's M-estimators, a test statistic is obtained to compare the treatment means under long-tailed symmetric distribution. To examine power and robustness properties of the test statistic, the simulation study is conducted.

When a statistician has no opportunity to investigate the nature of the underlying distribution, Adaptive Modified Maximum Likelihood estimators are used. The AMML estimators for unbalanced two-way classification model with interaction are derived. The efficiency properties of AMML, MML and Huber's W24 estimators are compared. Moreover, the pairwise multiple comparison procedure is conducted via AMML estimators and power and robustness properties of test statistics based AMML and Huber's W24 estimators are examined.

Six types of parametric methods (t-test, Bayes t-test, ANOVA, Huber estimation, MMLE and AMMLE) and one non-parametric method (SAM) are compared by using both the three real microarray experiments and the simulated datasets are compared.

On the basis of this research, the following conclusions can be stated:

1) The MML estimators $\hat{\mu}$, $\hat{V}$, $\hat{G}$, $(\widehat{VG})$ and $\hat{\sigma}$ are unbiased and considerably more efficient than the corresponding LS estimators even for small sample sizes. The LS estimators have a disconcerting feature, i.e., their relative efficiency decreases as the sample size increases. For small p values which are more appropriate for heavy-tailed microarray data, MML estimators are enormously more efficient than LS estimators

2) The W-test has smaller Type I error and it is clearly more powerful than the traditional F-test (even for approximately normal distribution when $p = 10$).

3) The T-test developed for pairwise multiple comparisons of the treatment means maintains higher power compared to t-test. Also, it has smaller Type I error than the t-test.

4) The MML estimators and the test statistics obtained by using MML estimators are robust to deviations from the assumed distribution.

5) The AMML estimators $\hat{\mu}^a$, $\hat{V}^a$, $\hat{G}^a$, $(\widehat{VG})^a$ and $\hat{\sigma}^a$ are considerably more efficient than LS estimators even for small sample sizes. The relative efficiencies of LS estimators, $\tilde{\mu}$, $\tilde{V}$, $\tilde{G}$ and $(\widetilde{VG})$ decreases as sample size increases.

6) The T^a-test obtained for pairwise multiple comparisons of the treatment means by using AMML estimators has higher power than t^{W24}-test obtained by using W24 estimators. Moreover, it has smaller Type I error than the t^{W24}-test.

7) The AMML estimators and the test statistics obtained by using AMML estimators are robust to deviations from the assumed distribution.

8) When compared using both the three real microarray experiments and the simulated datasets, estimation and testing procedures based on AMML and MML estimation methods seem appropriate choices for microarray data analysis since in general they perform better than W24, ANOVA, SAM, t-test, Bayes t-test methods in finding the significant genes. AMML and MML methods are also robust to the deviations from the assumed situations.

As a future research, we'll compare the efficiency properties of $\hat{V}^a$, $\hat{G}^a$, $(\widehat{VG})^a$ with the corresponding W24 estimators since in this study we just compared the properties of $\hat{\mu}^a$ and $\hat{\sigma}^a$. Moreover, this study is planned to be extended by facilitating the mixed model approach as a future research.

REFERENCES

Akkaya, A. D. and Tiku, M. L. (2008a). Robust estimation in multiple linear regression model with non-Gaussian noise. Automatica, 44, 407-417.

Akkaya, A. D. and Tiku, M. L. (2011). Adaptive estimation and hypothesis testing for AR(1) models. JISAS (to be published).

Amaratunga, D. and Cabrera, C. (2004). Exploration and Analysis of DNA Microarray and Protein Array Data. Wiley–Interscience: New Jersey.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). Robust Estimates of Location. Princeton University Press: Princeton.

Andrews, D. F. (1974). A robust method for multiple linear regression. Technometrics, 16, 523-531.

Baldi, P. and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics, 17, 509-519.

Beaton, A. E. and Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. Technometrics, 16-147-186.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal-Royal Statistical Society Series B, 57, 289-300.

Bhattacharyya, G. K. (1985). The asymptotics of maximum likelihood and related estimators based on type II censored data. J. Amer. Statist. Assoc., 80, 398-404.

Birch, J. B. and Myers, R. H. (1982). Robust analysis of covariance. Biometrics, 38, 699-713.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., and Sondak, V. (2000). Molecular classification of cutanbeous malignant melanoma by gene expression. Nature, 406, 536-540.

Blazejczyk, M., Miron, M., and Nadon, R. (2007). FlexArray: A statistical data analysis software for gene expression microarrays (online). Genome Quebec, Montreal, Canada, URL: http://genomequebec.mcgill.ca/FlexArray (accessed 03/08/2010).

Box, G. E. P. and Andersen, S. L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. J. Roy. Statist. Soc., B 17, 1-34.

Box, G. E. P. and Watson, G. S. (1962). Robustness to non-normality of regression tests. Biometrika, 49, 93-106.

Broberg, P. (2002). Ranking genes with respect to differential expression. Genome Biology, 3.

Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P., and Rubin, E. M. (200). Microarray expression profiling identifies genes with altered expression in HDL deficient mice. Nature, 406, 536-540.

Churchill, G. A. (2002). Fundamentals of experimental design for cDNA microarrays. Nature Genet., 29, 355-356.
David, F. N. and Johnson, N. L. (1951). The effect of non-normality on the power function of the F-test in the analysis of variance. Biometrika, 58, 43-57.

Donaldson, T. S. (1968). Robustness of the F-test to errors of both kinds and the correlation between the numerator and denominator of the F-ratio. J. Amer. Statist. Assoc., 63, 600-676.

Dönmez, A. (2010). Adaptive estimation and hypothesis testing methods. Ph.D Thesis, Middle East Technical University: Ankara.

Dunnett, C. W. (1982), Robust multiple comparisons. Commun. Statist.-Theor. Meth., 11 (22), 2611-2629.

Eisen, M. (1999). Cluster and tree view manual. Standford University.

Gayen, A. K. (1950). The distribution of the variance ratio in random samples of any size drawn from non-normal universes. Biometrika, 37, 236-255.

Geary, R. C. (1947). Testing for normality. Biometrika, 34, 209-242.

Golub, T. R., Slonim, D. K., Tamajo, P., Huard, C., Gaosenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Dowing, J. R., Caliguri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286, 531-537.

Göhlmann, H. and Talloen, W. (2009). Gene Expression Studies Using Affymetrix Microarrays. Chapman & Hall/CRC: New York.

Gross, A. M. (1976). Confidence interval robustness with long tailed symmetric distributions. J. Amer. Statist. Assoc., 71, 409-416.
Gross, A. M. (1977). Confidence intervals for bisquare regression estimates. J. Amer. Statist. Assoc., 72, 341-354.

Gyorffy, B., Molnar, B., Lage, H., Szallasi, Z., and Eklund, C. E. (2009). Evaluation of microarray processing algorithms based on concordance with RT-PCR in clinical samples. Plos One, 5, 1-6.

Hack, H. R. B. (1958). An empirical investigation into the distribution of the F-ratio in samples from two non-normal populations. Biometrika, 45, 260-265.

Hamilton, L. C. (1992). Regression with Graphics: A Second Course in Applied Statsistics. Brroks/Cole: California.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. J. Amer. Statist. Assoc., 62, 1179-1186.

Hampel, F. R., Ronchetti, E. M., and Rousseeuw, P. J. (1986). Robust Statistics. John Wiley: New York.

Huber, P. J. (1964). Robust estimation of a location parameter. Ann. Math. Statist., 35, 73-101.

Huber, P. J. (1977). Robust Statistical Procedures. Regional Conference Series in Applied Mathematics, 27. Soc. Industr. Appl. Math.: Philadelphia.

Huber, P. J. (1981). Robust Statistics. Wiley: New York.

Islam, M. Q. and Tiku, M. L. (2004). Multiple linear regression model under non-normality. Commun. Stat.-Theory Meth., 33, 2443-2467.

Jeffery, G. T., Olson, J. M., Tapscott, S. J., and Zhao, L. P. (2001). An efficient approach to discover differentially expressed genes using genomic expression profiles. Genome Research, 11, 1227-1236
Kerr, M.K., Martin, M., and Churchill, G.A. (2000) Analysis of variance for gene expression microarray data. J. Comput. Biol., 7(6), 819–837.

Lee, K. R., Kapadia, C. H., and Dwight, B. B. (1980). On estimating the scale parameter of Rayleigh distribution from censored samples. Statist. Hefte, 21, 14-20.

Lee, M.-L. T., Kuo, F. C., Whitmore, G. A., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc. Nat. Acad. Sci. U.S.A., 97, 9834-9839.

Lee, M.-L. T. (2004). Analysis of Microarray Gene Expression Data. Kluwer Academic Publishers: Boston.
Low, B. B. (1959). Mathematics. Neill and Co: Edinburgh.

Neter, J., Wasserman, W., and Kutner, M. H. (1985). Applied Statistical Models. Richard D. Irwin, Inc.

Parmigiani, G., Garrett, E. S., Irizarry, R. A., and Zeger S. L. (2003). The Analysis of Gene Expression Data. Springer: New York.

Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. Biometrika, 23, 114-133.

Puthenpura, S. and Sinha, N. K. (1986). Modified maximum likelihood method for the robust estimation of system parameters from very noisy data. Automatica, 22, 231-235.

Reiner A., Yekutieli, D., and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 2003, 19, 368-375.

Sapir, M. and Churchill, G. A. (2000). Estimating the posterior probability of gene expression from microarray data. Unpublished.

Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., and Herzel, H. (2000). Normalization strategies for cDNA microarrays. Nucleic Acids Res., 28, e47.

Smith, W. B., Zeis, C. D., and Syler, G. W. (1973). Three parameter lognormal estimation from censored data. J. Indian Statistical Association, 11, 15-31.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology., 3.

Srivastava, A. B. L. (1959). Effect of non-normality on the power of the analysis of variance test. Biometrika, 46, 114-122.

Staudte, R. G. and Sheather, S. J. (1990). Robust Estimation and Testing. John Wiley & Sons: New York.

Tan, W. Y. (1985). On Tiku's robust procedure-a Bayesian insight. J. Statist. Plann. and Inf., 11, 329-340.

Tiku, M. L. (1964). Approximating the general non-normal variance ratio sampling distributions. Biometrika, 51, 83-95.

Tiku, M. L. (1967). Estimating the mean and Standard deviation from censored normal samples. Biometrika, 54, 155-165.

Tiku, M. L. (1968). Estimating the parameters of log-normal distribution from censored sample. J. Amer. Stat. Assoc., 63, 134-140.

Tiku, M. L. (1971). Power function of the F-test under non-normal situations. J. Amer. Statist. Assoc., 66, 913-916.

Tiku, M. L. (1980). Robustness of MML estimators based on censored samples and robust test statistics. J. Stat. Plann. Inf., 4, 123-143.

Tiku, M. L. and Kumra, S. (1981). Expected values and variance and covariances of order statistics for a family of symmetric distributions (Student's t). Selected Tables in Mathematica Statistics, 8, 141-270. American Mathematical Society, Providence, RI.

Tiku, M. L., Tan, W. Y., and Balakrishnan, N. (1986). Robust Inference. Marcel Dekker: New York.

Tiku, M. L. (1988). Order statistics in goodness of fit tests. Commun. Statist.-Theor. Meth., 17, 2369-2387.

Tiku, M. L. and Suresh, R. P. (1992). A new method of estimation for location and scale parameters. J. Stat. Plan. Inf., 30, 281-292.

Tiku, M. L. and Akkaya, A. D. (2004). Robust Estimation and Hypothesis Testing. New Age International Limited, Publishers: New Delhi.

Tiku, M. L. and Surucu, B. (2009). MMLEs are as good as M-estimators or better. Statistics and Probability Letters, 79, 984-989.

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

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A, 98, 5116-5121.

Vaughan, D. C. (1992a). On the Tiku-Suresh method of estimation. Commun. Statist. Theory Meth., 21, 451-469.

Vaughan, D. C. and Tiku, M. L. (2000). Estimation and hypothesis testing for a non-normal bivariate distribution with applications. J. Mathematical and Computer Modeling, 32, 53-67.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. S. (2001). Asssessing gene significance from cDNA microarray expression data via mixed models. J. Comput. Biol., 8, 625-637.

Yang, Y. H. and Speed, T. (2002). Design issues for cDNA microarray experiments. Nature Rev. Genet., 3, 579-588.

APPENDIX A

MATLAB CODE FOR ESTIMATION AND HYPOTHESIS TESTING FOR UNBALANCED TWO-WAY ANOVA WITH INTERACTION MODEL BASED ON MML TECHNIQUE

```
clear all
% Before compiling this program, data should have saved as a mat
% file where rows denote genes and columns denote varieties.
% First n(i) columns should coreespond to the n(i)
% replications of the i-th variety
load data;
K=input('number of varieties K=');
G=input('number of genes G=');
for i=1:K
   n(i)=input('Number of replications for varieties respectively =')
end
N=G*(sum(n));

% Matrix of replication indices for different varieties
nn=[];
nn(1)=1;
nn(2)=n(1);
for i=2:K
   nn(2*i-1)=nn(2*i-2)+1;
   nn(2*i)=nn(2*i-2)+n(i);
end

% LSE of mu
sum_y=sum(sum(y));
mu_lse=sum_y/N;
V_lse=[];
G_lse=[];
VG_lse=[];
% LSE of V
for k=1:K
   sum1=0;
```

```
    for g=1:G
        for l=nn(2*k-1):nn(2*k);
            sum1=sum1+y(g,l);
        end
    end
    V_lse(k)=sum1/(G*n(k))-mu_lse;
end

% LSE of G
G_lse=(sum(y')/sum(n))-mu_lse;

% LSE of VG
for k=1:K
    for g=1:G
        sum1=0;
        for l=nn(2*k-1):nn(2*k);
            sum1=sum1+y(g,l);
        end
    VG_lse(g,k)=sum1/n(k)-mu_lse-V_lse(k)-G_lse(g);
    end
end

% Computing residuals
 r=[];
 for k=1:K
    for l=nn(2*k-1):nn(2*k);
        for g=1:G
        r(g,l)=y(g,l)-mu_lse-V_lse(k)-G_lse(g)-VG_lse(g,k);
     end
    end
 end
 e=[];
 for i=1:sum(n)
     e=[e;r(:,i)];
 end
 skw=skewness(e);
 kur=kurtosis(e);

% MLE of sigma
 sigma_mle=(sum(e.^2))/(N-(K*G));

% MML (general)
 y_sorted=[];
 for k=1:K
```

```
    y_sorted=[y_sorted sort(y(:,nn(2*k-1):nn(2*k)),2)];
 end
 j=1;
 for p=1.6:0.5:6;
    q=2*p-3;
    t=[];
    alpha=[];
    delta=[];
    t=zeros(max(n),K);
    alpha=zeros(max(n),K);
    delta=zeros(max(n),K);
    for k=1:K
        t(1:n(k),k)=lts_t(n(k),p);
        for l=1:n(k)
            delta(l,k)=(1-(t(l,k)^2)/q)/((1+(t(l,k)^2)/q)^2);
            alpha(l,k)=(2*(t(l,k)^3)/q)/((1+(t(l,k)^2)/q)^2);
        end
    end

% Computing MML of mu
sum_GKL=0;
for k=1:K
   for g=1:G
      for l=1:n(k)
         sum_GKL=sum_GKL+delta(l,k)*y_sorted(g,(nn(2*k-1)+l-1));
      end
   end
end
mu_MML=sum_GKL/(G*sum(sum(delta)));

% Computing MML of V
V_MML=[];
for k=1:K
   sum_GL=0;
   for g=1:G
      for l=1:n(k)
         sum_GL=sum_GL+delta(l,k)*y_sorted(g,(nn(2*k-1)+l-1));
      end
   end
   V_MML(k)=(sum_GL/(G*sum(delta(:,k))))-mu_MML;
end

% Computing MML of G
G_MML=[];
```

```
for g=1:G
   sum_KL=0;
   for k=1:K
      for l=1:n(k)
         sum_KL=sum_KL+delta(l,k)*y_sorted(g,(nn(2*k-1)+l-1));
      end
   end
   G_MML(g)=(sum_KL/sum(sum(delta)))-mu_MML;
end

% Computing MML of VG
VG_MML=[];
for k=1:K
   for g=1:G
      sum_L=0;
      for l=1:n(k)
         sum_L=sum_L+delta(l,k)*y_sorted(g,(nn(2*k-1)+l-1));
      end
      VG_MML(g,k)=(sum_L/sum(delta(:,k)))-mu_MML-V_MML(k)-
G_MML(g);
   end
end

% Computing MML of sigma
B=0;
C=0;
for k=1:K
   for g=1:G
      for l=1:n(k)
         B=B+alpha(l,k)*(y_sorted(g,(nn(2*k-1)+l-1))-mu_MML-
V_MML(k)-G_MML(g)-VG_MML(g,k));
         C=C+delta(l,k)*((y_sorted(g,(nn(2*k-1)+l-1))-mu_MML-
V_MML(k)-G_MML(g)-VG_MML(g,k))^2);
      end
   end
end
 B=(2*p/q)*B;
 C=(2*p/q)*C;
 sigma_MML=(-B+sqrt((B^2)+(4*N*C)))/(2*sqrt(N*(N-(K*G))));

% Finding p that maximizes lnL
 L=0;
 for k=1:K
   for g=1:G
```

```
        for l=1:n(k)
          L=L+log((((y_sorted(g,(nn(2*k-1)+l-1))-mu_MML-V_MML(k)-
G_MML(g)-VG_MML(g,k))^2)/q)+1);
        end
    end
 end
 Z=(-1*log(q))-log(beta(0.5,p-0.5))-log(sigma_MML)-((p/N)*L);
 ln_L(j,1)=p;
 ln_L(j,2)=Z;
 j=j+1;
 end

% MML (final)
[maxln_L,I]=max(ln_L(:,2));
p=ln_L(I,1)
q=2*p-3;
     t=[];
     alpha=[];
     delta=[];
     t=zeros(max(n),K);
     alpha=zeros(max(n),K);
     delta=zeros(max(n),K);
     for k=1:K
         t(1:n(k),k)=lts_t(n(k),p);
         for l=1:n(k)
             delta(l,k)=(1-(t(l,k)^2)/q)/((1+(t(l,k)^2)/q)^2);
             alpha(l,k)=(2*(t(l,k)^3)/q)/((1+(t(l,k)^2)/q)^2);
         end
     end

% Computing MML of mu
sum_GKL=0;
for k=1:K
    for g=1:G
        for l=1:n(k)
            sum_GKL=sum_GKL+delta(l,k)*y_sorted(g,(nn(2*k-1)+l-1));
        end
    end
end
mu_MML=sum_GKL/(G*sum(sum(delta)));

% Computing MML of V
V_MML=[];
for k=1:K
```

```
    sum_GL=0;
    for g=1:G
        for l=1:n(k)
            sum_GL=sum_GL+delta(l,k)*y_sorted(g,(nn(2*k-1)+l-1));
        end
    end
    V_MML(k)=(sum_GL/(G*sum(delta(:,k))))-mu_MML;
end

% Computing MML of G
G_MML=[];
for g=1:G
    sum_KL=0;
    for k=1:K
        for l=1:n(k)
            sum_KL=sum_KL+delta(l,k)*y_sorted(g,(nn(2*k-1)+l-1));
        end
    end
    G_MML(g)=(sum_KL/sum(sum(delta)))-mu_MML;
end

% Computing MML of VG
VG_MML=[];
for k=1:K
    for g=1:G
        sum_L=0;
        for l=1:n(k)
            sum_L=sum_L+delta(l,k)*y_sorted(g,(nn(2*k-1)+l-1));
        end
        VG_MML(g,k)=(sum_L/sum(delta(:,k)))-mu_MML-V_MML(k)-
G_MML(g);
    end
end
% Computing MML of sigma
B=0;
C=0;
for k=1:K
    for g=1:G
        for l=1:n(k)
            B=B+alpha(l,k)*(y_sorted(g,(nn(2*k-1)+l-1))-mu_MML-
V_MML(k)-G_MML(g)-VG_MML(g,k));
            C=C+delta(l,k)*((y_sorted(g,(nn(2*k-1)+l-1))-mu_MML-
V_MML(k)-G_MML(g)-VG_MML(g,k))^2);
        end
```

```
    end
end
 B=(2*p/q)*B;
 C=(2*p/q)*C;
 sigma_MML=(-B+sqrt((B^2)+(4*N*C)))/(2*sqrt(N*(N-(K*G))));
 v=2*p-1;
 lts_data=trnd(v,1,N)*sqrt(q/v)*sigma_MML;
 pe=0.05:0.01:0.995;
 q_lts=quantile(lts_data,pe);
 q_data=quantile(e,pe);
 plot(q_lts,q_data,'*');
 R=corrcoef(q_lts,q_data);
 R_square=R.^2;
%Variances of LSE and MMLE(multiplied by 1/sigma^2)
var_mu_lse=1/N;
var_mu_MML=((q^(3/2))*(p+1))/(2*N*p*(p-1/2));
var_V_lse=[];
var_V_MML=[];
var_G_lse=[];
var_G_MML=[];
var_VG_lse=[];
var_VG_MML=[];
for k=1:K
    var_V_lse(k)=(sum(n)-n(k))/(G*n(k)*sum(n));
    var_V_MML(k)=((q^(3/2))*(p+1))/(2*G*n(k)*p*(p-1/2));
end
for g=1:G
    var_G_lse(g)=(G-1)/N;
    var_G_MML(g)=((q^(3/2))*(p+1))/(2*sum(n)*p*(p-1/2));
end
for k=1:K
    for g=1:G
        var_VG_lse(g,k)=(N-sum(n)-(G*n(k)))/(N*n(k));
        var_VG_MML(g,k)=((q^(3/2))*(p+1))/(2*n(k)*p*(p-1/2));
    end
end
var_VG_MML=var_VG_MML.*(sigma_MML^2);
 % Hypothesis testing (W-test)
V_test=0;
for k=1:K;
    V_test=V_test+(sum(delta(:,k))*(V_MML(k)^2));
end
V_test=((2*p/q)*G*V_test)/(sigma_MML^2*(K-1));
G_test=0;
```

```
for g=1:G;
   G_test=G_test+(G_MML(g)^2);
end
G_test=sum(sum(delta))*G_test;
G_test=((2*p/q)*G_test)/(sigma_MML^2*(G-1));
VG_test=0;
for k=1:K;
   for g=1:G;
      VG_test=VG_test+((VG_MML(g,k)^2)*sum(delta(:,k)));
   end
end
VG_test=((2*p/q)*VG_test)/(sigma_MML^2*(G-1)*(K-1));
p_V_test=p_W(V_test);
p_G_test=p_W(G_test);
p_V_test=p_W(V_test);
p_VG_test=p_W(VG_test);

% Pairwise multiple comparisons (MML)
MML_t_test=[];
MML_group_mean=[];
MML_group_var=[];
for k=1:K
   for g=1:G
      sum_L=0;
      for l=1:n(k)
         sum_L=sum_L+delta(l,k)*y_sorted(g,(nn(2*k-1)+l-1));
      end
      MML_group_mean(g,k)=(sum_L/sum(delta(:,k)));
      MML_group_var(g,k)=(q*(sigma_MML)^2)/(2*p*sum(delta(:,k)));
   end
end

for i=1:K;
   for j=1:K;
      if i<j
         l=l+1;
         MML_t_test(:,l)=(MML_group_mean(:,i)-
MML_group_mean(:,j))./(sqrt(MML_group_var(:,i)/n(i)+MML_group_var
(:,j)/n(j)));
         df_t_test(l)=n(i)+n(j)-2;
      end
   end
end
p_t_test=p_t(MML_t_test);
```

APPENDIX B

MATLAB CODE FOR ESTIMATION AND HYPOTHESIS TESTING FOR UNBALANCED TWO-WAY ANOVA WITH INTERACTION MODEL BASED ON AMML TECHNIQUE

```
clear all
% Before compiling this program, data should have saved as a mat
% file where rows denote genes and columns denote varieties.
% First n(i) columns should coreespond to the n(i)
% replications of the i-th variety

load data;
K=input('number of varieties K=');
G=input('number of genes G=');
for i=1:K
   n(i)=input('Number of replications for varieties respectively =')
end
N=G*(sum(n));

% Matrix of replication indices for different varieties
nn=[];
nn(1)=1;
nn(2)=n(1);
for i=2:K
   nn(2*i-1)=nn(2*i-2)+1;
   nn(2*i)=nn(2*i-2)+n(i);
end

p=16.5;
q=2*p-3;
T0=[];
S0=[];
t=zeros(max(n),K);
for g=1:G
   for k=1:K
```

```
        a=[];
        for l=1:n(k)
            a(g,l)=y_sorted(g,nn(2*k-1)+l-1)
        end
        T0(g,k)=median(a);
        S0(g,k)=1.483*median(abs(a-T0));
        t(g,k)=(a-T0(g,k)
    end
end

    t=[];
    alpha=[];
    delta=[];
    t=zeros(max(n),K);
    alpha=zeros(max(n),K);
    delta=zeros(max(n),K);
    for k=1:K
        t(1:n(k),k)=lts_t(n(k),p);
        for l=1:n(k)
            delta(l,k)=(1-(t(l,k)^2)/q)/((1+(t(l,k)^2)/q)^2);
            alpha(l,k)=(2*(t(l,k)^3)/q)/((1+(t(l,k)^2)/q)^2);
        end
    end

% Computing MML of mu
sum_GKL=0;
for k=1:K
    for g=1:G
        for l=1:n(k)
            sum_GKL=sum_GKL+delta(l,k)*y_sorted(g,(nn(2*k-1)+l-1));
        end
    end
end
mu_MML=sum_GKL/(G*sum(sum(delta)));

% Computing MML of V
V_MML=[];
for k=1:K
    sum_GL=0;
    for g=1:G
        for l=1:n(k)
            sum_GL=sum_GL+delta(l,k)*y_sorted(g,(nn(2*k-1)+l-1));
        end
```

```
    end
    V_MML(k)=(sum_GL/(G*sum(delta(:,k))))-mu_MML;
end

% Computing MML of G
G_MML=[];
for g=1:G
    sum_KL=0;
    for k=1:K
        for l=1:n(k)
            sum_KL=sum_KL+delta(l,k)*y_sorted(g,(nn(2*k-1)+l-1));
        end
    end
    G_MML(g)=(sum_KL/sum(sum(delta)))-mu_MML;
end

% Computing MML of VG
VG_MML=[];
for k=1:K
    for g=1:G
        sum_L=0;
        for l=1:n(k)
            sum_L=sum_L+delta(l,k)*y_sorted(g,(nn(2*k-1)+l-1));
        end
        VG_MML(g,k)=(sum_L/sum(delta(:,k)))-mu_MML-V_MML(k)-
G_MML(g);
    end
end

% Computing MML of sigma
B=0;
C=0;
for k=1:K
    for g=1:G
        for l=1:n(k)
            B=B+alpha(l,k)*(y_sorted(g,(nn(2*k-1)+l-1))-mu_MML-
V_MML(k)-G_MML(g)-VG_MML(g,k));
            C=C+delta(l,k)*((y_sorted(g,(nn(2*k-1)+l-1))-mu_MML-
V_MML(k)-G_MML(g)-VG_MML(g,k))^2);
        end
    end
end
 B=(2*p/q)*B;
```

```
C=(2*p/q)*C;
sigma_MML=(-B+sqrt((B^2)+(4*N*C)))/(2*sqrt(N*(N-(K*G))));
v=2*p-1;
lts_data=trnd(v,1,N)*sqrt(q/v)*sigma_MML;
pe=0.05:0.01:0.995;
q_lts=quantile(lts_data,pe);
q_data=quantile(e,pe);
plot(q_lts,q_data,'*');
R=corrcoef(q_lts,q_data);
R_square=R.^2;

%Variances of LSE and MMLE(multiplied by 1/sigma^2)
var_mu_lse=1/N;
var_mu_MML=((q^(3/2))*(p+1))/(2*N*p*(p-1/2));
var_V_lse=[];
var_V_MML=[];
var_G_lse=[];
var_G_MML=[];
var_VG_lse=[];
var_VG_MML=[];
for k=1:K
   var_V_lse(k)=(sum(n)-n(k))/(G*n(k)*sum(n));
   var_V_MML(k)=((q^(3/2))*(p+1))/(2*G*n(k)*p*(p-1/2));
end
for g=1:G
   var_G_lse(g)=(G-1)/N;
   var_G_MML(g)=((q^(3/2))*(p+1))/(2*sum(n)*p*(p-1/2));
end
for k=1:K
   for g=1:G
      var_VG_lse(g,k)=(N-sum(n)-(G*n(k)))/(N*n(k));
      var_VG_MML(g,k)=((q^(3/2))*(p+1))/(2*n(k)*p*(p-1/2));
   end
end
var_VG_MML=var_VG_MML.*(sigma_MML^2);

% Hypothesis testing (W-test)
V_test=0;
for k=1:K;
   V_test=V_test+(sum(delta(:,k))*(V_MML(k)^2));
end
V_test=((2*p/q)*G*V_test)/(sigma_MML^2*(K-1));
G_test=0;
```

```
for g=1:G;
   G_test=G_test+(G_MML(g)^2);
end
G_test=sum(sum(delta))*G_test;
G_test=((2*p/q)*G_test)/(sigma_MML^2*(G-1));
VG_test=0;
for k=1:K;
   for g=1:G;
      VG_test=VG_test+((VG_MML(g,k)^2)*sum(delta(:,k)));
   end
end
VG_test=((2*p/q)*VG_test)/(sigma_MML^2*(G-1)*(K-1));
p_V_test=p_W(V_test);
p_G_test=p_W(G_test);
p_V_test=p_W(V_test);
p_VG_test=p_W(VG_test);

% Pairwise multiple comparisons (MML)
MML_t_test=[];
MML_group_mean=[];
MML_group_var=[];
for k=1:K
   for g=1:G
      sum_L=0;
      for l=1:n(k)
         sum_L=sum_L+delta(l,k)*y_sorted(g,(nn(2*k-1)+l-1));
      end
      MML_group_mean(g,k)=(sum_L/sum(delta(:,k)));
      MML_group_var(g,k)=(q*(sigma_MML)^2)/(2*p*sum(delta(:,k)));
   end
end
for i=1:K;
   for j=1:K;
      if i<j
         l=l+1;
         MML_t_test(:,l)=(MML_group_mean(:,i)-
MML_group_mean(:,j))./(sqrt(MML_group_var(:,i)/n(i)+MML_group_var
(:,j)/n(j)));
         df_t_test(l)=n(i)+n(j)-2;
      end
   end
end
p_t_test=p_t(MML_t_test);
```

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Ülgen, Burçin Emre
Nationality: Turkish (TC)
Data and Place of Birth: 11 September 1982, Ankara
email: b.e.ulgen@gmail.com

EDUCATION

Degree	Institution	Year of Graduation
MS	METU Statistics	2005
BS	METU Statistics	2002
High School	Özel Yükseliş Lisesi, Ankara	1998

Academic Experience

Year	Place	Enrollment
2002-2009	METU Department of Statistics	Research Asst.

FOREIGN LANGUAGES

English (advanced)

Conference Proceedings

1. Ulgen, B. E. (2009). Analysis of Variance in Microarray Data with Replication Proceedings, 57th Session of the International Statistical Institute, 242, South Africa.

2. Ulgen, B. E., Akkaya, A., Sener, C., and Kocair, C. (2009). Seismic Risk Assessment: A Grid-Based Approach for the South-East European Region SEE-GRID-SCI User Forum, Istanbul.