MIXTURE MODELING AND OUTLIER DETECTION IN MICROARRAY DATA ANALYSIS

A Dissertation

by

NYSIA I. GEORGE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2008

Major Subject: Statistics

MIXTURE MODELING AND OUTLIER DETECTION IN MICROARRAY DATA ANALYSIS

A Dissertation

by

NYSIA I. GEORGE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Naisyin Wang
Committee Members,	Raymond J. Carroll
	Robert Chapkin
	Erning Li
	F. Michael Speed
Head of Department,	Simon J. Sheather

August 2008

Major Subject: Statistics

ABSTRACT

Mixture Modeling and Outlier Detection in Microarray Data Analysis.

(August 2008)

Nysia I. George, B.S., Texas A&M University;

M.S., Texas A&M University

Chair of Advisory Committee: Dr. Naisyin Wang

Microarray technology has become a dynamic tool in gene expression analysis because it allows for the simultaneous measurement of thousands of gene expressions. Uniqueness in experimental units and microarray data platforms, coupled with how gene expressions are obtained, make the field open for interesting research questions. In this dissertation, we present our investigations of two independent studies related to microarray data analysis.

First, we study a recent platform in biology and bioinformatics that compares the quality of genetic information from exfoliated colonocytes in fecal matter with genetic material from mucosa cells within the colon. Using the intraclass correlation coefficient (ICC) as a measure of reproducibility, we assess the reliability of density estimation obtained from preliminary analysis of fecal and mucosa data sets. Numerical findings clearly show that the distribution is comprised of two components. For measurements between 0 and 1, it is natural to assume that the data points are from a beta-mixture distribution. We explore whether ICC values should be modeled with a beta mixture or transformed first and fit with a normal mixture. We find that the use of mixture of normals in the inverse-probit transformed scale is less sensitive

toward model mis-specification; otherwise a biased conclusion could be reached. By using the normal mixture approach to compare the ICC distributions of fecal and mucosa samples, we observe the quality of reproducible genes in fecal array data to be comparable with that in mucosa arrays.

For microarray data, within-gene variance estimation is often challenging due to the high frequency of low replication studies. Several methodologies have been developed to strengthen variance terms by borrowing information across genes. However, even with such accommodations, variance may be inflated by the presence of outliers. For our second study, we propose a robust modification of optimal shrinkage variance estimation to improve outlier detection. In order to increase power, we suggest grouping standardized data so that information shared across genes is similar in distribution. Simulation studies and analysis of real colon cancer microarray data reveal that our methodology provides a technique which is insensitive to outliers, free of distributional assumptions, effective for small sample size, and data adaptive.

To my Lord and Savior, Jesus Christ. I am nothing without you.

ACKNOWLEDGEMENTS

I have tremendous and unreserved appreciation for my advisor, Dr. Naisyin Wang. Finite words cannot express how blessed I am by her mentorship in my life. Dr. Wang is a very special person and many times went above and beyond the call of duty. She saw qualities in me that I never saw in myself and fully committed herself to investing in me as a future colleague. Through her guidance, I have gained confidence to explore and develop my statistical ideas and intuition. I learned a lot through our talks, which always complimented her insight, expertise, and passion for professional growth.

I also salute Drs. Michael Speed and Raymond Carroll for the special role each played during my graduate studies. I thank both of them for being a resource and reaching out to me in ways that stretched far beyond academia. Along with Dr. Speed and Dr. Carroll, I would like to thank the other members of my committee - Dr. Robert Chapkin and Dr. Erning Li. I could not have asked for a more supportive committee.

Praise God for my family who is probably most excited to celebrate with me - I am no longer a professional student! My family has shown me unconditional love and has forever supported my education and career aspirations. I am thankful to have them rooting me on. Undeniably, my mom is the best cheerleader anyone could hope for.

Additionally, I extend a heartfelt thanks to my spiritual family. I could not imagine completing this journey without them. The endless prayer, words of encouragement, and sound advice were my anchor. They were always available when I needed support and most importantly, continuously reminded of the love, grace, and

power of God. I offer special thanks to Danielle and Ra'sheedah, my roommates and spiritual sisters, for their love, friendship, and comic relief.

Finally, I thank God for imparting in me all that He is. He truly opened my eyes to new heights that I can reach through Him. Through this process, He taught me about endurance and standing firmly on the promises He has spoken to me. I am honored that He entrusted me with such an amazing accomplishment.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

TABLE Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Microarray technologies simultaneously measure the expression levels of thousands of genes and are widely used in biomedical research. Because these high-throughput instruments facilitate large-scale experiments and advanced research, microarray data analysis is constantly progressing. New statistical methodologies for analyzing gene expression data are emerging in order to gain biological insights. This dissertation presents two independent studies of microarray data analysis. The methods that we develop for both topics are very applicable and their considerations are necessary to accurately analyze microarray data sets.

Our first work is motivated by the need to evaluate fecal mRNA microarray reproducibility. We study a recent platform in biology and bioinformatics that compares the quality of genetic information from exfoliated colonocytes in fecal matter with genetic material from mucosa cells within the colon. Colon cancer is a leading cause of cancer death and it believed that attitudes towards the colonoscopy are a deterrent for colorectal cancer screening. The goal is to offer patients an alternative so that those who are at-risk for cancer are diagnosed and treated at an early stage. To address the issue of gene reproducibility between the two platforms, we use the intraclass correlation coefficient (ICC) as a measure of reproducibility. For measurements between 0 and 1, it is natural to assume that the data points are from a beta distribution. As an alternative, consider the fact that a uniform random variable after

---

The format and style follow that of *Biometrics*.

being transformed by the inverse-probit function will be normally distributed, where the probit function is the cumulative distribution function (CDF) of the standard normal distribution. This suggests that a plausible distributional assumption for the inverse-probit transformed ICC values is the normal distribution. Both are considered to be acceptable approaches. However, will both lead to comparable results? If not, then under what circumstances does one model fail? These are the questions we wish to address.

Chapters II and III of this dissertation are devoted to our first study. In Chapter II we give a complete description of the problem and discuss the details of each methodology used in the data analysis. In Chapter III we explore whether ICC values should be modeled with a beta mixture or transformed first and fit with a normal mixture. We carry out a simulation study and use the chi-square goodness of fit test to determine the accuracy of each model. We are particularly interested in the effect of transformation under model mis-specification.

Secondly, we introduce statistical tools to improve variance estimation and outlier identification in microarray data. A core goal of microarray analysis is to identify an informative subset of differentially expressed genes under different experimental conditions. Typically, this is done through hypothesis testing, which relies on test statistics that properly summarize and evaluate information in the sample(s). A reliable variance estimator that is applicable to all genes is important for analysis. In microarray analysis we often find that genome-scale expression analysis generates large data sets with a small number of replicates for each gene. The widespread statistical limitations due to low replication make it necessary to devise adaptive methods for estimating gene-specific variance. Further complicating variance estimation is the frequent presence of outliers in microarray data. Not only is outlier identification critical for reliable estimation of variance, but we also require accurate

variance estimation in order to successfully identify outliers.

In Chapter IV, we introduce several variance estimation methodologies and outlier identification procedures. We propose a robust modification of optimal shrinkage variance estimation (Tong and Wang, 2007). Our variance estimator is uninfluenced by outliers and allows for gene-specific, rather than pooled, estimates of variance. Additionally, we stabilize estimators by allowing each variance estimate to be the product of a gene-specific and common variance estimate. In order to increase power, we estimate the common variance term by grouping standardized data so that information shared by genes post-standardization can be more efficiently utilized. For outlier detection we adopt a technique which is based on the false discovery rate approach.

Chapter V describes the setup of two simulation studies. The first setup does not assume any relationship between gene variability and location center, while the second is structured such that variance is modeled as a quadratic function of mean. These studies are used to compare the performances of numerous variance estimators and adaptive methodology.

In Chapter VI, we investigate the performance of variance estimation and outlier identification on colon cancer microarray data. We introduce the between extreme expression deviation to MAD (BEED-MAD) ratio statistic as an assessment tool for outlier classification when the truth is unknown.

Finally, we provide a summary of all our findings in Chapter VII.

CHAPTER II

MIXTURE MODELING OF THE INTRACLASS CORRELATION COEFFICIENT

2.1 Introduction

Microarrays, which measure gene expressions at the transcription level where RNA is made from DNA, take us from the days of detecting messenger RNA (mRNA) expression of a single gene to the current stage in which scientists can simultaneously measure the expression of thousands of genes. Daily improvement in this technology frequents the production of new assays and new microarray data platforms. Among them, and of particular interest, is a recent development that enables the collection of genomic information from exfoliated colonocytes in fecal matter. It is known that an early detection of cancerous colon cells results in high cure and survival rates among colon cancer patients. However, people tend to shy away from invasive procedures such as the colonoscopy. Consequently, it is of great interest to develop non-invasive early detection instruments.

Although evidences exist in the fecal platform that partially degraded mRNA in fecal samples can produce meaningful measurements (Schoor et al., 2003), and the conclusions by Davidson et al. (2003) and Kanaoka et al. (2004) suggest that it is possible to isolate intact fecal eukaryotic mRNA, it is unknown whether one can expect the same quality from the large amount of fecal microarray data. The current study, to the best of our knowledge, is the first one that investigates and reports the reproducibility of fecal microarray data.

Biological variation in gene expression data can be assessed with subject to subject replication. In order to determine if one can successfully obtain the same findings

from the same biological sample when the experiment is repeated, it is necessary to determine whether the gene expression levels of a gene from the same subject behave more similarly to each other compared to those of the same gene from different subjects. One can observe this type of similarity even when the biological samples from the same subject are processed through two independent bioassays, as is done for samples using different biological materials. Precisely, under the best scenario that the same biological sample is used, we evaluate the similarity between independent results produced by the same bioassay in a lab. While we focus only on subject to subject variation, we acknowledge that there are other types of replication in gene expression data (Nguyen et al., 2002).

In order to assess the agreement between measurements from microarray data collected from the same subject we use the intraclass correlation coefficient as a reliability index (Carrasco and Jover, 2002). Intraclass correlation (ICC), defined in simplest terms as a measure of reproducibility, is used as a statistical measure to assess methodological and biological variation in DNA microarray analysis. The larger the intraclass correlation coefficient, the more differentiation among gene readings collected from different biological samples relative to that among readings using the same biological material. Thus, an ICC value near 1 signifies a strong indication of reproducibility and agreement between experiments. On the other hand, if the ICC is near 0, then within-subject variance is relatively large compared to between-subject variance and it is likely that one can not obtain the same expression level in a repeated experiment.

Considering how replicative arrays are produced, it may be harder to recognize the phenomenon commonly associated with "reproducibility" when a gene is neither up nor down regulated. Although the same biological materials are produced, the dominating variation could be caused by the two different bioassay processes. If this

is true, then we expect to observe at least a small proportion of genes to always have low reproducibility, thus resulting in a mixture model for the distribution of ICC values. The use of mixture-modeling in bioinformatic research is not new. Researchers have devoted much attention to methodology that can appropriately separate gene expressions into meaningful groups. Allison et al. (2002) and Ji et al. (2005) use beta-mixture modeling to describe distributional properties of different genes' correlation coefficients. Like measurements of ICC, the values of correlation coefficients are between 0 and 1. On the other hand, He et al. (2006) and McLachlan et al. (2006) prefer the use of normal mixture distributions which eliminate the (0,1)-range constraint.

In a study comparing the fecal and mucosa bioarray platform we obtained conflicting results when modeling inverse-probit transformed ICC (IPT-ICC) values with a two-component normal distribution and when modeling ICC values with a two-component beta distribution. It is our conjecture that, considering the boundary problem of the beta distribution, normal mixture modeling might be less sensitive toward model mis-specification. We have observed components of the beta mixture to be strictly decreasing with the density $f(y|\alpha, \beta)$ approaching infinity. This phenomenon causes the maximum likelihood estimate (MLE) of β parameters to be unstable. In order to address which of the two mixture models more accurately analyzes ICC values of gene expression levels, we conduct a simulation study. Our ultimate goal is to select the better of the two systems to ascertain whether the fecal array samples share similar reproducibility as the mucosa array samples.

2.2 Methods

In order to carry out this analysis, we rely on numerous statistical methods. These methodologies are described in detail in the following subsections.

2.2.1 The Intraclass Correlation Coefficient

As with most measurements, measuring gene expression levels even from the same biological materials involves measurement error. In order to assess the agreement between measurements, we look to intraclass correlation coefficients whose use in genomic study was promoted by Carrasco and Jover (Carrasco and Jover, 2003). Intraclass correlation, in its simplest term, is defined as a measure of reproducibility. Consider the following simple model where the response $Y_{ij} = a_i + e_{ij}$ is the jth measurement collected from the ith subject. Further, variables a_i and e_{ij} are independent with means 0 and variance σ_a^2 and σ_e^2, respectively. ICC is the ratio of the variance between subjects to the total variance and is given by the following equation:

$$ICC = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}, \tag{2.1}$$

where σ_a^2 represents the between-subject variation, σ_e^2 is the within-subject variance, and $0 \leq ICC \leq 1$. In the situation that samples from the same subject give an identical reading, then $\sigma_e^2 = 0$ and ICC = 1. We have perfect reproducibility. On the other hand, if all the a_i are close to zero such that the different measurements corresponding to the same subject give the dominating variation, then we expect the ICC to be small.

Here, we use the ICC as a statistical measure to assess methodological and biological variation in DNA microarray analysis. The larger the intraclass correlation coefficient, the more differentiation among gene readings collected from different biological samples relative to that among readings using the same biological material. Thus, an ICC value near 1 signifies a strong indication of reproducibility and agreement between experiments. On the other hand, if the ICC is near 0, then within-subject variance is relatively large compared to between-subject variance and it is likely that the actual gene expression is irreproducible.

Shrout and Fleiss (1979) give guidelines for choosing among six different forms of the ICC, where each form is specifically defined by the experimental design and intent of the study. Since our design has each subject under different treatments measured by 1-2 randomly selected microarrays, we let the measurements from the same subject share the same random intercept and let the different treatments be the fixed effect. We then use this mixed-effects model to obtain the overall and random intercept variation and set them to be the denominator and numerator of the ICC value.

Classifying the ICC as a measure of reproducibility has long been in debate. Lin (1989) discusses two drawbacks that discredits the ICC as a reliable reproducibility index. First, it allows duplicate readings to be interchangeable in the sense that duplicate readings are considered as replicates rather than two distinct readings. Secondly, it is faulted for assessing uncorrelated paired readings with negative values. However, Carrasco and Jover (2003) argue that the ICC is a valid measure of agreement among microarrays and identify it as one of the most popular aggregate procedures used in measuring the agreement of continuous-scaled data. An aggregate procedure is characterized by its use of a single measure to assess agreement, whereas a disaggregate approach calculates agreement for each component of the measurement model separately (Carrasco and Jover, 2003).

2.2.1.1 Obtaining ICC Values for Genes on a Microarray Chip

We define a data observation $Y_{ijk}^{[g]}$ as being the gene expression g for subject i, treatment j, and array k. The observations are modeled by

$$Y_{ijk}^{[g]} = \mu_j^{[g]} + a_i^{[g]} + e_{ijk}^{[g]}, \tag{2.2}$$

for

$$i = 1, 2, \ldots, I, \;\; j = 1, 2, \ldots, J, \;\; \text{and } k = 1, 2, \ldots, K_{ij}$$

This describes a microarray experiment where we consider I subjects, J treatments, and K_{ij} arrays for subject i under treatment j. Also, μ_j is the overall mean for the j^{th} treatment, $a_i \sim N(0, \sigma_a^2)$ is the random effect due to the different subjects, and $e_{ijk} \sim N(0, \sigma_e^2)$ is the random effect due to array to array replication. We assume that the error terms, e_{ijk}, are iid.

After having formulated the model for the data observations taken from the microarray, we can easily characterize the ICC for each gene. The following expression for ICC allows us to quantify the reproducibility index of gene g.

$$ICC_g = \frac{\sigma_{a,g}^2}{\sigma_{a,g}^2 + \sigma_{e,g}^2}, \;\; g = 1, 2, \ldots, G, \tag{2.3}$$

where G is the number of genes.

2.2.2 Two-Component Mixture Models

The numerical findings of ICC and IPT-ICC values clearly show that the data comes from a mixture of two populations. Thus, we analyze the data as clusters to minimize within-group variance and maximize between-group variance.

When data is modeled by a mixture of two distributions we suppose that an observation comes from distribution 1 with probability π or from distribution 2 with probability $1 - \pi$. Suppose Z_i is a random indicator variable such that

$$Z_i = \begin{cases} 1 \;, \;\; \text{with prob } = \pi \\ 0 \;, \;\; \text{with prob } = 1 - \pi \end{cases}$$

Let W_i be the actual outcome observed through the process. Then W_i is distributed as follows:

$$W_i \sim \begin{cases} f_1(w) \;, \;\; Z_i = 1 \\ f_2(w) \;, \;\; Z_i = 0 \,, \end{cases}$$

where f_1 and f_2 are the probability density functions of distributions 1 and 2, respectively. If we consider the joint distribution of (W,Z), then

$$f(w, z) = f(w|z)f(z).$$

Thus,

$$f(w) = \sum_z f(w|z)f(z)$$

and the resulting probability distribution function is given by

$$f(w) = \pi f_1(w) + (1 - \pi)f_2(w) \tag{2.4}$$

Furthermore, if we observe $W = W_1, \ldots, W_n$, the likelihood function is

$$\prod_{i=1}^{n}[\pi f_1(w_i) + (1 - \pi)f_2(w_i)]. \tag{2.5}$$

2.2.3 Parameter Estimation using Expectation-Maximization Algorithm

We use the expectation-maximization (EM) algorithm (Dempster et al., 1977) to obtain parameter estimates of the mixture distributions. The EM algorithm is an iterative approach for estimation of incomplete data problems. Given starting values of the model parameters, the EM algorithm iteratively updates the estimates until a specified convergence is reached. In Sections 2.2.3.1 and 2.2.3.2 we describe procedures for estimating the two-component mixture of beta and normal distributions, respectively.

2.2.3.1 Mixture of Beta Distributions

Suppose $y_1, \ldots y_n$ are n independent observations from $f_Y(y|\theta_B)$, where f_Y is the density of a beta distribution and $\theta_B = (\pi, \alpha_1, \alpha_2, \beta_1, \beta_2)$. Let the random vector $X = (Z, Y) = \{z_i, y_i\}$, where z_i is an indicator variable which assumes the value 1 (0) when the i^{th} observation comes from the first (second) component for $i = 1, \ldots, n$.

In the algorithm, we iteratively perform the "E" and "M" steps with the 'complete' data likelihood function, $L(\theta_B|y_i)$, for θ_B being

$$L(\theta_B|y_i) = \prod_{i=1}^{n} \pi f(y_i|\alpha_1, \beta_1) + (1-\pi) f(y_i|\alpha_2, \beta_2) \tag{2.6}$$

and the corresponding log-likelihood being

$$\ell(\theta_B|y_i) = \sum_{i=1}^{n} log[\pi f(y_i|\alpha_1, \beta_1) + (1-\pi) f(y_i|\alpha_2, \beta_2)]. \tag{2.7}$$

In the E-step, $\mathbf{z}$ is updated with its conditional expectation given the observed data $\mathbf{y}$. Consequently,

$$\begin{aligned} z_i^{(k)} &= E[z_i|y_i, \hat{\pi}^{(k)}, \hat{\alpha}_1^{(k)}, \hat{\alpha}_2^{(k)}, \hat{\beta}_1^{(k)}, \hat{\beta}_2^{(k)}] \\ &= \frac{\hat{\pi}^{(k)} f(y_i|\alpha_1^{(k)}, \beta_1^{(k)})}{\hat{\pi}^{(k)} f(y_i|\alpha_1^{(k)}, \beta_1^{(k)}) + (1-\hat{\pi}^{(k)}) f(y_i|\alpha_2^{(k)}, \beta_2^{(k)})}, \end{aligned} \tag{2.8}$$

where the super index, k, denotes an estimate at the k^{th} iteration.

In the M-step of the EM algorithm we use $z_i^{(k)}$ to estimate the mixing proportion, where

$$\hat{\pi}^{(k+1)} = \frac{\sum_{i=1}^{n} z_i^{(k)}}{n}, \tag{2.9}$$

and obtain maximum likelihood estimates of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, and $\hat{\beta}_2$ numerically. The E- and M-steps are iterated until the convergence criteria is met.

The starting values for α_1, α_2, β_1, and β_2 were set to 0.01 and $\{z_i\}$ was initialized by setting one half of the indicator variables equal to 0 and the other half equal to 1 so that $\hat{\pi}^{(0)}$=0.50. We utilize the 'optim' function in R to obtain parameter estimates for the two beta density functions. The procedure was repeated until we observed a negligible change in the value of the log-likelihood given in (2.7).

2.2.3.2 Mixture of Normal Distributions

Let $x_1, \ldots, x_n$ be n iid observations from $f_X(x|\theta_N)$, where f_X is the density of a normal distribution and $\theta_N = (\pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$. In order to estimate the parameters for a two-component normal mixture, we use the MCLUST software package for R (Fraley and Raftery, 1999). MCLUST implements the EM algorithm (Section 2.2.3.1) to carry out the computations of a maximum likelihood approach for normal mixture modeling. For model selection, Mclust determines the number of clusters and the clustering model by maximizing the Bayesian Information Criterion (BIC) (Schwartz, 1978).

See Fraley and Raftery (1999) and Fraley and Raftery (2002) for more details regarding the MCLUST software package.

2.2.4 Distribution of Transformed Random Variables

In our simulation study of mixture model mis-specification, it is necessary to define the distribution of transformed random variables. In Section 2.2.4.2, we describe the distribution of probit transformed normal random variables (Normal $\rightarrow$ Beta). Likewise, in Section 2.2.4.1 we describe the distribution of inverse-probit transformed beta random variables (Beta $\rightarrow$ Normal).

2.2.4.1 Normal $\rightarrow$ Beta

Let X be a random variable from a two-component normal mixture model with probability density function (pdf) f_N given by

$$f_N(x) = \pi\ \phi(x; \mu_1, \sigma_1^2) + (1 - \pi)\ \phi(x; \mu_2, \sigma_2^2), \tag{2.10}$$

where $0 < \pi < 1$ and $\phi(x; \mu_i, \sigma_i^2)$ is the pdf of a normal random variable with mean μ_i and variance σ_i^2, $i = 1, 2$.

Furthermore, consider transforming the data via the probit transformation given by $Y = \Phi(X)$. Then the density function of Y is given by

$$\begin{aligned} f_B(y) &= f_N(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| \\ &= f_N(g^{-1}(y)) \left| \frac{1}{g'\{g^{-1}(y)\}} \right| \\ &= f_N(\Phi^{-1}(y)) \left| \frac{1}{\phi\{\Phi^{-1}(y)\}} \right|. \end{aligned} \tag{2.11}$$

2.2.4.2 Beta → Normal

Let Y be a random observation from a two-component beta mixture model with pdf f_B given by

$$f_B(y) = \pi \ f(y|\alpha_1, \beta_1) + (1 - \pi) \ f(y|\alpha_2, \beta_2), \tag{2.12}$$

where $0 < \pi < 1$ and

$$\begin{aligned} f(y|\alpha_i, \beta_i) &= \frac{y^{\alpha_i - 1}(1-y)^{\beta_i - 1}}{\int_0^1 t^{\alpha_i - 1}(1-t)^{\beta_i - 1} dt} \\ f(y|\alpha_i, \beta_i) &= \frac{y^{\alpha_i - 1}(1-y)^{\beta_i - 1}}{B(\alpha_i, \beta_i)} \end{aligned} \tag{2.13}$$

is the pdf of a beta random variable with shape parameters α_i, β_i, for $i = 1, 2$

We consider transforming the observations using the inverse-probit transformation by letting $X = g(Y)$ and $g(\cdot) = \Phi^{-1}(\cdot)$. Then the range of X becomes $(-\infty, \infty)$ and its density function is expressed as

$$\begin{aligned} f_N(x) &= f_B(g^{-1}(x)) \left| \frac{d}{dx} g^{-1}(x) \right| \\ &= f_B(\Phi(x)) \left| \phi(x) \right|. \end{aligned} \tag{2.14}$$

2.2.5 Chi-square Goodness of Fit

Let $X_1, \ldots, X_n$ be an observed dataset. Suppose we divide the range of the data into k bins. By comparing the number of observations that fall into a given bin with

the expected number of observations for that bin, we are able to use the Pearson's chi-square (χ^2) goodness of fit test to assess how well the proposed distribution fits the observed data. The χ^2 statistic for testing the null hypothesis H_0 : The data follow the specified distribution, is

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{2.15}$$

where O_i and E_i are the observed and expected, respectively, frequencies for bin i. To ensure that the expected frequency count is never zero, data is binned according to the following quantiles of observed data: 0, $k-1$ equally spaced values between 0.025 and 0.975, and 1. This results in k disjoint bins.

If a dataset is fit with a mixture of normal distributions, then the density function defined in (2.10) is used to determine the expected frequencies. Likewise, we use (2.14) to calculate expected frequencies when a dataset is fit with a mixture of betas. The cdf of (2.14) does not have a closed form solution. Thus, for both distributions, the area of a given bin is approximated with the trapezoidal rule for computing a Riemann sum. The trapezoid approximation of $\int_a^b f(x)dx$ associated with a partition $a = x_0 < x_1 < \ldots < x_n = b$ is

$$T = \frac{1}{2}\left[f(x_0) + 2f(x_1) + \ldots + 2f(x_{n-1}) + f(x_n)\right]\Delta x. \tag{2.16}$$

Each of the k intervals is divided into 4 equal parts so that $n = 4$ and $\Delta x = \frac{b-a}{4}$.

2.2.6 Likelihood Ratio Test

We use the likelihood ratio test in order to test for distributional differences in the reproducibility of fecal and mucosa samples. Let L_1 be the maximum value of the likelihood without placing any assumptions on the model parameters. L_1 is evaluated by substituting the maximum likelihood estimates for the unknown unrestricted

parameters. Let L_0 be the maximum value of the likelihood function when the parameters are restricted by assumptions placed on the model. We can then define the likelihood ratio statistic by

$$\lambda = \frac{L_1}{L_0}.$$

Furthermore, let us assume that k parameters were lost by moving from the unrestricted to the restricted setting. Under the restricted model, as $n \to \infty$, $2 \log \lambda \to \chi^2_k$ in distribution. The likelihood ratio test rejects the testing assumption if $2 \log \lambda > \chi^2_{k,\alpha}$, where α is the level of significance.

CHAPTER III

DATA ANALYSIS AND SIMULATION STUDY OF ICC VALUES

3.1 Introduction

We begin this chapter with a description of the fecal and mucosa data used in the study. In Section 3.3 we present the nature of the problem by showing discrepancies in the beta mixture fit of ICC values and the normal mixture fit of IPT-ICC values. Section 3.4 discusses a simulation study that was carried out to analyze sensitivity to model mis-specification. Finally, in Section 3.5 we compare the ICC distributions of fecal and mucosa samples in order to evaluate the quality of reproducible genes between the two platforms.

3.2 Fecal and Mucosa Data Description

Gene expression levels from the colon mucosa and fecal data samples were collected using the CodeLink System. From the thousands of genes included in both data sets, our working data set of statistically significant genes consisted of 2171 genes for the fecal data and 2241 genes for the mucosa data. The bioassays that were used to extract fecal mRNA were developed much later than the mucosa data used in this study, which was collected earlier in a different experiment. Although we did not have access to the original dataset, the available summary statistics were sufficient for us to produce ICC measurements.

3.2.1 Fecal Data

The fecal array data were collected from rat fecal samples in a study designed to explore the affect that diet has on genes being differentially expressed after exposure

to carcinogen/radiation (Liu et al., 2005). Rats in the study were exposed to carcinogen azoxymethane (AOM) and randomly assigned to one of four different treatments resulting from a 2×2 factorial design. The two experimental factors were diet - fish oil/pectin (D1) and corn oil/ cellulose (D2), and radiation - with radiation exposure (IRT) and without radiation exposure (RCT). Fecal samples were collected 14 weeks after the last exposure to carcinogen AOM. There are respectively 7, 6, 8, and 7 bioarrays collected under IRT-D1, IRT-D2, RCT-D1, and RCT-D2, respectively. Genes which were not disqualified and which had at least 3 usable replicates were kept.

3.2.2 Mucosa Data

Spraque Dawley rats used in the study to obtain mucosa array data were randomly assigned in a $3 \times 2 \times 2$ factorial experiment to a treatment with diet, exposure, and time points as factors (Davidson et al., 2004). Corn oil/n-6 polyunsaturated fatty acid (PUFA) or fish oil/n-3 PUFA or olive oil/n-9 monounsaturated fatty acid (MUFA) was used as the dietary fat source; carcinogen AOM was used as the exposure source; time points were either 12 hours or 10 weeks after the first injection. The units were terminated at the appropriate time point in order to remove the mucosal layer from each colon so that RNA could be extracted from the mucosal samples.

3.3 Preliminary Data Application

The original ICC values were fit with a two-component beta mixture using the EM algorithm, producing the following density estimation for the fecal and mucosa data, f_B^f and f_B^m respectively,

$$f_B^f(.;\hat{\theta}_B) = 0.50\ Beta(0.30, 0.64) + 0.50\ Beta(0.27, 0.63)$$
$$f_B^m(.;\hat{\theta}_B) = 0.53\ Beta(2.20, 2.40) + 0.47\ Beta(0.25, 1.22).$$

After transforming the original ICC values via the inverse-probit transformation, we estimate the following two-component normal mixture densities for the fecal and mucosa data, f_N^f and f_N^m respectively,

$$f_N^f(.;\hat{\theta}_N) = 0.72\ N(0.04, 0.84) + 0.28\ N(-3.50, 0.07)$$
$$f_N^m(.;\hat{\theta}_N) = 0.81\ N(-0.29, 0.64) + 0.19\ N(-3.35, 0.12).$$

A simple observation of the difference in proportion estimates for fecal and mucosa data leads us to question the accuracy of the two fits. It is unclear what the proportion of reproducible genes (upper component of the two mixtures) for fecal samples should be, 0.50 or 0.72? Unfortunately, the answer to this question depends on the mixture model we use to fit the data.

It is well known that when $\alpha < 1$ ($\beta < 1$), the beta distribution increases to infinity at the lower (upper) endpoint, respectively. We find this to be the case with components of the beta mixture for both data sets. This phenomenon is easily seen in the graphs displayed in Figure 1, where we plot the fitted beta mixture superimposed on the histogram of ICC values for the fecal and mucosa data. Because the beta distribution has such a boundary issue, we suspect that a simple violation of distributional assumption near the boundary could have profound effects on maximum likelihood estimates. In comparisons, the fitted normal mixture superimposed on the histogram of IPT-ICC values is plotted in Figure 2. It is worth noting that the visual evaluation of Figures 1 and 2 might not be helpful to the comparisons of these two modeling approaches. We investigate the veracity of the comparisons with numerical studies.

In light of the numerical outcomes from our Monte Carlo investigation, we plot three estimated density functions in Figure 3. The solid curves in each plot of Figure 3 provide the kernel estimated density functions of the fecal and mucosal IPT-ICC

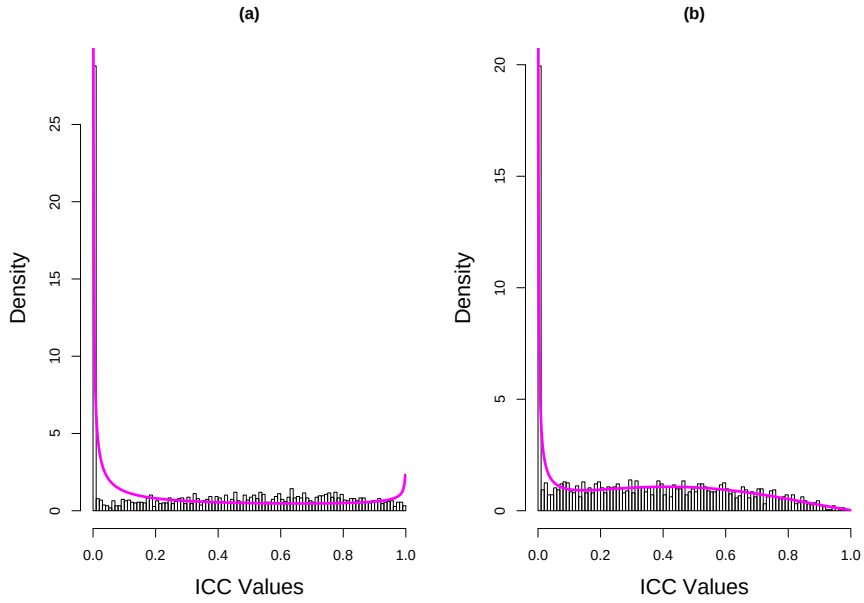


Figure 1: *Histogram of ICC values. The density of the fitted two-component beta mixture to the (a) fecal data and (b) mucosa data is superimposed.*

values, respectively. The estimated density functions based on the normal-mixture models are given by the dashed lines. Finally, the estimated density function calculated using the transformation theory give the estimated density functions of IPT-ICC values in the dotted lines, when the ICC values were fitted with beta-mixtures. Even though not perfectly, the kernel density estimates and the normal-mixture based estimates correspond roughly well with each other. However, the transformed beta-mixture based density estimates misfit the lower mixture component for the mucosa data. For fecal data, this approach almost concluded that there was a single component – a feature which could not be clearly seen in Figure 1.

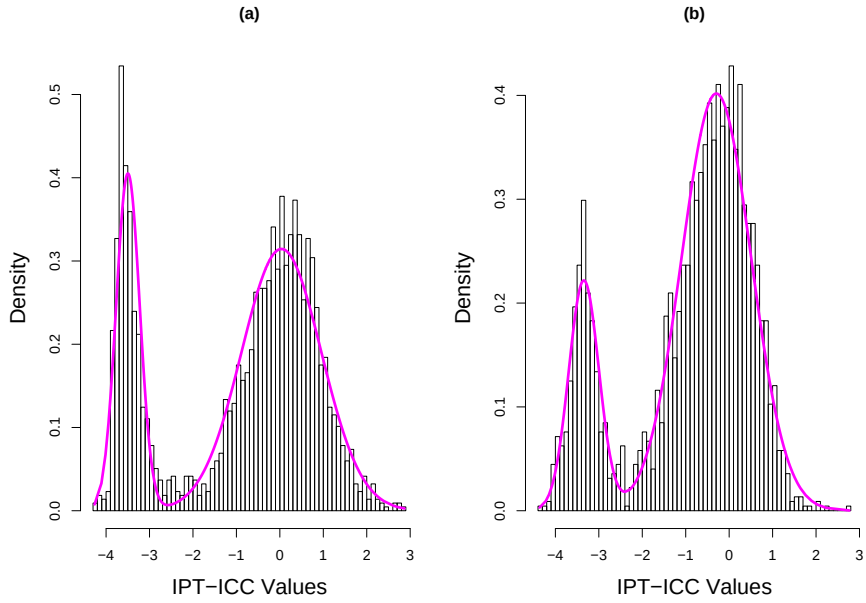


Figure 2: *Histogram of IPT-ICC values. The density of the fitted two-component normal mixture to the (a) fecal data and (b) mucosa data is superimposed.*

3.4 Simulation Study

To investigate the fit of a beta mixture to probit transformed normal random variables and the fit of a normal mixture to inverse-probit transformed beta random variables, we conduct a Monte Carlo simulation study for each of the fecal and mucosa data sets. Our goal is to determine how well the densities fit data under model mis-specification. In other words, we want to assess the loss in accuracy if data is really normal but we transform and fit with beta or otherwise, if data is truly beta but we transform and fit with normal. Simulation for the fecal data is described as follows:

Simulation 1: Data Generated from Beta-mixtures, Fit with Normal-mixtures

(1) Generate $Y_1, \ldots, Y_n$ from $\tilde{f}_B^f = 0.7\ Beta(2.6, 1.7) + 0.3\ Beta(0.2, 0.8)$.

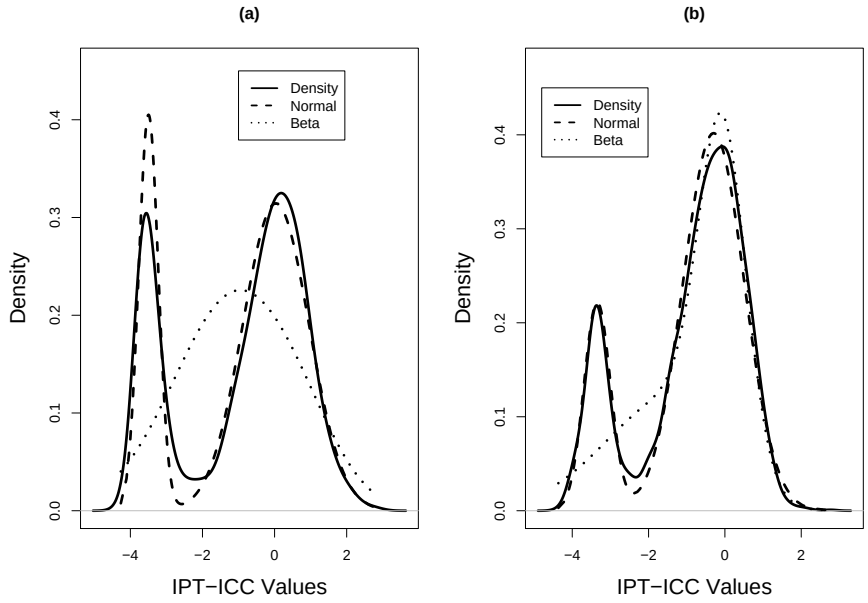


Figure 3: *Plot of the IPT-ICC values, fitted mixture of normal distribution, and pdf of transformed beta random variables for the (a) fecal and (b) mucosa data.*

(2) Transform $Y_1, \ldots, Y_n$ using the inverse-probit transformation and fit the transformed data with a two-component normal mixture.

Simulation 2: Data Generated from Normal-mixtures, Fit with Beta-mixtures

(1) Generate $X_1, \ldots, X_n$ from $\tilde{f}_N^f = 0.7\ N(0.04, 0.80) + 0.3\ N(-3.5, 0.07)$.

(2) Transform $X_1, \ldots, X_n$ using the probit transformation and fit the transformed data with a two-component beta mixture.

We repeat each simulation s=250 times for sample size n=1600 and use the EM algorithm to obtain estimates $\hat{\theta}_B$ and $\hat{\theta}_N$. The steps above are repeated for the mucosa dataset where the beta random variables are generated from $\tilde{f}_B^m = 0.8\ Beta(2.3, 2.3) + 0.2\ Beta(0.3, 1.3)$ and the normal random variables are gener-

ated from $\tilde{f}_N^m = 0.8\ N(-0.30, 0.60) + 0.2\ N(-3.3, 0.10)$.

We could not compare the outcomes of *Simulations 1* and *2* directly when the estimated parameters are for normal-mixtures and beta-mixtures, respectively. To ease the comparisons, we transform the resulting estimates in *Simulation 2* so that the outcomes correspond to means and variances of distributions that would give observations on the whole real line.

Consider the following approach:

1. *Retrieve* $(\hat{\alpha}_L, \hat{\beta}_L)$ *and* $(\hat{\alpha}_U, \hat{\beta}_U)$ *from the EM algorithm.*
2. *Generate 10,000 random variables from* $Z_L \sim Beta(\hat{\alpha}_L, \hat{\beta}_L)$ *and 10,000 random variables from* $Z_U \sim Beta(\hat{\alpha}, \hat{\beta})$.
3. *Transform* Z_L *and* Z_U *using the inverse-probit transform.*
4. *Calculate the mean and variance of the transformed random variables.*

Steps 1-4 are repeated for each simulation.

We present Monte Carlo statistics corresponding to the two components of the mixture distribution. Summary statistics for both simulation scenarios are presented in Table 1 and Table 2. We identify the target estimate of a scenario as "Truth" and report Monte Carlo estimates of mean, bias, standard deviation, and the square root of mean squared error (RMSE).

When comparing the true estimates to those obtained from the fitted distribution, we find that summary statistics from fitting transformed normal random variables with a beta mixture closely resemble the phenomenon observed when analyzing the fecal and mucosa data. Namely, it is the case that although the true proportions for the upper components of the fecal and mucosa data are 0.7 and 0.8, respectively, estimates of π_U resulting from the fit of a two-component beta distribution average

0.5. Moreover, the measure of bias in parameter estimation for fitting transformed normal random variables with a beta mixture is at least twice the bias in fitting transformed beta random variables with a normal mixture. This is true for almost every parameter estimation. These results lead us to believe that the two-component normal mixture is more robust to model mis-specification. Naturally, the normal mixture model fits normal data well. However, we also find that even if the data is truly beta distributed, we still retain significant accuracy if we alternatively transform the data into z-scores and fit with a normal distribution. On the other hand, assuming that the data is beta distributed could be costly if, in fact, it is not. We find that a biased conclusion could be reach if we model the two sets of ICC values using the mixture of betas.

Table 1: Monte Carlo mean, bias, standard deviation, and square-root MSE (RMSE) of estimates from simulation study "Data Generated from Beta-mixtures, Fit with Normal-mixtures."

		Data Generated from Beta-mixtures, Fit with Normal-mixtures				
Dataset		$\hat{\pi}_U$	$\hat{\mu}_U$	$\hat{\sigma}^2_U$	$\hat{\mu}_L$	$\hat{\sigma}^2_L$
Fecal	Truth	0.700	0.328	0.446	-1.771	3.330
	Mean	0.725	0.302	0.440	-1.951	3.321
	Bias	0.025	-0.026	-0.006	-0.180	-0.009
	Std Dev	0.018	0.023	0.028	0.152	0.283
	RMSE	0.031	0.035	0.029	0.235	0.283
Mucosa	Truth	0.800	-0.033	0.391	-2.090	2.722
	Mean	0.816	-0.049	0.398	-2.254	2.823
	Bias	0.016	-0.016	0.007	-0.164	0.101
	Std Dev	0.015	0.022	0.022	0.157	0.272
	RMSE	0.022	0.027	0.023	0.227	0.290

We further analyze the simulated outcomes and compare the sensitivity of each modeling approach toward distributional mis-specification through performing goodness of fit tests against assumed models. Analysis of goodness of fit (Section 2.2.5) test statistics resulting from the simulation study are given in Table 3. Precisely, for each simulated data set, we let the null hypothesis, H_0, be that the observed ICC (or

Table 2: Monte Carlo mean, bias, standard deviation, and square-root MSE (RMSE) of estimates from simulation study "Data Generated from Normal-mixtures, Fit with Beta-mixtures."

		Data Generated from Normal-mixtures, Fit with Beta-mixtures *(beta estimates valued on real line)*				
Dataset		$\hat{\pi}_U$	$\hat{\mu}_U$	$\hat{\sigma}^2_U$	$\hat{\mu}_L$	$\hat{\sigma}^2_L$
Fecal	Truth	0.700	0.328	0.446	-1.771	3.330
	Mean	0.453	0.282	0.521	-1.995	3.409
	Bias	-0.247	-0.046	0.075	-0.224	0.079
	Std Dev	0.010	0.036	0.032	0.050	0.138
	RMSE	0.247	0.059	0.082	0.229	0.159
Mucosa	Truth	0.800	-0.033	0.391	-2.090	2.722
	Mean	0.527	-0.149	0.387	-1.691	2.546
	Bias	-0.273	-0.116	-0.004	0.399	-0.176
	Std Dev	0.011	0.031	0.023	0.049	0.111
	RMSE	0.273	0.120	0.023	0.402	0.208

IPT-ICC) values are from the assumed model. We then compare the observed and the expected counts of observations within k bins, where $k = 5$, 8, 12, respectively, using Pearson's chi-square goodness of fit tests with significance level α=0.05 and $k-1$ degrees of freedom. For large values of the test statistic, namely $X^2 > \chi^2_{0.05,k-1}$, we reject the null hypothesis that the data comes from the assumed distribution.

Ideally, if the H_0 is true, there should be no more than 5% chance to reject the H_0 when α=0.05. Except for when $k = 5$, the proportion of tests that reject H_0 with normal-mixture modeling are all less than nominal level of 0.05. Further, in all cases, the outcomes obtained by normal-mixture modeling are comparable to those obtained when the true underlying distributions are assumed. The same does not hold for beta-mixture modeling. When the data are not generated according to the beta-mixture scheme, the goodness of fits tests are rejected close to or equal to 100% throughout. That is, the best fits of beta-mixtures still could not provide sufficiently close approximations that could pass the goodness of fit tests under *Simulation 1.*

Table 3: $P(X^2 > \chi^2_{0.05,k-1})$ for fecal (mucosa) data using 5, 8, and 12 bins.

		True	
	Fit	Beta	Normal
5	Beta	0.12 (0.08)	0.98 (0.01)
	Normal	0.13 (0.09)	0.36 (0.01)
8	Beta	0.00 (0.01)	1.00 (1.00)
	Normal	0.00 (0.01)	0.04 (0.02)
12	Beta	0.02 (0.01)	1.00 (1.00)
	Normal	0.02 (0.00)	0.03 (0.01)

3.5 ICC Comparisons of Fecal and Mucosa Data

Since our findings from the simulation study suggest that we use a two-component normal distribution to fit the probit transformed ICC values, we can accurately compare the fecal and mucosa array platform. In order to measure the quality of fecal array data, we first test for a distributional difference between ICC values from colon fecal and mucosa samples. With a p-value equating to 0, the likelihood ratio test (Section 2.2.6) rejects the null hypothesis. Thus, as expected, we deduce that the ICC values of genes obtained from the mucosa data are differentially expressed from those obtained from the fecal array data.

We further explore the extent of these distributional differences using bootstrapping for hypothesis testing. The first bootstrap analysis is designed to test for a difference in the proportion of irreproducible genes contained in each data set. Specifically, we test $H_a : \pi_F - \pi_M \neq 0$, where π_F and π_M are the proportion of irreproducible genes (genes with lower ICC values) in the fecal and mucosa data, respectively. Secondly, we determine whether there is a difference in the quality of information for reproducible genes. We test the hypothesis $H_a : \mu_F - \mu_M \neq 0$, where μ_F and μ_M are means of the upper mixture components for the fecal and mucosa data, respectively. Bootstrapped confidence intervals for the two respective tests are calculated to be (0.06,0.10) and

(0.27,0.40). As a result, we find that while the fecal array has a higher proportion of irreproducible genes, it averages ICC values for reproducible genes that are no worse than those obtained from the mucosa platform.

CHAPTER IV

VARIANCE ESTIMATION AND OUTLIER DETECTION METHODOLOGY

4.1 Introduction

Chapter IV is divided into two major components. First, we review variance estimation and introduce competing methodologies. Special focus is given to techniques which group genes in order to strengthen estimation. We discuss grouping G=50 genes to estimate variance. However, we also considered grouping 25 and 75 genes. While G=25 led to suboptimal power in detecting true outliers, there was no added benefit in G=75. Secondly, we give an overview of outlier detection algorithms.

4.2 An Overview of Variance Estimation

Microarray experiments are generally large in scale because a single array hybridization can generate thousands of data. However, since microarrays are costly and RNA samples are limited, replication of experiments is typically low in number. An issue of major concern in data analysis is the ability to estimate gene-specific variances from a small number of samples. Statistical tests such as the traditional t-test, which rely heavily on the sample variance, will have low power to detect differentially expressed genes if tests are carried out gene by gene. For example, a gene with small estimated variance could, by chance, have a large test statistic and be classified as differentially expressed even when the fold-change is small. This hinders the ability to draw reliable biological conclusions. Previous work (Arfin et al., 2000) suggested estimating a global variance by pooling information across all genes. If variances are homogeneous across genes, then this is a suitable approach. However, this assumption is likely to be untrue since variation of expression levels are known to vary from gene to gene.

Various methods have been devised to stabilize gene-specific variances by borrowing information across genes. Alternative variance-stabilization approaches, such as the statistical analysis of microarrays (SAM) t-test (Tusher et al., 2001), adjust gene-specific variances by adding a small constant to each variance estimate. The methodology proposed by Baldi and Long (2001) assumes a dependent relationship between mean and standard deviation in array data and models the statistics jointly using a conjugate normal-inverse gamma prior distribution. Each estimate of variance is a weighted contribution of gene-specific and global variation. Lonnstedt and Speed (2002) formulate the B-statistic by using an empirical Bayes approach and combining information across many genes. Kendziorski et al. (2003) also use Bayesian techniques in their consideration of a hierarchical gamma-gamma model.

Using the idea that there are strengths in numbers, various approaches seek to improve variance estimation by grouping genes according to intensity values and applying nonparametric smoothing techniques. Kamb and Ramaswami (2001) group genes by increasing average intensity and use regression to estimate gene-by-gene variance. Huang and Pan (2002) compare variance estimation obtained by (1) regression using equal weights, (2) loess regression giving less weight to more distant observations, and (3) nonparametric smoothing of the sample variance. Jain et al. (2003) propose local-pooled-error (LPE) estimation of within-gene expression error by pooling variance estimates for genes with similar expression intensities as other gene expressions under the same experimental condition. Lin et al. (2003) use smoothed medians and smoothed MAD's to estimate center and spread, respectively and construct standardized test statistics. Comander et al. (2004) consider a different intensity-variation relationship and pool together variance estimates of genes with similar minimum intensity in hopes of pooling together genes that are likely to have similar variances.

Several additional studies not only acknowledge, but also model the mean-variance

dependency that has been observed in some microarray data. Rocke and Durbin (2001) assert that the variance of raw spot intensities increases with their mean and model intensities with a two-component model. Durbin et al. (2002) also develop an error model to quantify variance as a function of mean intensity. Based on the assumption of a quadratic relationship between center and spread, Huber et al. (2002) propose variance stabilizing methodology in order to stabilize variance at low intensities.

Two recent methodologies by Cui et al. (2005) and Tong and Wang (2007) make no presumptions about a variance-intensity relationship for microarray data and estimate gene-specific variance components using shrinkage estimators. The method proposed by Cui et al. (2005) (referred to as the CHQBC estimator) presents estimates based on the James-Stein estimator (Lindley, 1962). Tong and Wang (2007) extend the shrinkage estimator methodology and suggest an optimal shrinkage parameter to replace the James-Stein shrinkage factor. Throughout, we refer to this technique as the TW estimator.

We propose gene-specific variance estimates based on shrinkage estimation of robust variance estimates. In the presence of outliers, performances of the CHQBC and TW methods deteriorate because each depend on the sample variance. To overcome this drawback, our approach replaces the sample variance by a robust variance estimator. In order to present a baseline comparison, we compare our proposed methodology to one which uses grouped estimation of variance determined by robust standard deviation of residuals (Motulsky and Brown, 2006). Both methods borrow information across genes in order to increase power. They are described in detail in sections 4.3.1 and 4.3.2.

It is our belief that grouping genes by similar intensities does not guarantee that data pooled together share similar variability. If this is true and grouped estimates

of variation are inaccurate representations of the observed variation for genes within that group, then any conclusions drawn will be misleading. Furthermore, we are likely to underestimate (overestimate) gene-specific variances for genes with high (low) variation. Our goal is to stabilize variance estimates in order to detect outliers, so that we can better quantify gene-specific variation in microarray data.

4.3 Variance Estimation Methodologies

We describe leading variance estimation techniques, and variants thereof, in the subsections to follow.

4.3.1 Grouped Estimation of the Robust Standard Deviation of the Residuals

Motulsky and Brown (2006) consider a robust non-linear regression setting and use residuals of the curve fitting to estimate variance. Since it is expected that 68.27% of the values in a Gaussian distribution fall within one standard deviation of the mean, Motulsky and Brown quantify variation in the residuals by calculating the 68.27 percentile of the absolute values of the residuals. The robust standard deviation of the residuals (RSDR) has a breakdown point of 32%.

Rather than use all the data in a large-scale microarray expression data set, we propose an alternative method based on local grouped estimation of the RSDR (grpRSDR). Under an ideal setup we should have that each gene expression is scattered around its median expression value. Thus, residuals are computed from subtracting a gene's median from its expression value. The steps for computing grpRSDR are:

1. Pool data by grouping 50 genes together and let I be the number of groups; I = G/50.

2. $RSDR_i$ = the 68.27th percentile of absolute residuals obtained from gene expression data in group i.

Thus, the grpRSDR estimation of variance for genes g in group i is given by $RSDR_i$.

4.3.2 Tong & Wang's Optimal Shrinkage Estimation of Variance

In order to estimate variance, we adopt Tong and Wang's procedure for optimal shrinkage variation estimation (Tong and Wang, 2007). They propose a shrinkage estimator for gene-specific variance components which borrows information across variances.

4.3.2.1 Gene-specific Variance Estimation

Let X_g be the residual sum of squared errors (SSE) and σ_g^2 be the true variance of gene g. It is assumed that X_g/σ_g^2 are independent and Chi-square distributed with ν degrees of freedom, for $g = 1, \ldots, G$ genes. Thus,

$$X_g \sim \sigma_g^2 \chi_\nu^2.$$

After natural logarithmic transformation on X_g, the above expression is equivalent to the following location model:

$$X_g' = ln\ \sigma_g^2 + \epsilon_g', \tag{4.1}$$

where $X_g' = ln(X_g/\nu) - m$, $\epsilon_g' = ln(\chi_\nu^2/\nu) - m$, and $m = E(ln(\chi_\nu^2/\nu))$.

Cui et al. (2005) extend Stein's theory for estimation of multiple means to multiple variances and use the James-Stein shrinkage estimator to shrink the variance component of each gene towards the bias corrected geometric mean of variances. The James-Stein shrinkage estimator of $ln\ \sigma_g^2$ is given by

$$\bar{X}' + \left(1 - \frac{(G-3)V}{\sum(X_g' - \bar{X}')^2}\right)_+ (X_g' - \bar{X}'), \tag{4.2}$$

with shrinkage factor $\left(1 - (G-3)V/\sum(X_g' - \bar{X}')^2\right)_+$. The estimator is proven to have uniformly smaller mean square error than the maximum likelihood estimator. It

also requires no assumptions about the distribution of variances across genes. However, the authors do note that the sampling distribution of the logarithm of variance estimates is assumed to be normal.

The CHQBC estimate of σ_g^2 that Cui et al. (2005) propose emerges upon transforming (4.2) back to the original scale. We have that,

$$\tilde{\sigma}_g^2 = (\prod_{g=1}^{G}(X_g/\nu)^{1/G})\ B \times exp\left[\left(1 - \frac{(G-3)*V}{\sum(ln\ X_g - \overline{ln\ X_g})^2}\right)_+ \times (ln\ X_g - \overline{ln\ X_g})\right], \tag{4.3}$$

where $V = var(\epsilon_g')$, $\overline{ln\ X_g} = \sum_{g=1}^{G} ln(X_g)/G$ and $B = exp(-m)$. If we let $Z_g = X_g/\nu$, $Z_{pool} = \prod_{g=1}^{G} Z_g^{1/G}$, and $\hat{\alpha}_0 = 1 - (1 - (G-3)V/\sum(ln\ X_g - \overline{ln\ X_g})^2)_+$. With $\alpha = \hat{\alpha}_0$, the CHQBC estimator may be written as

$$\tilde{\sigma}_g^2(\alpha) = B(Z_{pool})^{\alpha}(Z_g)^{1-\alpha}. \tag{4.4}$$

Tong and Wang revise (4.4) by combining two unbiased estimators of σ_g^2. If variance homogeneity holds, then $\sigma_g^2 = \sigma^2$ for all g, $E(Z_{pool}) = \sigma^2/B$, and BZ_{pool} is an unbiased estimator of σ_g^2. It is also the case that Z_g is an unbiased estimate of σ_g^2. Hence, they present the following modification

$$\hat{\sigma}_g^2(\alpha) = (BZ_{pool})^{\alpha}(Z_g)^{1-\alpha}, \qquad 0 \leq \alpha \leq 1. \tag{4.5}$$

The estimator $\hat{\sigma}_g^2$ is referred to as the TW estimator.

4.3.2.2 Estimation of the Shrinkage Parameter

In lieu of the shrinkage factor given in (4.2), Tong and Wang (2007) derive optimal estimation of the shrinkage estimator α. We implement the authors' adaptation of optimization under the Stein loss function. The Stein Loss function,

$$L(\sigma^2, \hat{\sigma}^2) = \hat{\sigma}^2/\sigma^2 - ln(\hat{\sigma}^2/\sigma^2) - 1, \tag{4.6}$$

converges to infinity as $\hat{\sigma}^2$ approaches zero and as $\hat{\sigma}^2$ approaches infinity. Thus, gross overestimation and underestimation of the true variance are equally penalized.

The authors present a family of shrinkage estimators for $(\sigma_g^2)^t$ given by

$$\hat{\sigma}_g^{2t}(\alpha) = (h_G(t)Z_{pool}^t)^\alpha \ (h_1(t)Z_g^t)^{1-\alpha}, \qquad 0 \leq \alpha \leq 1, \tag{4.7}$$

where

$$h_n(t) = (\frac{\nu}{2})^t \left(\frac{\Gamma(\frac{\nu}{2})}{\Gamma(\frac{\nu}{2} + \frac{t}{n})} \right)^n \tag{4.8}$$

and $\Gamma(\cdot)$ is the gamma function. It is the case that when $t = 1$ and G is large, (4.7) reduces to (4.5).

The optimal α under Stein loss function minimizes the average risk for each gene, which is given by

$$\begin{aligned} R(\sigma^{2t}, \hat{\sigma}^{2t}) &= \frac{1}{G} \sum_{g=1}^{G} E(L(\sigma_g^{2t}, \hat{\sigma}_g^{2t})) \\ &= \frac{h_G^\alpha(t) h_1^{1-\alpha}(t)}{h_1^{G-1}(\frac{\alpha t}{G}) h_1((1 - \alpha + \frac{\alpha}{G})t)} (\sigma_{pool}^2)^{\alpha t} \frac{1}{G} \sum_{g=1}^{G} (\sigma_g^2)^{-\alpha t} \\ &\quad -ln \ (h_G^\alpha(t) h_1^{1-\alpha}(t)) - t\Psi(\frac{\nu}{2}) + tln \ (\frac{\nu}{2}) - 1, \end{aligned} \tag{4.9}$$

where $t > \nu/2$, $\Psi(t) = \Gamma'(t)/\Gamma(t)$ is the digamma function. The optimal estimator is denoted as $\hat{\sigma}_{Z,g}^2(\alpha_1^*)$, where $\alpha_1^* = \underset{\alpha \varepsilon [0,1]}{\operatorname{argmin}} \ R(\sigma^{2t}, \hat{\sigma}^{2t})$.

4.3.2.3 Some Important Algorithm Details

In order to obtain an estimate of the optimal shrinkage parameter, it is necessary to assume that $Z_g \rightarrow \sigma_g^2$ *a.s.* Let $b(\boldsymbol{\sigma}^2) = (\sigma_{pool}^2)^{\alpha t} \frac{1}{G} \sum_{g=1}^{G} (\sigma_g^2)^{-\alpha t}$. Then we can estimate $b(\boldsymbol{\sigma}^2)$ with $b(\boldsymbol{Z})$ in (4.9). For small ν, we also find it necessary to implement an alternative two-step procedure:

1. Estimate $b(\boldsymbol{\sigma}^2)$ with $b(\boldsymbol{Z})$ in (4.9) and compute a temporary optimal shrinkage parameter and resulting temporary optimal shrinkage estimators, $\hat{\sigma}_*^2$.

2. Substitute $b(\hat{\sigma}_*^2)$ for $b(\boldsymbol{\sigma}^2)$ in (4.9) in order to find the final optimal shrinkage parameter and estimators.

As the authors suggest, we truncate the smallest 1% of Z_g's in the procedure so that estimation of α remains stable. We use the built-in optimization code 'nlminb' within the R computing environment to estimate α.

4.3.3 Proposed New Methodology: TW(mix/mad)

When outliers are present in the data, the gene-specific estimator of variance that Tong and Wang propose is prone to inaccurate estimation. The estimator relies heavily on the sample variance, which overestimates variance when the data is contaminated by outliers. Alternatively, we propose to replace Z_g in (4.7) with a vector of robust variance estimates. The square of the median absolute deviation (MAD) would be a natural consideration for robust estimation of variance, but this statistic alone is insufficient. MAD has a tendency to underestimate standard deviation even when no outliers are present and could potentially create a problem with high counts of false positives. Our alternative uses information in the MAD and sample standard deviation to find the best variance estimate for a given gene.

To estimate the likelihood of an outlier in each gene's expression data, we assess the relative change in standard deviation between an estimate robust to outliers and one influenced by outliers. Let

$$r_g = \frac{S_g - MAD_g}{MAD_g}, \tag{4.10}$$

where S_g and MAD_g are the sample standard deviation and MAD, respectively, of gene g. The MAD is defined to be

$$MAD = k \times median\{x_i - median\{x_i\}\}, \tag{4.11}$$

where $k \approx 1.4826$ for normally distributed data, unless otherwise stated. We will observe small values of r_g when there is little deviation between MAD and sample standard deviation, which suggests that the gene data may be free of outliers. On the other hand, we expect to observe large values of r_g for genes with significant outlying expressions.

We consider shrinking the following vector of variance estimates:

$$V_g = \begin{cases} \frac{1}{2}(S_g^2 + MAD_g^2) & \text{if } r_g <= R \\ MAD_g^2 & \text{if } r_g > R. \end{cases} \tag{4.12}$$

In order to specify the cutoff for the piecewise function, and also to justify why we choose to send V_g into Tong and Wang's optimal shrinkage algorithm, we use the following illustration: Simulate n=6 random observations from a $N(0,1)$ distribution for $G = 10,000$ genes. Possible values of R were determined by percentiles of r_g ratios obtained from the simulated data. We ultimately choose R=3.6, which was the approximated 99^{th} percentile of r_g ratios. For all genes g with r_g values exceeding 3.6, we avoid any sensitivity to outlier observations since the relative change in variation is large. Instead, we quantify variation solely using the square of the MAD statistic. On the other hand, for all genes g with r_g values at most 3.6, we send into the algorithm the average of sample variance and the square of the MAD statistic.

Ideally, we want to shrink estimates of variance which are not already believed to be distorted. In Figure 4, we plot the distribution of deviance from the true variance when using the square of the MAD statistic, the sample variance, and also, the average of the two to estimate variance. When there are no outliers in the data, the sample variance is most centered at zero deviation. We retain most of the same accuracy even when using the average of the sample variance and the square of the MAD statistic. However, the square of the MAD statistic is shown to consistently underestimate the true variance. For this reason, we choose not to rely solely on

the square of the MAD statistic in computing optimal shrinkage variance estimators. Certainly the sample variance will overestimate variance in the presence of outliers; however, the MAD statistic underestimates variance even when the data is free of outliers. By using the average of the two, we are able to protect against extreme overestimation and underestimation of the true variance. Thus, we derive V_g as the population of variance estimates we wish to shrink.

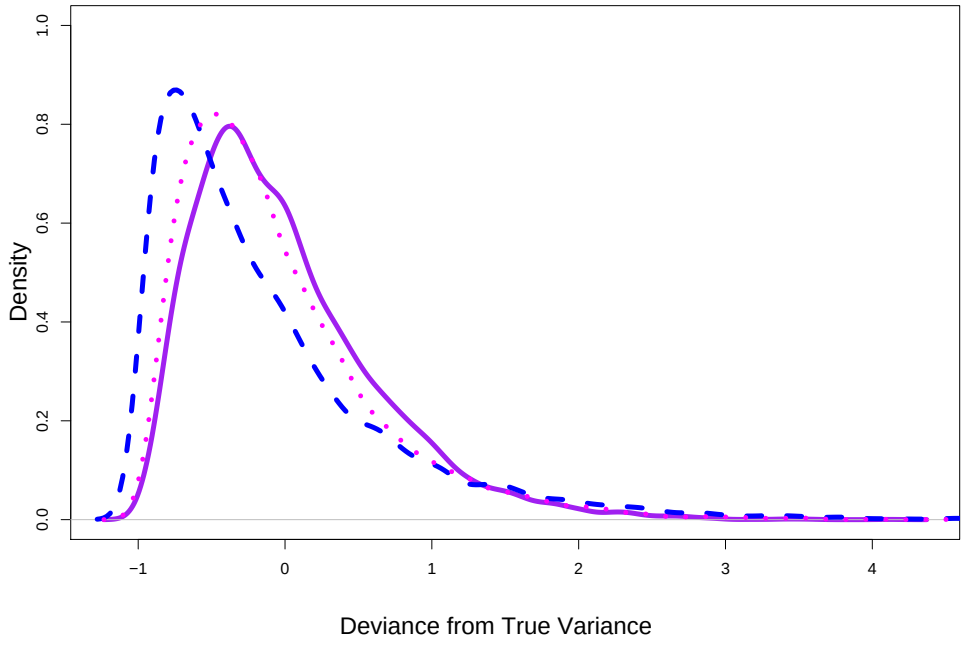


Figure 4: *Density plots of the difference between estimated variances and the true variance for simulated N(0,1) gene expression data. Variance is estimated using three statistics - (1) square of the MAD statistic (dashed line), (2) sample variance (solid line), and (3) the average of sample variance and the square of the MAD statistic (dotted line).*

The optimal shrinkage estimators based on our proposed methodology arise naturally after substituting $b(\boldsymbol{V})$ for $b(\boldsymbol{Z})$ in estimating $b(\boldsymbol{\sigma}^2)$ (section 4.3.2.3). We denote these optimal shrinkage estimators as $\hat{\sigma}^2_{V,g}$, which we reference as the TW(mix/mad)

estimate.

4.3.3.1 Grouped Estimation of TW(mix/mad) Methodolgy

As an extension of the previous section, we suggest further stabilizing variance estimation by grouping together residuals that are standardized with robust optimal shrinkage variance estimators, $\hat{\sigma}^2_{V,g}$ (Section 4.3.3). The estimate of common variance measures grouped estimation of the robust standard deviation of the standardized residuals (RSDSR). Because standardized residuals follow an approximately standard normal distribution, we are now able to pool data from the same distribution. We refer to this approach as grpTW(mix/mad). It is detailed as follows:

1. Pool data by grouping 50 genes together and let I be the number of groups; I = G/50.

2. Compute $RSDSR_i = 68.27^{th}$ percentile of absolute standardized residuals from expression data in group i.

Thus, the grpTW(mix/mad) estimation of variance for genes g in group i is given by

$$\tilde{\sigma}^2_{V,g} = \hat{\sigma}^2_{V,g} * RSDSR^2_i, \tag{4.13}$$

where $\hat{\sigma}^2_{V,g}$ is the optimal shrinkage estimator of gene g using the TW(mix/mad) methodology.

4.3.4 Grouped Estimation of Variance as a Function of Mean

Modeling variance as a quadratic function of mean is a common approach in modeling gene expression data. Strimmer (2003), Rocke and Durbin (2001), and Huber et al. (2002) (and references therein) each suggest a quadratic variance-mean dependency in microarray data. We follow this trend in proposition of grouped estimation of variance as a function of mean. The data fitting routine for each treatment requires

the robust fitting of linear models. We use the 'rlm' function contained in the $MASS$ package of R. In 'rlm', fitting is done by iterated re-weighted least squares using M-estimation. Initial values for the coefficients are found using "lts," an unweighted least-trimmed squares fit with 200 samples.

We estimate the variance of Y_{gj} (observation j of gene g) with its squared residual and substitute the mean for median, a robust measure of center. Let X_{gj} be the median of gene g for observation j. Consequently, the estimated gene-specific variances $\hat{Y}_{gj}$ are the fitted values from robust quadratic regression of Y_{gj} on X_{gj}^2. We use the square root of $\hat{Y}_{gj}$ to standardize residuals and estimate RSDSR by:

1. Pool data by grouping 50 genes together and let I be the number of groups; I = G/50.

2. Compute $RSDSR_i = 68.27^{th}$ percentile of absolute standardized residuals from expression data in group i.

Thus, the grpVM estimation of variance for genes g in group i is given by $\hat{Y}_{gj} * RSDSR_i^2$.

4.4 An Overview of Outlier Detection

Outliers are common in microarray data and generally arise from either biological variation or measurement error from systematic sources of variability. While both anomalies present observations that are inconsistent with the remainder of the data (Barnett and Lewis, 1984), we are solely interested in identifying the latter since any variation resulting from true biological differences is important for analysis. The high-dimensionality of microarray datasets coupled with low gene-by-gene repetition makes outlier identification considerably challenging. The process of detecting outliers is essentially automated since visual inspection of such large data sets is impractical.

The "3σ rule" is often used to detect outliers in data of any nature. This is true for microarray data as well. Under the assumption that the data is independent, i.i.d., and normally distributed, the probability that an observation lies more than 3 standard deviations away from the mean is 0.3%. Extreme points lying within the tails of this distribution are considered to be outliers. We determine this by calculating a z-score for every observation,

$$z_{gj} = \frac{X_{gj} - \bar{X}_g}{S_g}, \tag{4.14}$$

where X_{gj} is expression data for gene g of array j, and $\bar{X}_g$ and S_g are the sample mean and standard deviation, respectively, of gene g. Any observation with an absolute z-score that exceeds three is classified as an outlier. There are other variations of the "3σ rule" in which data producing absolute z-scores above a certain threshold are flagged as outliers. Moffitt et al. (2005) adopt this rule in order to detect outliers on a gene-by-gene basis. Prolla (2002) and Du et al. (2005) also rely on standard z-scores to detect anomalous gene expressions.

However, if outliers are indeed present, then measures of the sample mean and sample standard deviation are highly influential and it is likely to experience both masking and swamping. As an alternative, the *Hampel identifier* (Davies and Gather, 1993) substitutes the mean and standard deviation for outlier resistant statistics - the median and median absolute deviation (MAD), respectively in (4.14). The median and MAD are robust estimates of center and spread, respectively. Yang et al. (2006) use resistant z-scores to identify outliers and then classify outlier arrays based on the percentage of detected outliers. The *Hampel identifier* presents an outlier resistant alternative to standard z-scores especially when the sample size is small. However, it may cease to be useful when the distribution of data is not symmetric.

Departing from z-scores altogether, Mariani et al. (2003) compute fold-changes

for pairwise comparisons of gene expressions within a treatment and classify outlier genes based on the number of fold-changes that exceed a 2 point cutoff. Li and Wong (2001) use a statistical model to identify array, probe, and single Perfect Match (PM) - Mismatch (MM) outliers through an iterative, sequential process. A drawback to this approach is that a large number of arrays is required to obtain accurate standard error estimates for parameters in the model.

We consider the process of detecting outliers in replicated microarray data similar to that of detecting multiple outliers in regression. The "pseudo" regression curve is constructed by the median values of genes within a treatment and we are interested in identifying points that deviate far from the curve. Motulsky and Brown (2006) propose a method for identifying outliers when fitting data with nonlinear regression that mimics the false discovery rate (FDR) approach for multiple comparisons. We adopt their outlier elimination method and apply it holistically to all the arrays within the same treatment group.

By analyzing arrays from different treatments, we are able to retain only those gene expressions that are consistent with other measurements for the same gene under the same experimental setup. We believe that any striking variation due mainly to noise is corrected in the normalization step of microarray analysis. Furthermore, gene expressions under the same treatment are believed to be consistent enough across arrays so that when analyzed for systematic outliers, the biological variation is preserved and only those expressions which do not provide valuable information will be classified as outliers.

4.5 Outlier Detection Methodology

Motulsky and Brown (2006) outline an algorithm for detecting outliers based on the false discovery rate (FDR) approach to multiple hypothesis testing (Benjamini and

Hochberg, 1995). They liken the problem of determining when a point is far enough from the curve to be considered an outlier similar to the problem of determining when a p-value is small enough to be labeled statistically significant. In this way, we are able to incorporate the empirical distribution of p-values in deciding a decision threshold.

When making multiple comparisons, correction techniques like the Bonferroni adjustment, which sets a significance cutoff equal to the family-wide error rate divided by the number of comparisons, can be too conservative when analyzing a large number of tests. Benjamini and Hochberg developed the FDR procedure to control the expected proportion of false positives among all tests declared significant. This test is most prevalent in microarray analysis when detecting differentially expressed genes. It mimics a data reduction problem in that we are able to narrow a search from thousands of genes to a reduced set. Since we are more concerned with making certain the reduced data set contains all possible differentially expressed genes, we are willing to accept a set which includes some false positives.

4.5.1 Choosing a Value of Q

The number of detected outliers, and consequently the proportion of false positives, is controlled by specifying a value Q for the FDR. Choosing an FDR of 5% means that 5% of the points we detect as outliers are actually false positives. Decreasing (increasing) the value of Q will simultaneously decrease (increase) the number of false positives and decrease (increase) the number of true positives. Thus, any criterion specification offers a tradeoff between the number of false positive and false negative errors.

Motulsky and Brown consider values of Q equal to 0.1%, 1%, and 10%. While Q=10% is too aggressive, Q=0.1% is too conservative. They make a subjective deci-

sion to choose Q=1%, which we follow for our analysis.

4.5.2 Algorithm

Let $\bar{\sigma}_g^2$ be the estimate of variance for gene g as determined by any particular methodology *e.g.* grpRSDR, grpTW(mix/mad), grpVM etc. The standardized residual for gene expression g is given by

$$t_{igj} = \frac{y^*_{igj} - y^{med}_g}{\bar{\sigma}_g}, \tag{4.15}$$

where y^*_{igk} is the normalized gene expression of rat i, array j(i), and gene g and y^{med}_g is the median expression of gene g. We can now compute a p-value for each test statistic testing the null hypothesis that the statistic follows a t-distribution. The steps for outlier detection are as follows:

(i) For each test statistic, t_{igj}, in group k of grouped gene expression data:

1. Compute p_{igj}, the two-tailed p-value from a $t_{df=N_k-2}$, where N_k is the number of observations in group k

(ii) For all p_{igj}:

1. Order all p_{igj} from lowest to highest and let $p_{[t]}$ denote the t-th order-statistic.
2. Find the largest t for which $p_{[t]} < (Q * t) \,/\, \sum_j N_j$, where N_j is the number of observations in array j and Q is pre-specified, e.g. 1%.
3. Classify $p_{[1]}, \ldots, p_{[t]}$ as outliers.

We leave it to the researcher of subject-level expertise to decide whether to delete the anomalous data or further analyze in order to identify the false positives.

CHAPTER V

ANALYSIS OF OUTLIER DETECTION FOR SIMULATED MICROARRAY DATA

5.1 Introduction

This chapter presents analysis of simulated microarray data. In Sections 5.2 and 5.3 we discuss two simulation scenarios and present analytical results. For each simulation study, we perturb the data with outliers and evaluate the performances of each method primarily based on positive predictive values and false negative rates. We initially consider ten different methodologies for estimating variance, but currently narrow our focus to a comparison of grpRSDR, grpTW(mix/mad), and grpVM.

5.2 Simulation I: Independent Gene Variance-Intensity Relationship

In this section we conduct a study based on a simulation setup described in Tong and Wang (2007) which assumes a completely independent relationship between a gene's location center and spread. We consider several methods of estimating gene-specific variance, which include approaches documented in existing literature. The methods are evaluated for data simulated with and without outliers. First, we simulate expression data for G=5000 genes by generating σ_g^2 , $g = 1, \ldots, G$ from a $U(0.05, 0.30)$ distribution. For each σ_g^2, we simulate $n = 6$ observations from $N(\mu_g, \sigma_g^2)$, where each μ_g is a random sample from $N(0, 1)$. We repeat simulation $s = 200$ times.

Second, we perturb each simulated data set with 2500 outliers and measure the positive predictive value (PPV) (Altman and Bland, 1994) and false negative rate (FNR) (Fleiss, 1981) of each methodology. PPV and FNR measure the proportion of correct-detection among all detected outliers and the proportion of false negatives

among true outliers, respectively. Precisely,

$$PPV = \frac{true\ positives}{true\ positives + false\ positives} \tag{5.1}$$

and

$$FNR = \frac{false\ negatives}{true\ positives + false\ negatives} \tag{5.2}$$

In real data, outlier identities and features are random and unknown. Hence, we contaminate data by moving observations 3 and 10 standard deviations away. Specifically, we introduce outliers by considering a subset of every other gene. For each of these genes, we find the observation, $y_{g,i}$, farthest from the median in absolute value and replace that gene expression with

$$y_{g,i} = \begin{cases} y_{g,i} - c\sigma_g, & \text{if } y_{g,i} - y_g^{med} < 0 \\ y_{g,i} + c\sigma_g, & \text{if } y_{g,i} - y_g^{med} > 0, \end{cases}$$

where c=3 for all subsetted g or c=10.

5.2.1 Methodologies

We consider ten different methodologies for estimating variance. They are listed in Table 4.

As an overview, the prefix 'grp' indicates that an estimate of robust standard deviation of the residuals is obtained by grouping 50 genes in sequential order. The '-OBM' suffix indicates that in lieu of sequential order, we group genes by increasing median intensity. For optimal shrinkage estimation the statistic inside the parentheses represents what we send into the algorithm detailed by Tong and Wang (TW).

The design of simulation studies presented in the following subsections assist in investigating properties of the different variance estimators. Our goals are to examine the effects of grouping, examine the effectiveness of gene-specific variances, and to

Table 4: Descriptive procedures for ten methodologies of estimating gene-specific variance.

Methodology	Description
MAD	divide each residual by the the estimate of MAD for that gene
grpMAD	standardize each expression using MAD; estimate RSDSR
grpSTD	standardize each expression using sample standard deviation;
—	estimate RSDSR
TW(var)	divide each residual by the TW shrinkage estimators
grpTW(var)	standardize each expression using the TW shrinkage estimators;
—	estimate RSDSR
grpTW(mad)	standardize each expression by sending MAD^2 into the
—	TW algorithm; estimate RSDSR
grpTW(mad50/std50)	standardize each expression by sending $\frac{1}{2}(S^2 + MAD^2)$ into the
—	TW algorithm; estimate RSDSR
grpTW(mix/mad)	as described in section 4.3.3.1
grpRSDR-OBM	as described in section 4.3.1
grpVM-OBM	as described in section 4.3.4

determine the strengths and weaknesses of estimators when we assume no functional relationship between gene variance and intensity. By perturbing the simulated data with outliers we are better able to characterize each of these properties.

5.2.2 Simulated Data: No Outliers

With no outliers we direct our attention to quantifying the average number of false positives, see Table 5. Results seen here offer a baseline of what to expect from these variance estimators once outliers are added. With large data sets, we expect to observe some false positives even when there are no outliers. We observed in Figure 4 that the sample standard deviation and the average of the square of the MAD and sample standard deviation accurately estimate variance for non-outlier data. We find that methodologies which depend on these statistics are extremely conservative and fail to misclassify any observations. We also observe that grpMAD and grpTW(mad), which group residuals that are standardized with MAD estimates, are several times more liberal than the other methods. This supports our findings of negative deviance when estimating variance with MAD. The effects of underestimation are demonstrated with

Table 5: The average number of detected outliers for simulated data with no outliers.

Methodology	# of FP
MAD	10
grpMAD	699
grpSTD	0
TW(var)	0
grpTW(var)	0
grpTW(mad)	413
grpTW(mad50/std50)	0
grpTW(mix/mad)	54
grpRSDR-OBM	55
grpVM-OBM	54

an increase in the detection of false positives.

5.2.3 Simulated Data: With Outliers

With non-outlier simulated data we were able to generalize tendencies in the methodologies. Additionally, we perturb the data with outliers so that we are better able to assess the effects of grouping and draw additional conclusions about the variance estimators themselves. Table 6 shows the positive predictive values and false negative rates of simulated data with outliers for each methodology listed in Table 4.

grpTW(mix/mad) and grpRSDR-OBM are most robust to outlier distance and do well in performance measures i.e. high predictive probability and low false negative rate. In general, we see a vast improvement in performance when we allow the variance term to be supplemented with local grouped estimation of a common variance. For example, grpMAD is much more stable in outlier detection than MAD. We find that using MAD to estimate gene-specific performs great when outliers are known to be considerably far from non-outlier data (i.e. c=10). However, this limits analysis to an unreasonable constraint. Although grpMAD has smaller positive predictive value, its stability over outlier distance provides an advantage in real data

Table 6: Positive Predictive Value (PPV) and False Negative Rate (FNR) of outlier detection methodologies for simulated data perturbed by outliers.

	Outliers (c=3)		Outliers (c=10)	
Methodology	*PPV*	*FNR*	*PPV*	*FNR*
MAD	0.923	0.862	0.966	0.033
grpMAD	0.781	0.075	0.792	0
grpSTD	0	1	0.886	0.420
TW(var)	0	1	0	1
grpTW(var)	0.978	0.177	0.897	0.017
grpTW(mad)	0.870	0.023	0.872	0
grpTW(mad50/std50)	0.958	0.047	0.822	0
grpTW(mix/mad)	0.922	0.058	0.946	0
grpRSDR-OBM	0.943	0.088	0.946	0
grpVM-OBM	0	0.166	0	0.166

analysis. Grouped estimation of the TW estimator, grpTW(var), is less powerful than grpTW(mix/mad) and grpRSDR-OBM because it relies on the sample variance, which is unreliable in the presence of outliers. It is within reason to anticipate an increase in performance when outliers are forced to be substantially far from the remainder of the data. However, grpTW(var) decreases in positive predictive probability as outliers are moved from 3 to 10 standard deviations away. The effects of grouping are further demonstrated in the inability of the ungrouped TW estimator to detect any outliers. Recall that grpVM-OBM methodology is constructed under the assumption that variance increases as a quadratic function of mean. Since gene spread and center are simulated completely independent of one another, we see the repercussions - zero PPV and relatively high FNR - in using parametric models of variance-intensity to fit data which is nonparametric in structure.

Equally important to reports of PPV and FNR is understanding where the false positives (FP) and false negatives (FN) occur. This is especially of interest in comparing grpRSDR-OBM and grpTW(mix/mad). Both methodologies consistently perform well; however, there is a tradeoff in PPV and FNR for c=3. While grpRSDR-OBM

has higher positive predictive probability, grpTW(mix/mad) has lower false negative rate. For one simulated data set, we record counts of misclassified observations based on properties of the data. Specifically, we subset data using a 5×5 MAD versus median grid system, where MAD and median values are divided into 5 equally spaced intervals. Data in Figure 5 summarizes the counts of false negatives for this grid system. We find that grpRSDR-OBM is extremely conservative and shows difficulty in detecting outliers in the lowest variable region. Throughout the range of median intensities for lowest variability, it fails to detect 215 observations compared to the 68 that grpTW(mix/mad) fails to detect.

Data in Figure 6 reports the counts of false positives over the same grid system for the same simulated data set. We acknowledge that grpTW(mix/mad) is more liberal with classifying outliers in the lowest variable region. However, we find this to be less alarming since one is afforded the opportunity to examine a smaller subset of genes in order to verify outlier classification. We have developed a diagnostic tool which allows for further investigation of these observations which we discuss in Chapter VI. Figure 6 also demonstrates that grpRSDR-OBM continues to falsely classify non-outlier observations as the measure of spread increases. This is particularly alarming since the ability to distinguish between outliers and non-outliers naturally diminishes as the variability increases.

We gather from Figures 5 and 6 that grpRSDR-OBM suffers from classifying false negatives for low variable genes and false positives for high variable genes. There is also evidence to suggest that grpTW(mix/mad) mistakenly classifies non-outliers as outliers for low variable genes. We extend the investigation of misclassified observations to the entire simulation study and report the average number of false positives and false negatives based on quantiles of the true variance, see Figure 7. We divide true variances into 10 equal frequency bins and report the average FN and FP counts

over all simulations. Since FP and FN averages for grpTW(mix/mad) are near equal across variance quantiles, it is impossible to draw meaningful conclusions. On the other hand, it is clear that grpRSDR-OBM struggles to correctly identify outliers for low variable genes and incorrectly detects twice as many outliers for the most variable genes. Moreover, the high false negative counts seen with grpRSDR-OBM indicate that a significant number of true outliers are undetected and misclassified as non-outlier data. This drawback is certain to inflate gene-specific variance estimates and distort any biological conclusions.

5.2.4 Additional Analysis

We also consider perturbing simulated data with a total 500 and 250 outliers by adding outliers to every 10^{th} and 20^{th} gene, respectively. As the number of true outliers decreases, we observe a decrease in PPV and for c=3, we generally see an increase in FNR. Overall, we observe similar relative relationships between methodologies. It is still the case that grpRSDR-OBM and grpTW(mix/mad) appear to be competing methodologies. However, the additional studies with fewer outliers suggest that grouped estimation of TW(mad50/std50) may be an acceptable alternative when there are fewer outliers, especially when outliers are more distinct and further from non-outlier data. For c=3, our methodologies suffer in false negative rate. We take a closer look at the methodologies for our second simulation scenario and for real data analysis. Complete tables of PPV and FNR reports for the ten methodologies are presented in Appendix VII for each additional simulation.

5.3 Simulation II: Gene Variance-Intensity Dependency

We include an additional simulation study to reflect a mean-variance relationship in expression data. Based on real data structure, we model variance as a quadratic

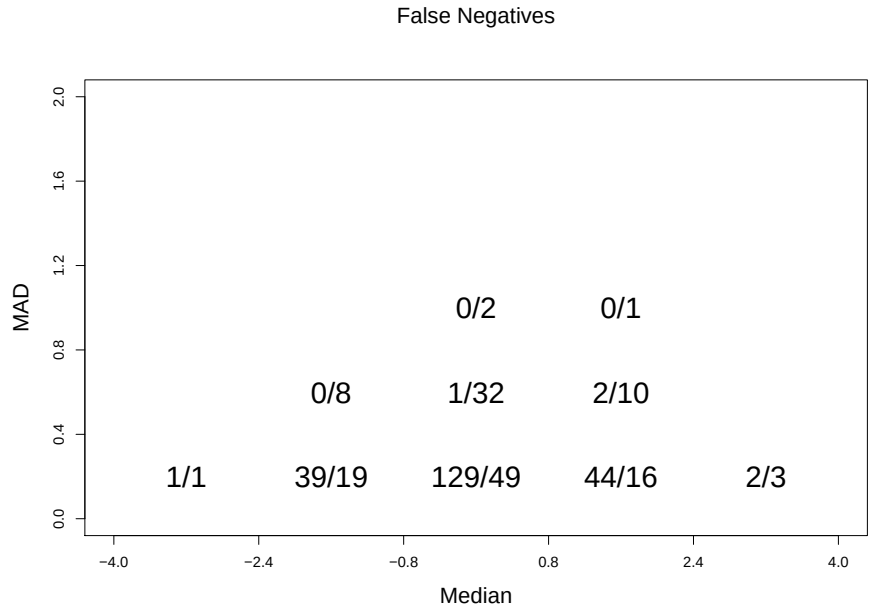


Figure 5: *The number of false negatives classified by two methods using one gene expression data set simulated to have no relationship between gene center and spread. Counts are reported using a* 5×5 *median vs MAD grid system. The results for grpTW(mix/mad)/grpRSDR-OBM are shown to the left/right of the slash, respectively.*

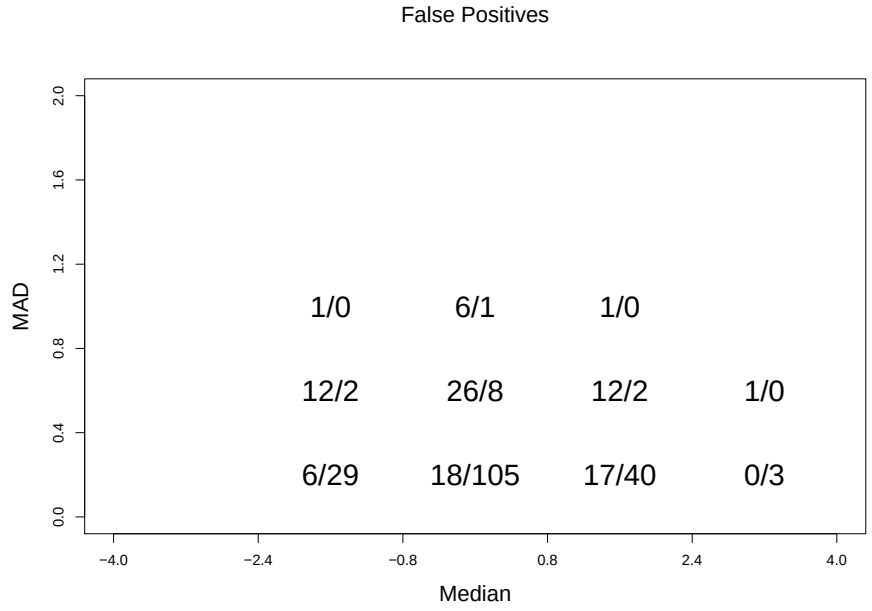


Figure 6: *The number of false positives classified by two methods using one gene expression data set simulated to have no relationship between gene center and spread. Counts are reported using a* 5×5 *median vs MAD grid system. The results for grpTW(mix/mad)/grpRSDR-OBM are shown to the left/right of the slash, respectively.*

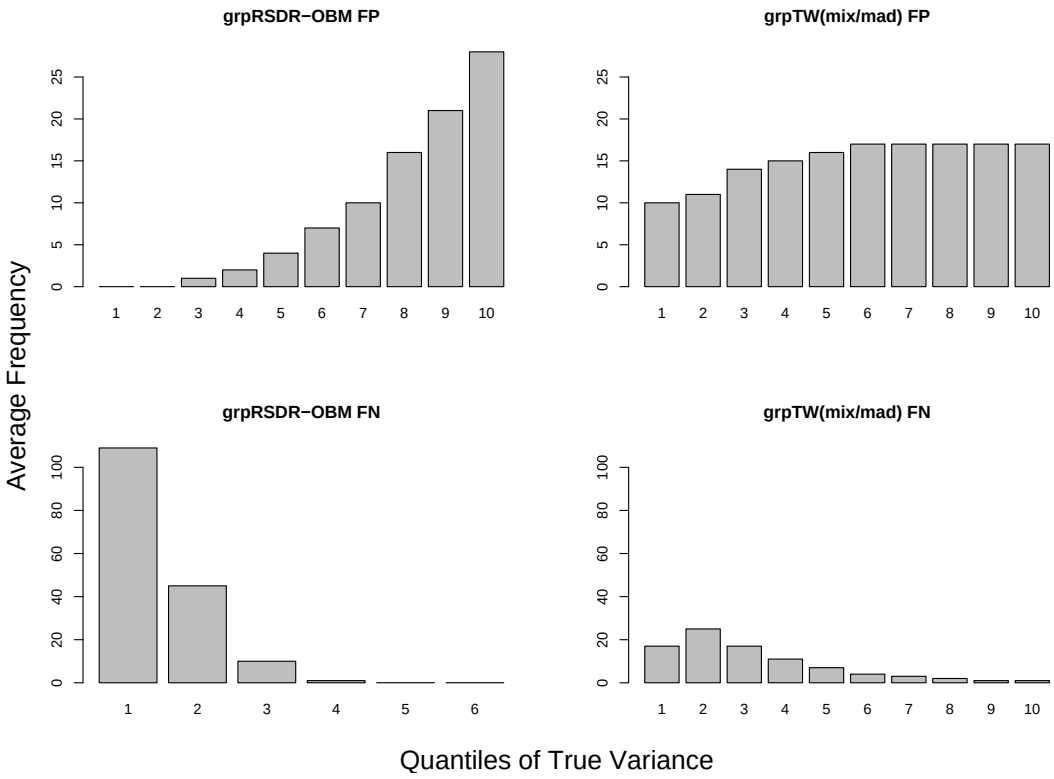


Figure 7: *A report of false negatives and false positives for gene expression data simulated to have no relationship between gene center and spread. The average number of false positives (upper panel) and false negatives (lower panel) over all simulations are shown for 10 equally spaced quantiles of true variance estimates. Results are provided for two methodologies - grpRSDR-OBM (left) and grpTW(mix/mad).*

function of mean using the following piecewise function

$$Y_g = \begin{cases} 0.054 & \text{if } X_g < 6, \\ 0.09 - 0.004(X_g - 9)^2 & 6 \leq X_g \leq 9 \\ 0.09 & \text{otherwise,} \end{cases} \tag{5.3}$$

with X_g generated from a gamma distribution with shape parameter 21 and rate parameter 3 for $g = 1, \ldots, 5000$. For each g, we randomly sample 6 observations from $N(X_g, Y_g)$. We adopt the same protocol as outlined in the previous section for outlier perturbation of the data.

The PPV and FNR for competing methodologies are given in Table 7. We include grpRSDR-OBM, grpTW(mix/mad)-OBM, and grpVM-OBM from the previous simulation. We also include grouped estimation of variance as a nonparametric function of mean, which we call grpLoess-OBM. For this methodology we use loess regression to obtain the fitted values of gene-specific variance before computing estimates of RSDSR. In order to further allow for grouped genes to share similar variation, we estimate robust standard deviation of the residuals by ordering genes according to increasing median intensity for all four methodologies. Though least influential for residuals that are standardized first, this modification could offer some stability since data is simulated with a mean-variance dependency. Table 7 also presents the PPV and FNR for grpTW(mix/mad) methodology in which genes are grouped simply in natural, sequential order. These statistics vary minimally from those reported for grpTW(mix/mad)-OBM, suggesting that grouped estimation of TW(mix/mad) protects against any contingency on the belief similar expression intensity will imply similar variability.

By modeling variance as a function of mean, grpRSDR-OBM, grpVM-OBM, and grpLoess-OBM perform considerably well. Each attains extremely high positive predictive values and near zero false negative rates. By design, we expect grpVM-

Table 7: Positive Predictive Value (PPV) and False Negative Rate (FNR) of outlier detection methodologies for simulated data with mean-variance relationship perturbed by outliers.

	Outliers (c=3)		Outliers (c=10)	
Methodology	*PPV*	*FNR*	*PPV*	*FNR*
grpRSDR-OBM	0.975	0	0.971	0
grpTW(mix/mad)-OBM	0.892	0.067	0.94	0
grpVM-OBM	0.970	0.004	0.970	0
grpLoess-OBM	0.970	0.004	0.970	0
*grpTW(mix/mad)	0.883	0.134	0.936	0.013

OBM to do well. grpVM-OBM showed no power for detecting outliers when gene variance and intensity were completely independent; however, the efficacy of the methodology significantly improves when analyzing data which is modeled with a quadratic variance-mean dependency. Also, the data structure of this simulation study guarantees that large (small) mean will imply large (small) variance. This strengthens the performance of grpRSDR-OBM since unstandardized data grouped by increasing intensity will be similar in variation. Loess regression also relates variance to mean so that grpLoess-OBM is able to efficiently utilize the structure in data for accurate variance estimation. Although grpTW(mix/mad)-OBM does well with more distant outliers, its performance suffers when c=3. Thus, for methods which do make use of data structure when one is clearly defined, we actually lose power.

Additionally, we investigate PPV and FNR for simulated data with outliers added to every 10^{th} and 20^{th} gene; results are presented in Appendix VII. As with the previous simulation study, we find that decreasing the number of outliers does not change the relative relationships between methodologies; however, the probability of correct detection among all detected outliers decreases as the number of outliers decreases. Based on the additional analysis for our first simulation study, we incorporate grpTW(mad50/std50)-OBM into our analysis of variance-mean dependency data. We

find that it outperforms the other methodologies under this scenario. Though not presented in Table 7, the PPV/FNR for simulated data with 2500 true outliers are 0.974/0 (0.821/0) for c=3 (c=10). Thus, its comparative performance varies depending on the number of outliers and outlier distance.

In the following chapter, we explore the performance of grpRSDR-OBM, grpTW(mix/mad)-OBM, and grpVM-OBM on colon cancer microarray data. We exclude grpLoess-OBM in real data analysis since its performance is equivalent to grpRSDR-OBM when analyzing data simulated to reflect real data. In addition, both are non-parametric methodologies and grpRSDR-OBM is more easily implemented and computationally efficient. We also exclude grpTW(mad50/std50)-OBM due to analytical results of the real data.

CHAPTER VI

ANALYSIS OF OUTLIER DETECTION FOR REAL DATA

6.1 Introduction

The purpose of this chapter is to illustrate the performance of grpRSDR, grpTW(mix/mad), and grpVM using real microarray data sets. Simulation studies in Chapter V present two opposing arguments for mean-variance relationship. The first assumes no relationship and the second assumes a completely dependent and quadratic relationship. Though good arguments, neither are believed to model the true non-trivial relationship between mean and variance for gene expression data. The data, which is taken from the lab of Dr. Laurie Davidson, is described in detail in Section 6.2. We outline steps for data normalization in Section 6.3. Section 6.4 is devoted entirely to data analysis and results of our findings.

In real data analysis we find within-gene distribution to be an issue which further complicates the problem, yet contributes to successful outlier classification. The distribution of within-gene expression data affects the MAD constant k (4.11). While some gene data is normally distributed, others are more uniformly distributed. Both require a different constant k (Rousseeuw and Croux, 1993) which plays a significant role in shrinkage estimation. It is beyond the scope of this dissertation to tackle the difficulties in determining the distribution of gene-by-gene data. However, we offer some demonstration of the problem and show success in the necessity of accurate MAD k constants.

Initially, we remove genes with any missing data. For each treatment, there remains over 12,500 genes for analysis. In Section 6.4.2 we present analysis of outlier detection for all genes with at least 5 observations.

6.2 Data Description

Colon cancer polysomal microarray data will be used in this paper. Researchers simulated the colonic environment in vitro in order to determine if different media could detect microbial changes influenced by cancer promotive and cancer preventative factors. The study was carried out using a $2 \times 2 \times 2$ factorial design of experiment. Male Sprague-Dawley rats were given dietary treatments composed of a fiber source - either cellulose or pectin, and fatty acids - either corn oil (enriched with Omega-6) or fish oil (enriched with Omega-9). The rats were weighed and assigned to one of the four diet treatments. Group assignments were done so that each group measured an equal initial weight. After five weeks of age, the rats were injected subcutaneously with either azoxymethane (AOM) or an equal amount of saline. Azoxymethane is a carcinogen or cancer producing agent that is colon specific. Saline is used to control for any confounding variables that would arise from injection itself. When data was collected at 6 months of age, the rats were housed in double gridded polycarbonate cages in order to prevent them from tampering with bedding or feces. At 10 months of age, the rats were relocated to wire mesh hanging cages. Rats were given unlimited access to diet and water and all animal handling procedures were approved by the Texas A&M University Laboratory Animal Care Committee.

The 8 treatments resulting from this design are denoted as Acp, Acc, Acp, Afc, Afp, scc, scp, sfc, and sfp. We compose each treatment acronym by using the first letter of the factor level. For exposure, diet, and fiber we have A or s, c or f, and c or p, respectively. The total number of arrays used in analysis for each treatment are Acc - 6, Acp - 8, Afc - 6, Afp - 8, scc - 6, scp - 5, sfc - 6, and sfp - 8.

6.3 Data Normalization

Normalization is a critical step in microarray data processing. Due to the nature of microarray analysis, measured gene expression intensities are subject to noise arising from various sources. These factors include, but are not limited to, tissue-handling, labeling efficiency, image scanning, and hybridization efficiency. The purpose of normalization is to correct microarray data for systematic and technical biases.

There are numerous normalization procedures available today. Yang et al. (2000) discuss various normalization methods in detail. Some common normalization procedures for one-channel microarray data include the globalization method, median normalization, local regression, and quantile normalization. We choose the latter in order to adjust the multiple high density oligonucleotide arrays used in this study. Quantile normalization uses the whole data set and makes no assumption about the distribution of data. By matching percentiles of each array, its effect aligns the distribution of expression intensities across arrays so that they are the same. Bolstad et al. (2003) demonstrate that the quantile method of normalization leads in performance based of bias, variance, and computational efficiency.

We utilize quantile normalization software developed by Ben Bolstad and made available in the Bioconductor *affy* package. After normalization, data are transformed via log2 transformation. This scaling adjusts the variances to be the same for all intensities. For the remainder of this paper, we refer to normalized gene expressions as data that has been both normalized and transformed.

6.3.1 Additional Bias Correction

In order to further minimize artifacts between arrays within the same treatment, we make use of regression techniques. Normalized data for each array was regressed onto a vector of median values for genes under the same treatment. Regression was carried

out using least trimmed squares (LTS) (Rousseeuw and Leroy, 1987) to estimate unknown parameters of the linear regression model. Least trimmed squares is a robust alternative to ordinary least squares (OLS) regression. It is a high breakdown point method and is able to accommodate highly contaminated data. Rather than estimating coefficients based on minimizing the sum of squared residuals, coefficients from LTS regression are estimated by minimizing the sum of the h smallest squared residuals. Namely,

$$\underset{\hat{\theta}}{\operatorname{argmin}} \sum_{i=1}^{h} (r^2)_{i:n}, \tag{6.1}$$

where $r^2_{1:n} \leq \ldots \leq (r^2)_{n:n}$ are the ordered squared residuals.

The effects of normalization are shown by plotting array data against median expression values for arrays under the same treatment. As shown in Figure 8, array bias is virtually eliminated with bias corrected normalization of the raw data for treatment Acp. We include bias correction as a precautionary step to control for array intensity based biases should there be a need after log2 transformation of the normalized data. The other treatments show similar trends.

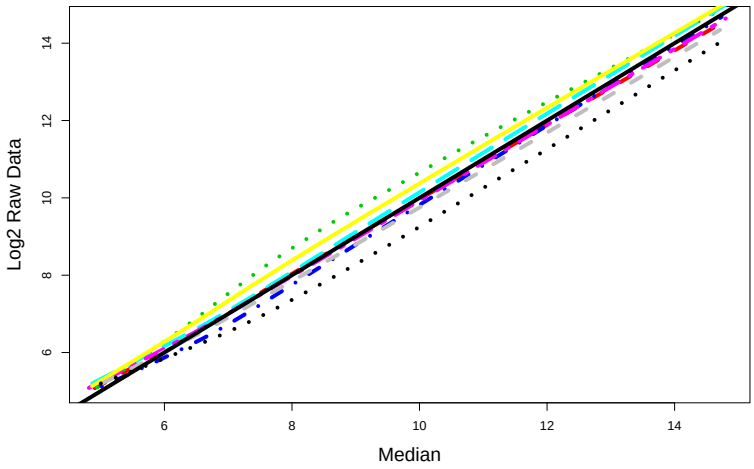
6.3.1.1 Bias Correction Algorithm

Let $\tilde{\mathbf{Y}}$ be a vector of median expressions for G genes of treatment k. For each array j of treatment k, let $\mathbf{Y}_j$ be the normalized gene expressions of G genes. We define the steps for bias correction of a single treatment to be:

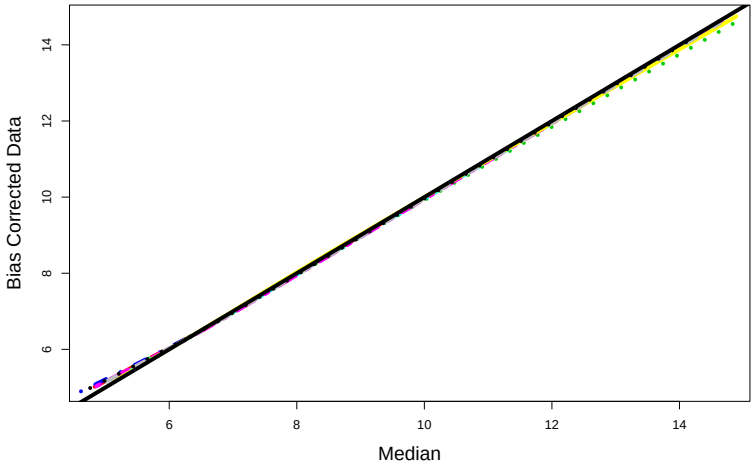
1. Obtain the fitted values, $\hat{\mathbf{Y}}_j$, from regressing $\mathbf{Y}_j$ on $\tilde{\mathbf{Y}}$,

$$\hat{\mathbf{Y}}_j = f(\tilde{\mathbf{Y}}), \tag{6.2}$$

where f is the linear function resulting from LTS regression.



(a)



(b)

Figure 8: *Two methods of reducing systematic bias in gene expression data. Data are taken from treatment Acp. A different loess curve was fit to each array data plotted against median expression values of data under the same treatment. (a) Log2 transformation of raw data. (b) Array bias corrected data as described in 6.3.1. A perfect diagonal is represented by the solid line.*

2. Calculate bias corrected values by removing array intensity based biases and adding the center back.

$$\mathbf{Y}_j^* = \mathbf{Y}_j - (\hat{\mathbf{Y}}_j - \tilde{\mathbf{Y}}) \tag{6.3}$$

3. Repeat steps 1 and 2 for each array j.

For completeness, repeat steps 1-3 for each experimental condition. We denote bias-corrected array data as $\{Y^*\}$.

6.4 Colon Cancer Microarray Data Analysis

In this section we present analyses from applying the proposed methodology to colon cancer microarray data. We consider grpRSDR-OBM, grpTW(mix/mad)-OBM, and grpVM-OBM to be competing methodologies since each perform well under certain realized data structures. grpRSDR-OBM accommodates the belief that genes with similar intensity will share similar variation and if true, will demonstrate superior performance whether the relationship between gene spread and location is parametric or nonparametric. On the other hand, grpVM-OBM is ideal if the variance of a gene truly increases as a quadratic function of its central measure. Finally, grpTW(mix/mad)-OBM presents a platform for which neither assumption must hold true.

The total number of detected outliers for data under each experimental treatment is presented in Table 8 for each methodology. After additional investigation of outlier classification for grpRSDR-OBM and grpVM-OBM, it was determined that their performance on real data is identical for nearly every gene expression. For any treatment where the counts for grpRSDR-OBM and grpVM-OBM are not the same, the method with more outliers simply picked up a few that the other did not. Thus, for the remainder of this paper we compare grpTW(mix/mad)-OBM with grpRSDR-OBM, noting that the same conclusions can be drawn for grpVM-OBM. We proceed

Table 8: The number of outliers detected at Q=0.01 in real data when using grpRSDR-OBM, grpTW(mix/mad)-OBM, and grpVM-OBM to estimate within-gene variance.

Treatment	grpRSDR-OBM	grpTW(mix/mad)-OBM	grpVM-OBM
Acc	536	142	536
Acp	1446	165	1446
Afc	1010	334	1010
Afp	746	127	746
scc	973	284	975
scp	665	989	665
sfc	674	226	674
sfp	679	121	677

with a comparative study of two methods which assume no functional relationship between gene variance and intensity.

We focus our analysis of real data on two experimental conditions - Acp and scp. This allows for the study of treatments that are both biologically and statistically interesting. Biologically, we are able to study the effect of colon cancer promotive corn oil diets in combination with pectin fiber to enhance colonic apoptosis against azoxymethane-induced colon cancer. From a statistical standpoint, results presented in Table 8 show that grpRSDR-OBM and grpTW(mix/mad)-OBM differ most in the number of outliers detected in Acp data. We also find that the only time grpTW(mix/mad)-OBM returns more outliers is when analyzing treatment scp. Taking a closer look at these two unique data sets facilitates our assessment of accuracy and reliability of the variance estimation methodologies. More specifically, we want to make certain that there is no flaw in the systematic implementation of grpTW(mix/mad)-OBM which makes it prone to a limited number of detected outliers.

6.4.1 Between Extreme Expression Deviation to MAD Ratio

The uncertainty of truth concerning outlier classification presents a challenge in real data analysis. As one tool for assessing accuracy, we introduce a statistic that measures the relative deviation level of outliers. We denote this measure as the between extreme expression deviation to MAD (BEED-MAD) ratio. For each outlier gene g, the BEED-MAD ratio is calculated as the absolute difference between the outlier expression and its closest non-outlier neighbor divided by the gene's measure of MAD. Precisely, we denote the BEED-MAD ratio of outlier gene g as

$$BEED\text{-}MAD_g = \frac{|y_{g,i} - y_{g,j}|}{MAD_g}, \tag{6.4}$$

where $y_{g,i}$ is the outlier expression and $y_{g,j}$ is the expression value of its closest non-outlier neighbor. Without loss of generality, MAD is computed with k=1 (4.11) for all g. If there are multiple outliers detected for a gene, then the BEED-MAD ratio is taken to be the average of BEED-MAD ratios for all detected outliers in that gene.

The average BEED-MAD ratio of detected outliers in each experimental condition is shown in Table 9 for both methodologies. Based on the reported averages, grpRSDR-OBM is consistently less powerful in detecting outliers that deviate most from non-outlier observations. On average, grpTW(mix/mad)-OBM detects outliers with BEED-MAD ratios that exceed those of grpRSDR-OBM by more than a factor of 3. It is important to stress that we do not devise the BEED-MAD ratio as a tool for detecting outliers. Doing so would require dependency on a cutoff value for classification. Rather, we promote a technique that is data adaptive and use the BEED-MAD ratio to give insight on the plausibility of detected outliers.

For further analysis, we pool all outliers detected by grpRSDR-OBM and grpTW(mix/mad)-OBM and order the BEED-MAD ratios from lowest to highest. A comparison of gene expression scatter for genes with the 40 lowest and 40 highest

Table 9: Average BEED-MAD ratio for detected outliers when using grpRSDR-OBM and grpTW(mix/mad)-OBM to estimate gene variability.

Treatment	grpRSDR-OBM	grpTW(mix/mad)-OBM
Acc	2.99	17.39
Acp	3.62	12.96
Afc	5.48	17.57
Afp	4.04	13.62
scc	4.87	17.61
scp	6.49	24.25
sfc	4.6	18.38
sfp	3.99	13.38

ratios for treatments Acp and scp is shown in Figures 9 and 10, respectively. Outliers detected by grpRSDR-OBM dominate the lowest BEED-MAD ratio region. For each of these genes, the between extreme expression deviation is minimal compared to within-gene variability. As a result, we observe outliers that appear to be false positive observations. This would explain the relatively high counts of detected outliers reported in Table 8 for grpRSDR-OBM and grpVM-OBM. Outliers with highest BEED-MAD ratios are shown in the bottom panel of Figures 9 and 10. For Acp data, there are many distant outliers that both methodologies detect, many outliers that grpTW(mix/mad)-OBM uniquely detects, and only two outliers that are uniquely detected by grpRSDR-OBM. For scp, the grpTW(mix/mad)-OBM methodology is superior in detecting outliers which deviate most from their non-outlier neighbors. Thus, estimating variance using grouped estimation of TW(mix/mad) does well in identifying observations that appear to be true outliers.

The top panels of Figures 9 and 10 affirm our conjecture and support previous findings that grpRSDR-OBM is too liberal in outlier detection. Ironically, it tends to be most conservative when handling genes for which positive outlier identification is likely to be feasible. The grpTW(mix/mad)-OBM methodology is shown to suffer

some when gene expressions are more uniformly than normally distributed. This is evident with the high BEED-MAD ratio outliers in Figure 9 that are uniquely detected by grpRSDR-OBM. We believe that these points are rejected as outliers by grpTW(mix/mad)-OBM because the distribution of expressions within those genes are more uniform. After setting the MAD constant to 1.15 (an appropriate constant for making the estimator consistent in estimating σ for uniformly distributed data) for only those two genes and rerunning the algorithm, we analyze the same genes and find that the two observations in question have also been identified by grpTW(mix/mad)-OBM. The remainder of observations in Figure 9 had no change in outlier classification by either method. This fact suggests that further improvement can be obtained if an improved robust measurement of the heterogeneity level of gene-specific variation can be proposed.

6.4.2 Outlier Detection for Genes with Sample Size ≥ 5

In order to detect outliers in genes with missing data, we make a small modification to the average risk function (4.9) used in determining the optimal shrinkage parameter. This function was originally specified for gene expression data with a fixed degrees of freedom. However, this constraint does not allow for variance estimation of genes with varying sample sizes so we de-generalize the average risk function so that genes with missing data, but sufficient sample size can be included in analysis. The updated risk function is given by

$$\begin{aligned} R(\sigma^{2t}, \hat{\sigma}^{2t}) &= \frac{1}{G}\sum_{g=1}^{G} E(L(\sigma_g^{2t}, \hat{\sigma}_g^{2t})) \\ &= \frac{1}{G}\sum_{g=1}^{G} \frac{h_G^{\alpha}(t,\nu_g)h_1^{1-\alpha}(t,\nu_g)}{h_1^{G-1}(\frac{\alpha t}{G},\nu_g)h_1((1-\alpha+\frac{\alpha}{G})t,\nu_g)}(\sigma_{pool}^2)^{\alpha t}(\sigma_g^2)^{-\alpha t} \\ &\quad -ln\ (h_G^{\alpha}(t,\nu_g)h_1^{1-\alpha}(t,\nu_g)) - t\Psi(\frac{\nu_g}{2}) + tln\ (\frac{\nu_g}{2}) - 1, \end{aligned} \tag{6.5}$$

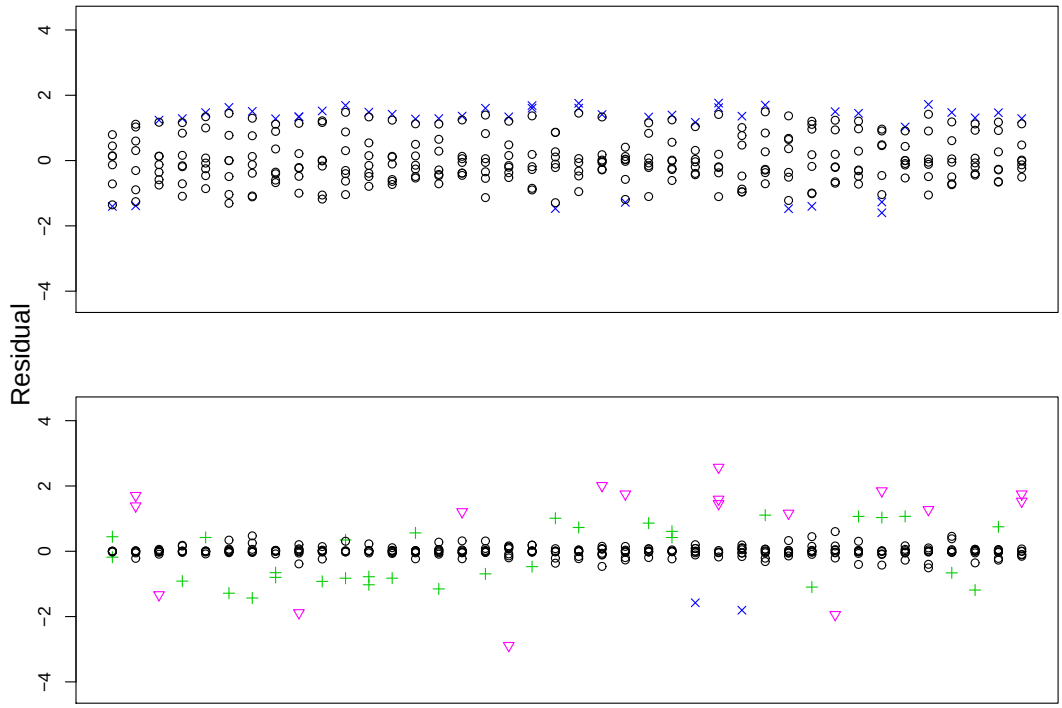


Figure 9: *Scatterplot of residual values for genes with outlier expressions for Acp data. Outliers detected by grpRSDR-OBM and grpTW(mix/mad)-OBM are pooled together and ordered by BEED-MAD ratios. Outlier genes with the 40 lowest and 40 highest ratios are plotted in the top and bottom panels, respectively. Genes in the top panel are ordered by increasing BEED-MAD ratio and genes in the bottom panel are ordered by decreasing BEED-MAD ratio. Outliers are classified by either grpRSDR-OBM (×), grpTW(mix/mad)-OBM (plus sign), or both (triangle).*

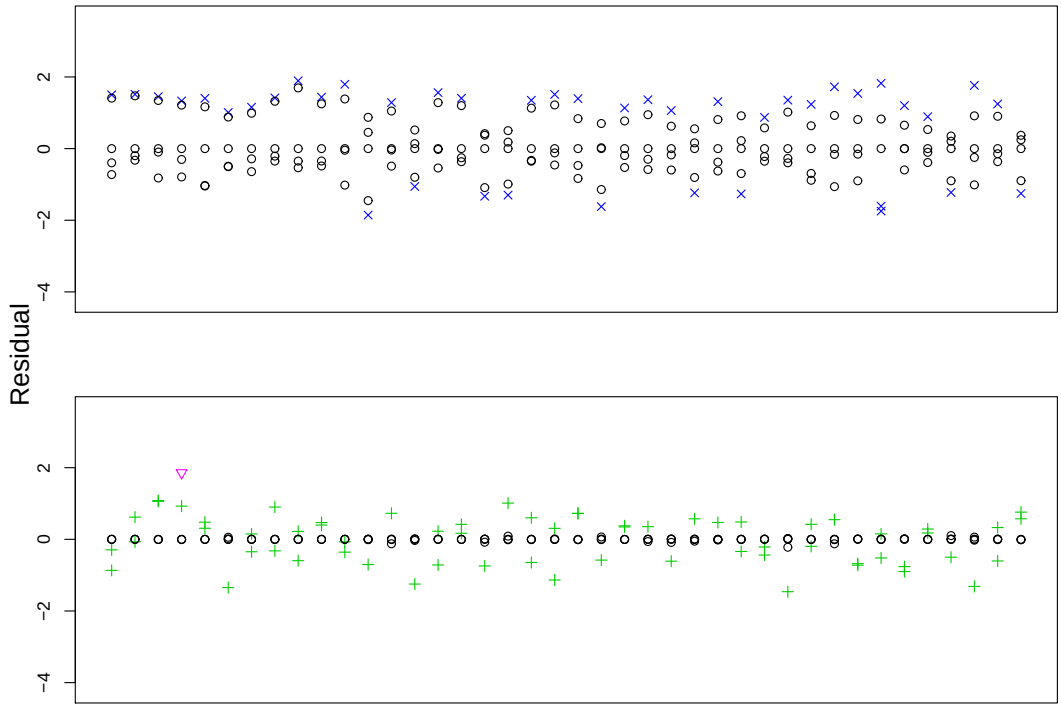


Figure 10: *Scatterplot of residual values for genes with outlier expressions for scp data. Outliers detected by grpRSDR-OBM and grpTW(mix/mad)-OBM are pooled together and ordered by BEED-MAD ratios. Outlier genes with the 40 lowest and 40 highest ratios are plotted in the top and bottom panels, respectively. Genes in the top panel are ordered by increasing BEED-MAD ratio and genes in the bottom panel are ordered by decreasing BEED-MAD ratio. Outliers are classified by either grpRSDR-OBM (×), grpTW(mix/mad)-OBM (plus sign), or both (triangle).*

Table 10: Number of detected outliers (average BEED-MAD ratio) for detected outliers when using grpRSDR-OBM and grpTW(mix/mad)-OBM to estimate gene variability for genes with sample size ≥ 5.

Treatment	grpRSDR-OBM	grpTW(mix/mad)-OBM
Acc	694 (2.96)	336 (17.12)
Acp	1795 (3.58)	407 (12.56)
Afc	1241 (5.41)	548 (17.06)
Afp	1012 (3.99)	307 (13.47)
scc	1124 (4.96)	477 (17.25)
scp	665 (6.49)	989 (24.25)
sfc	767 (4.52)	411 (18.02)
sfp	963 (3.98)	325 (13.19)

where

$$h_n(t,\nu) = (\frac{\nu}{2})^t \left(\frac{\Gamma(\frac{\nu}{2})}{\Gamma(\frac{\nu}{2}+\frac{t}{n})} \right)^n, \tag{6.6}$$

$t = 1$, $\Psi(t) = \Gamma'(t)/\Gamma(t)$ is the digamma function, and ν_g is the degrees of freedom for gene g.

We incorporate into our analysis genes with at least 5 observations for sufficient sample size. In Table 10 we report the updated number of detected outliers and the average BEED-MAD ratio for these outliers. Again, we find that the average BEED-MAD ratio for outliers detected by grpTW(mix/mad)-OBM exceed the average BEED-MAD ratio for outliers detected by grpRSDR-OBM by a factor of 3 for every treatment. Our findings also show consistency in the quality of outlier observations for both methodologies. The average BEED-MAD ratio for outliers detected in genes with varying sample sizes are congruous with those obtained when analyzing only genes with complete data. We still find that grpRSDR-OBM detects observations that appear to be false positives, while on average, grpTW(mix/mad)-OBM classifies outlier observations which deviate most from their non-outlier neighbors.

6.4.3 Additional Analysis and Discussion

Through our consideration to perturb simulated data with fewer outliers, we find that introducing a cutoff criterion for specifying which genes to send only the square of the MAD statistic into our modification of the TW algorithm may not prove superior based on the number of true outliers. Thus, we analyze outlier detection for real data using grpTW(mad50/std50)-OBM to estimate variance. In Appendix VII we present tables which report the number of detected outliers and the average BEED-MAD ratio for each experimental treatment. grpTW(mad50/std50)-OBM not only detects considerably less outliers than grpTW(mix/mad)-OBM, but also averages lower BEED-MAD ratios. It appears that properties of real data which depart from structured simulated data make it necessary to assess the likelihood of an outlier gene in order to determine which variance statistics should be sent into Tong & Wang's optimal shrinkage algorithm. Thus, for real colon cancer data, we analyze the performance of grpRSDR-OBM and grpTW(mix/mad)-OBM.

The numbers of detected outliers presented in Tables 8 and 10 definitely reveal discrepancies in the competing methodologies. On average, grpRSDR-OBM and grpVM-OBM detect hundreds more outliers than grpTW(mix/mad)-OBM. However, based on the BEED-MAD averages which range between 3 and 6.5 for grpRSDR-OBM and grpVM-OBM and 13 and 24 for grpTW(mix/mad)-OBM, we can deduce that many of these observations are indeed false positives. This is especially clear in the gene expression scatter plot for selected outlier genes presented in Figures 9 and 10.

The consistency in grpTW(mix/mad)-OBM's ability to classify relatively distant observations as outliers convince us that differences in the number of detected outliers across treatments are ultimately explained by uniqueness in data. As an

example let us consider the number of outlier observations for our representative treatments. grpTW(mix/mad)-OBM detects 407 outlying observations in treatment Acp and detects 989 outliers in treatment scp. Yet, the average BEED-MAD ratio of outliers is approximately 3 times the average BEED-MAD ratio for outliers detected by grpRSDR-OBM. This is true for both treatments. We find that despite the differences in numbers, the quality of information is equally powerful because grpTW(mix/mad)-OBM is data adaptive.

It is clear that the benefits of grouping are impressive; however, grouping genes to estimate variance is only effective if genes share similar variability. If variance is non-homogenous across genes within a group, then methodologies which depend on this assumption are subject to problems which stem from inaccurate variance estimation. For grouped estimation of robust standard deviation of the residuals, the effect of variance heterogeneity is amplified since the procedure groups non-standardized expressions. For these microarray data, most genes have low within-gene variability. Thus for grpRSDR-OBM, the 68% percentile of residuals is likely to be small. The variance for genes which are truly more variable, but share similar median intensity, will be underestimated. Consequently, test statistics for these high variable gene expressions will be inflated, resulting in an increased number of false positives.

In addition, it appears that grpRSDR-OBM classifies an observation as an outlier based on its deviation from the median while completely ignoring the within gene variability. We see this to be evident in Figures 9 and 10 where outliers detected by grpRSDR-OBM are more reflective of a threshold criterion. This would be acceptable if genes shared similar variability; however, we observe variance to be heterogeneous across grouped genes. If the measure of variability for a gene is large, then an expression with a large residual may not be a legitimate outlier. Moreover, if the within-gene variability is small, then an observation with a mild residual value

could very well be an outlier. Thus, we expect to see some data with small residuals classified as outliers. This is especially true when within-gene variability varies. We are not interested in selecting outliers based on a cutoff for residuals. This approach is known to favor high variable genes. Rather, our intent is to truly identify expressions that are statistically distant from their within-gene neighbors.

Outlier detection by grpVM-OBM methodology also suffers in estimation of the true variance. Recall that its detection of outliers is almost completely synonymous with that of grpRSDR-OBM. Thus, it, too, is extremely liberal with classification of high variable genes. We observed superior performance in the proportion of correct detection when the procedure was implemented on simulated data generated to have a quadratic relationship between location center and spread. Undeniably it does well under this scenario; however, results of real data analysis are indicative of drawbacks in the methodology when the model is misspecified.

Our proposed methodology shows strength here because we abandon any reliance on non-random data structure. Although genes are ordered by median, the methodology itself is not constructed under this assumption. For simulated data, grpTW(mix/mad)-OBM has considerable power in detecting outliers although it suffers some when the relationship between gene variance and intensity is perfectly defined. We observe superior performance in real data analysis for which gene centrality and variability has neither a completely random or completely specified relationship. In microarray data analysis, assumptions of variance heterogeneity across genes and gene variance-intensity dependency is important in gene-specific variance estimation. Such assumptions are powerful and can either improve or fracture the performance of methodologies depending on realized data structure. When the true data structure and outlier features are unknown, we observe substantial power in using robust methodology for optimal shrinkage variance estimation which requires no assumptions

about data structure.

CHAPTER VII

CONCLUSION

The success of both colon cancer screening methodologies hinges on the ability of mRNA to move the information contained in DNA to the translation machinery. Obtaining gene expressions from mucosa cells presents no issue of partially degraded mRNA; unfortunately, we do not have the same assurance with fecal array data. Colon cancer takes lives of more than fifty thousand Americans every year because only 38% of cases are diagnosed at an early stage. The potential of being able to recover information from fecal samples is important since technological advances seek to provide at-risk patients with a non-invasive alternative for testing their susceptibility to colon cancer.

Although formal tests have shown that genetic material collected from feces is not as good as that collected from mucosal cells within the colon, it is still believed that partially degraded RNA samples can produce meaningful measurements (Schoor et al., 2003). Research is currently underway to develop normalization techniques that can accurately measure gene expression levels from partially degraded genetic materials. Liu et al. (2005) report a new two stage semiparametric normalization method that performs favorably when compared to the global median and quantile normalization methods. Also promising to this area of research is the demonstration by Kanaoka et al. (2004) in isolating intact fecal eukaryotic mRNA.

Our investigations suggest that even though there tends to be a higher proportion of genes that have low reproducibility in the fecal array data than in the mucosa array data, for the group of genes which possess high ICC values in the fecal data, their reproducibility is no worse than that of the mucosa data. We note that the inverse

probit transformation of ICC values enables us to easily adopt the mixture of normal modeling approach carried out by MCLUST. Our findings also suggest that this modeling approach allows us to separate the low and high ICC components so that meaningful conclusions can be reached. In our study investigating mixture modeling of intraclass correlation coefficients, we demonstrate that the normal mixture model is less sensitive toward model mis-specification than that of mixture of betas. By modeling the IPT-ICC values with normal mixtures, we are able to obtain accurate density parameter estimates. We also avoid the problem of estimating parameters for a beta distribution which may increase to infinity at the boundaries. We risk losing valuable information if we incorrectly model data with a mixture of betas.

The objective of the second study was to develop a robust methodology for estimating gene-specific variance. It is motivated by the need for adaptive techniques that are able to accommodate limitations seen in microarray gene expression data. We believe that a suitable method should be insensitive to outliers, free of any assumptions, effective for small sample size, and data adaptive. Ultimately, we desire accurate estimation of variance in order to detect outlier observations. The successful identification and handling of outliers will, in turn, lead to reliable variance estimation which is essential for detecting differentially expressed genes.

Numerous innovative approaches have been devised to strengthen the variance term, many of which center around the idea of gathering information across genes. Although a common trend in estimation procedures, methodologies which group genes by similar median or mean intensity will suffer greatly from inaccurate variance estimation when gene-specific variance is heterogeneous across genes. Likewise, methodologies which assume a quadratic variance-mean dependency in the data will also be less powerful in detecting outliers when the assumption does not hold. From simulation studies based on parametric and nonparametric data structures, we observe

the strengths and weaknesses of both approaches. When data is generated to model variance as a quadratic function of center location then both methodologies perform considerably well. However, both suffer to some extent when data is generated to have a completely random data structure. The former shows suboptimal false negative rates and the latter is unable to successfully identify almost any outliers.

In response to these drawbacks, we extend the optimal shrinkage variance estimation work of Tong and Wang (2007). We formulate a robust modification which is necessary for microarray data analysis. Simulation studies and real data analysis in this dissertation demonstrate that our methodology is capable of handling both functional and non-functional relationships in data structure. Our proposed methodology performs well with genes of sample size as small as 5. For real data analysis we support our findings based on a statistic that measures the relative deviation level of outliers. On average, this measure for outliers detected by our methodology, is more than 3 times that of the other methodologies. We find that our methodology consistently detects observations that are relatively distant from their non-outlier neighbors.

Our work also contributes to previous work which promotes the argument that there are strengths in numbers. Even methodology which is believed to be extremely effective in estimating variance for small sample size performs poorly unless we consider an additional grouped estimated variance. There are substantial power increases when gene-by-gene variance is computed to be $c_i\sigma^2$, where c_i and σ^2 are gene-specific and local common variance estimates, respectively. We emphasize that grouping will be ineffective when similar intensity does not imply similar variation. Methodologies which produce gene-specific variance estimates have an advantage here because gene-by-gene variance estimates allow for standardized test statistics, which have a standard normal distribution. By grouping data that come from the same distribu-

tion, the information shared by genes after standardization is efficiently utilized.

REFERENCES

Allison, D., Gadbury, G., Heo, M., Fernandez, J., Lee, C., Prolla, T., and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis* **39**, 1–20.

Altman, D. G. and Bland, J. M. (1994). Diagnostic tests 2: predictive values. *British Medical Journal* **309**, 102.

Arfin, S., Long, A., Ito, E., Tolleri, L., Riehle, M., Paegle, E., and Hatfield, G. (2000). Global gene expression profiling in Escherichia coli K12: the effects of integration host factor. *Journal of Biological Chemistry* **275**, 29672–29684.

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

Barnett, V. and Lewis, T. (1984). *Outliers in Statistical Data*. New York: John Wiley & Sons.

Benjamini, Y. and Hochberg, T. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **85**, 289–300.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193.

Carrasco, J. and Jover, L. (2002). Estimating the generalized concordance correlation coefficient through varince components. *Biometrics* **59**, 849–858.

Carrasco, J. and Jover, L. (2003). Estimating the generalized concordance correlation coefficient through variance components. *Biometrics* **59**, 849–858.

Comander, J., Natarajan, S., Gimbrone, M., and Garcia-Cardena, G. (2004). Improving the statistical detection of regulated genes from microarray data using intensity-based variance estimation. *BMC Genomics* **5**, 17.

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

Davidson, L., Lupton, J., Miskovsky, E., Fields, A., and Chapkin, R. (2003). Quantification of human intestinal gene expression profiles using exfoliated colonocytes: a pilot study. *Biomarkers* **8**, 51–61.

Davidson, L., Nguyen, D., Hokanson, R., Callway, E., Isett, R., Turner, N., Dougherty, E., Wang, N., Lupton, J., and Carroll, R. (2004). Chemopreventive n-3 polyunsaturated fatty acids reprogram genetic signatures during colon cancer initiation and progression in the rat. *Cancer Research* **64**, 6797–6804.

Davies, L. and Gather, U. (1993). The identification of multiple outliers. *Journal of the American Statistical Association* **88**, 782–801.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood for incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society Series B* **39**, 1–38.

Du, Y., Lenz, J., and Arvidson, C. (2005). Global gene expression and the role of sigma factors in neisseria gonorrhoeae in interactions with epithelial cells. *Infection and Immunity* **73**, 4834–4845.

Durbin, B., Hardin, J., Hawkins, D., and Rocke, D. (2002). A variance stabilizing transformation for gene-expression microarray data.. *Bioinformatics* **18 Suppl. 1**, S105–S110.

Fleiss, J. (1981). *Statistical Methods for Rates and Proportions (2nd ed.)*. New York: John Wiley & Sons.

Fraley, C. and Raftery, A. (1999). Mclust: Software for model-based cluster analysis and discriminant analysis. Tech. rep. 342, University of Washington.

Fraley, C. and Raftery, A. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**, 611–631.

He, Y., Pan, W., and Lin, J. (2006). Cluster analysis using multivariate normal mixture models to detect differential gene expression with microarray data. *Computational Statistics and Data Analysis* **51**, 641–658.

Huang, X. and Pan, W. (2002). Comparing three methods for variance estimation with duplicated high density oligonucleotide arrays. *Functional & Integrative Genomics* **2**, 126–133.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18 Suppl. 1**, S96–S104.

Jain, N., Thatte, J., Braciale, T., Ley, K., O'Connell, M., and Lee, J. (2003). Local-pooled-error test for identifying differentially expressed genes with a small number of replicated arrays. *Bioinformatics* **19**, 1045–1951.

Ji, Y., Wu, C. nd Liu, P., Wang, J., and Coombes, K. (2005). Applications for beta-mixture models in bioinformatics. *Bioinformatics* **21**, 2118–2122.

Kamb, A. and Ramaswami, A. (2001). A simple method for statistical analysis of intensity differences in microarray-derived gene expression data. *BMC Biotechnology* **1**, 8.

Kanaoka, S., I., Y. K., Miura, N., Sugimura, H., and Kajimura, M. (2004). Potential usefulness of detecting cyclooxygenase 2 messanger RNA in feces for colorctal cancer screening. *Gastroenterology* **127**, 422–427.

Kendziorski, C., Newton, M., Lan, H., and Could, M. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22**, 3899–3914.

Li, C. and Wong, W. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences USA* **98**, 31–36.

Lin, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 255–268.

Lin, Y., Nadler, S., Lan, H., Attie, A., and Yandell, B. (2003). Adaptive gene picking with microarray data: detecting important low abundance signals. In *The Analysis of Gene Expression Data: Methods and Software,* G. Parmigiani, E.S. Garrett, R.A. Irizarry, and S.L. Zeger (eds), 291-312. New York: Springer.

Lindley, D. (1962). Discussion of Professor Stein's paper, 'confidence sets for the mean of a multivariate normal distribution'. *Journal of the Royal Statistical Society, Series B* **24**, 265–296.

Liu, L., Wang, N., Lupton, J., Turner, N., Chapkin, R., and Davidson, L. (2005). A

two-stage normalization method for partially degraded mRNA microarray data. *Bioinformatics* **21**, 4000–4006.

Lonnstedt, I. and Speed, T. (2002). Replicated microarray data. *Statistica Sinica* **12**, 31–46.

Mariani, T., Budhraja, V., Mecham, B., Gu, C., Watson, M., and Sadovsky, Y. (2003). A variable fold-change threshold determines significance for expression microarrays. *FASEB J.* **17**, 321–323.

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

Moffitt, R., Phan, J., Hemby, S., and Wang, M. (2005). Effect of outlier removal on gene marker selection using support vector machines. *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the IEEE*, 917–920.

Motulsky, H. and Brown, R. (2006). Detecting outliers when fitting data with nonlinear regression - a new method based on robust nonlinear regression and the false discovery rate. *BMC Bioinformatics* **7**, 123.

Nguyen, D., Arpat, A., Wang, N., and Carroll, R. (2002). DNA microarray experiments: biological and technological aspects. *Biometrics* **58**, 701–717.

Prolla, T. (2002). DNA microarray analysis of the aging brain. *Chemical Senses* **27**, 299–306.

Rocke, D. and Durbin, B. (2001). A model for measurement error for gene expression arrays. *Journal of Computational Biology* **8**, 557–569.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.

Rousseeuw, P. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association* **88**, 1273–1283.

Schoor, O., Weinschenk, T., Hennenlotter, J., Corvin, S., Stenzl, A.and Rammensee, H. G., and Stevanović, S. (2003). Moderate degragradation does not preclude micoarray analysis of small amounts of RNA. *BioTechniques* **35**, 1192–1201.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Shrout, P. and Fleiss, J. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* **86**, 420–428.

Strimmer, K. (2003). Modeling gene expression measurement error: a quasi-likelihood approach. *BMC Bioinformatics* **4**, 10.

Tong, T. and Wang, Y. (2007). Optimal shrinkage estimation of variances with applications to microarray data analysis. *Journal of the American Statistical Association* **10**, 113–122.

Tusher, V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences USA* **98**, 1516–1521.

Yang, S., Guo, X., Yang, Y., Papcunik, D., Heckman, C., Hooke, J., Shriver, C., Liebman, M., and Hu, H. (2006). Detecting outlier microarray arrays by correlation and percentage of outlier spots. *Cancer Informatics* **2**, 351–360.

Yang, Y., Dudoit, S., Luu, P., and Speed, T. (2000). Normalization for cDNA microarray data. Tech. rep. 589, University of California - Berkeley.

APPENDIX A

ADDITIONAL ANALYSIS OF SIMULATED STUDIES PRESENTED IN CHAPTER V

Simulation I: Independent Gene Variance-Intensity Relationship

In this section we report the PPV and FNR for gene data simulated to have no relationship between location center and spread. Our analyses in Chapter V reflect a study where outliers are added to every other gene, resulting in 2500 outliers. Alternatively, we consider the effects of adding a total of 500 and 250 outliers. Tables 11 and 12 present results where outliers are added to every 10^{th} and 20^{th}, respectively.

Table 11: Positive Predictive Value (PPV) and False Negative Rate (FNR) of outlier detection methodologies for simulated data perturbed by outliers added to every 10^{th} gene.

	Outliers (c=3)		Outliers (c=10)	
Methodology	*PPV*	*FNR*	*PPV*	*FNR*
MAD	0.620	0.964	0.887	0.215
grpMAD	0.338	0.103	0.361	0
grpSTD	0	1	0	1
TW(var)	0	1	0	1
grpTW(var)	0.932	0.891	0.997	0.115
grpTW(mad)	0.511	0.051	0.522	0
grpTW(mad50/std50)	0.951	0.352	0.968	0
grpTW(mix/mad)	0.798	0.434	0.867	0
grpRSDR-OBM	0.760	0.141	0.776	0
grpVM-OBM	0	0.167	0	0.167

Simulation II: Gene Variance-Intensity Dependency

In this section we report the PPV and FNR for data which is simulated such that the variance of a gene is a quadratic function of its mean intensity. Tables 13 and 14

Table 12: Positive Predictive Value (PPV) and False Negative Rate (FNR) of outlier detection methodologies for simulated data perturbed by outliers added to every 20^{th} gene.

	Outliers (c=3)		Outliers (c=10)	
Methodology	*PPV*	*FNR*	*PPV*	*FNR*
MAD	0.421	0.975	0.831	0.348
grpMAD	0.200	0.111	0.218	0
grpSTD	0	1	0	1
TW(var)	0	1	0	1
grpTW(var)	0.885	0.911	0.995	0.026
grpTW(mad)	0.347	0.063	0.360	0
grpTW(mad50/std50)	0.939	0.474	0.964	0
grpTW(mix/mad)	0.648	0.525	0.784	0
grpRSDR-OBM	0.651	0.167	0.677	0
grpVM-OBM	0	0.168	0	0.168

present results where outliers are added to every 10^{th} and 20^{th} gene, respectively.

Table 13: Positive Predictive Value (PPV) and False Negative Rate (FNR) of outlier detection methodologies for simulated data with mean-variance relationship perturbed by outliers added to every 10^{th} gene.

	Outliers (c=3)		Outliers (c=10)	
Methodology	*PPV*	*FNR*	*PPV*	*FNR*
grpRSDR-OBM	0.915	0.006	0.916	0
grpTW(mix/mad)-OBM	0.863	0.222	0.886	0
grpVM-OBM	0.916	0.005	0.916	0
grpLoess-OBM	0.916	0.090	0.915	0
grpTW(mad50/std50)-OBM	0.974	0	0.984	0
*grpTW(mix/mad)	0.844	0.318	0.882	0

Table 14: Positive Predictive Value (PPV) and False Negative Rate (FNR) of outlier detection methodologies for simulated data with mean-variance relationship perturbed by outliers added to every 20^{th} gene.

	Outliers (c=3)		Outliers (c=10)	
Methodology	*PPV*	*FNR*	*PPV*	*FNR*
grpRSDR-OBM	0.885	0.015	0.886	0
grpTW(mix/mad)-OBM	0.740	0.398	0.817	0
grpVM-OBM	0.885	0.014	0.885	0
grpLoess-OBM	0.885	0.015	0.884	0
grpTW(mad50/std50)-OBM	0.972	0.180	0.982	0
*grpTW(mix/mad)	0.727	0.433	0.813	0

APPENDIX B

ADDITIONAL ANALYSIS OF REAL DATA PRESENTED IN CHAPTER VI

The additional simulated studies (as reported in Appendix A) suggest that grouped estimation of TW(mad50/std50) may be a reasonable methodology for estimating gene-specific variance. We report the number of outliers and average BEED-MAD ratios for each treatment of the real data analysis in Table 15.

Table 15: Number of detected outliers (average BEED-MAD ratio) for detected outliers when using grpTW(mad50/std50)-OBM to estimate gene variability for genes with complete data and those with sample size ≥ 5.

Treatment	Complete Data	Sample Size ≥ 5
Acc	4 (11.95)	5 (13.04)
Acp	29 (8.15)	71 (6.8)
Afc	86 (9.89)	132 (9.12)
Afp	37 (10.01)	76 (9.01)
scc	60 (11.15)	72 (11.00)
scp	31 (16.43)	31 (16.43)
sfc	8 (7.46)	21 (9.39)
sfp	26 (9.11)	70 (8.06)

VITA

Nysia I. George was born in Baton Rouge, Louisiana in 1981. She received her high school education from the Talented and Gifted Magnet High School in Dallas, Texas. In August of 1999, she entered the undergraduate program in applied mathematics at Texas A&M University in College Station, Texas and graduated magna cum laude with a Bachelor of Science degree in May 2003. She continued her studies at Texas A&M, earning a Master of Science degree in statistics in December 2005 and then a Doctor of Philosophy degree in statistics in August 2008; both under the advisement of Dr. Naisyin Wang. Her permanent address is:

319 Broken Pine Ct.

Conroe, Texas 77304