

**THE USE OF MULTIPLE GROUP OUTLIER DETECTION METHODS
TO IDENTIFY INFORMATIVE BRAIN REGIONS IN MAGNETIC
RESONANCE IMAGES**

by

Nathan A. Pugh

B.S. in Biology, Juniata College, 2006

Submitted to the Graduate Faculty of

Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH
GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Nathan A. Pugh

It was defended on

April 17, 2013

and approved by

Dissertation Director:

Lisa Weissfeld, Ph.D., Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Committee Members:

Yan Lin, Ph.D., Research Assistant Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Joyce Chang, Ph.D., Associate Professor, Department of Medicine, School of Medicine,
University of Pittsburgh

Howard Aizenstein, M.D., Ph.D., Associate Professor, Department of Psychiatry, School of
Medicine, University of Pittsburgh

**THE USE OF MULTIPLE GROUP OUTLIER DETECTION METHODS TO
IDENTIFY INFORMATIVE BRAIN REGIONS IN MAGNETIC RESONANCE
IMAGES**

Nathan Pugh, PhD

University of Pittsburgh, 2013

ABSTRACT

The discovery of genetic markers that exhibit differential expression is of great interest in cancer studies. Researchers have now looked to find ways to identify genes with different expression patterns that exist only in a subset of the disease samples. Recently, a class of outlier detection methods has been developed to search for genes with outlier subsets. Using this approach, results in increased power to detect differences across groups relative to standard methods for group comparisons. Outlier detection has also been extended to handle multiple disease groups that are relevant to many more studies. The purpose of this research is to provide a comprehensive review of the class of two-group outlier detection methods which has been limited to date. From these results a modification is proposed to an existing method and the performance of this modification is examined via simulation studies. In addition, three extensions of two-group outlier detection methods are proposed to handle multiple group comparisons. Lastly, a novel application of these methods to structural magnetic resonance

imaging data to identify informative brain regions related to cognitive decline in elderly adults is discussed.

Public Health Significance: Outlier detection is a significant contribution to public health as a method that allows researchers to investigate high-dimensional data where issues such as heterogeneity and multiple comparisons are problematic. These methods allow for the identification of factors, such as genes or brain regions, that are related to group membership while identifying homogeneous subpopulations in the data.

Keywords: Outlier Detection, Structural Magnetic Resonance Imaging, High Dimensional Data, Differential Gene Expression, False Discovery Rates

TABLE OF CONTENTS

1.0	INTRODUCTION.....	1
2.0	BACKGROUND	4
3.0	AIMS	11
4.0	TWO-GROUP OUTLIER DETECTION.....	12
4.1	ABSTRACT.....	12
4.2	INTRODUCTION	13
4.3	STATISTICAL METHODS.....	15
4.3.1	T-statistic	15
4.3.2	Cancer Outlier Profile Analysis	16
4.3.3	Outlier Sum statistic.....	16
4.3.4	Outlier Robust t-statistic	16
4.3.5	Maximum Ordered Subset t-statistic.....	17
4.3.6	Least Sum of Ordered Subset Square t-statistic	17
4.3.7	Modified Least Sum of Ordered Subset Square t-statistic	18
4.4	SIMULATION STUDY.....	19
4.4.1	Methods to Evaluate Performance.....	19
4.4.2	Simulation Analysis.....	19
4.4.3	Simulation Results	20

4.5	APPLICATION	28
4.6	DISCUSSION.....	31
5.0	MULTI-GROUP OUTLIER DETECTION.....	33
5.1	ABSTRACT.....	33
5.2	INTRODUCTION	34
5.3	STATISTICAL METHODS.....	35
5.3.1	F-statistic	36
5.3.2	Outlier Robust F-statistic.....	37
5.3.3	Outlier F-statistic.....	37
5.4	PROPOSED METHODS	38
5.4.1	Multi-Group MOST	38
5.4.2	Multi-Group LSOSS.....	39
5.4.3	Modified Multi-Group LSOSS.....	41
5.5	SIMULATION STUDY.....	42
5.5.1	Simulation Setup.....	42
5.5.2	Simulation Results	43
5.6	APPLICATION	49
5.7	DISCUSSION.....	52
6.0	DISCUSSION	55
	BIBLIOGRAPHY.....	58

LIST OF TABLES

Table 1. Top 10 brain regions identified by each two-group outlier detection method.	29
Table 2. Top 15 brain regions identified by each multi-group outlier detection method	51

LIST OF FIGURES

Figure 1. <i>Contour plots of the AUC from each two-group outlier detection method.</i>	24
Figure 2. <i>Contour plots of AUC differences between the t-statistic and the outlier detection methods.</i>	25
Figure 3. <i>Contour plots of the differences in AUC between the outlier detection methods.</i>	26
Figure 4. <i>The false discovery rates versus the probability of all regions called significant.</i>	27
Figure 5. <i>Distribution of voxel counts of caudate right brain region overall and by outlier status.</i>	30
Figure 6. <i>Contour plots of the AUC from each multi-group outlier detection method.</i>	45
Figure 7. <i>Contour plots of the differences in AUC between the outlier detection methods and the F-statistic.</i>	46
Figure 8. <i>Contour plots displaying the difference in AUC in comparisons between the multi-group outlier detection methods.</i>	47
Figure 9. <i>False discovery rates of the two-group outlier detection methods.</i>	48
Figure 10. <i>Trajectory plot of 3 groups modeled using the neuropsychological memory test scores.</i>	50
Figure 11. <i>Distribution of voxel counts of Brodmann Area 28 right brain region overall, and by outlier status.</i>	52

1.0 INTRODUCTION

Structural magnetic resonance imaging (MRI) data measures the volume of brain matter that occupies the brain as well as the hundreds of smaller regions that are mapped. The scans can provide estimates of the volume of the different brain matter components, gray matter, white-matter hyperintensities and cerebral spinal fluid. It is believed that the volume of a particular region or changes to the volumes of these regions may be related to progression of cognitive decline in elderly adults. To understand how the size of brain regions is related to cognitive decline, it is necessary to identify which regions have different volumetric patterns between normal samples and cognitively abnormal samples which can be obtained from structural MRI scans. This task of identifying the differences between normal and abnormal samples is complicated by the existence of heterogeneous patterns. Standard methods have limitations in applications resulting in reduced power to identify significant differences relating to declining cognitive ability. This requires novel methodology to be applied to this data to enhance the discovery of informative brain regions involved in cognitive decline in elderly adults.

Structural MRI data is collected as high-dimensional data where several hundred brain regions or several hundred thousand voxel level markers are measured. The structure of these data introduces many challenges at the analysis level due to the large number of potential variables and limited number of samples making it difficult to identify the predictors of greatest interest. A common approach to analyzing high-dimensional data is to search for markers with

different expression patterns between groups, for example, identifying genes that are differentially expressed (DE) between cases and controls. Multiple comparisons occur when comparing normal versus diseased samples for differential expression levels among the thousands of genes measured by microarray chips. However, correcting for multiple comparisons is generally too conservative with high-dimensional data, resulting in very few or no significant findings. Another challenge in such studies is data within a group are heterogeneous or have a subset of outliers from an alternate distribution. In a microarray study, a gene may have a subgroup of differentially expressed samples which can avoid detection by standard methods that assume the entire group follows one distribution. In the two group comparison scenario, the t -statistic does not account for the possibility that samples from one classification may be derived from two distributions. As a result of this phenomenon of large variation, these tests will have low power.

Researchers have noted that oncogene activation does not occur in every disease sample. This oncogene activation pattern is analogous to any data for which expression levels differ within disease samples. In the microarray data analysis literature, a group of methods have been proposed to account for this heterogeneity of the cases, called outlier detection [1-5], while addressing the issues of high dimensional data that is problematic with standard methods. These methods aim to identify a subgroup of cases that determine which samples could be outliers and then construct statistics using this outlier information while standardizing this result using robust measures of the mean and variation of that particular marker. The statistic allows for the comparison of the set of markers analyzed which signify the strength of the existence of an outlier subset.

A few outlier detection methods have been developed in recent years and many have a similar structure to the t -statistic. Each method varies slightly in how they define the outliers and how they standardize the statistics. The first method by Tomlins et al., Cancer Outlier Profile Analysis (COPA), selects a percentile of the expression levels from the disease group to define outliers which include the most extreme values. The COPA statistic uses this percentile and subtracts the median of the pooled data and is divided by the median absolute deviation of the pooled data [1]. Tibshirani and Hastie built on the theory for detecting differentially expressed cancer genes with the outlier-sum statistic (OS) by gathering more information about the outlier [2]. The threshold for defining an outlier is calculated by the interquartile range (IQR) plus the 75th quartile (q^{75}) of the pooled samples. The statistic adds all outlier samples centered by the median of the pooled samples and then scaled by the median absolute deviation (MAD) of the pooled samples. Wu [3] followed up on this method by developing the Outlier Robust t -statistic (ORT) which is more robust than the OS statistic by using only the normal samples to calculate the median and the MAD. Lian [4] proposed the Maximum Ordered Subset t -statistic (MOST) which removed the definitiveness of the outlier threshold by examining all possible numbers of outliers and determines the optimal size of the outlier subset. For the latest outlier detection method, Wang & Rekaya [5] developed the Least Sum of Ordered Subset Square t -statistic (LSOSS) method. This method also considered every possible size of the subset in the group of interest and the optimal size of the subset is determined by the smallest combined sum of squares from the two subsets created from splitting the disease group.

2.0 BACKGROUND

The criteria for determining performance of these different methods are based on the sensitivity and specificity of the classification of simulated markers into two groups: data with an outlier subset and data without an outlier subset. The classification is dependent on the choice of the threshold and can be altered across the range of statistics when applied to many markers. This information is used to construct receiver operator characteristic (ROC) curve to measure the detection power of these methods. To control for multiple comparisons the FDR is at each threshold.

The number of samples in the outlier subset is bounded from $k = \{0, \dots, n_i\}$ while the magnitude of change in these samples, μ , is restricted to the real number line. The discovery of negative outliers is measured using the same approach as that used for the positive values so μ is further bounded by a positive number line in simulations. A small positive range $\mu \geq 0$ is examined to reflect such deviations that can be expected in real data examples.

The common quality of each of the published simulation studies is that a small set of parameter values were chosen to inspect the qualities of the methods. In each simulation study, the number of samples, n_i , per group were equal in both groups ranging from 15 to 25 per group. The data is simulated from a standard normal distribution while the outlier subset is created by modifying k samples by adding a constant μ . The size of the magnitude difference, μ , in the outlier subset ranged from 1 to 4 but focused on a moderate value $\mu = 2$.

The first simulation study in the outliers sum paper compared the OS statistic, COPA and the t -statistic. These data were simulated with 15 samples per group and 1000 genes from a standard normal distribution of which one was simulated to exhibit an outlier subset. The one simulated gene with an outlier subset was simulated with a μ of 2 and four different values of k , {2, 4, 8, 15}. The detection power was assessed by calculating the percent of simulated normal genes that scored a statistic higher than that of the gene with an outlier subset. The main result was that the OS statistic was best at identifying the single gene for small $k = 2$, and equivalent to the t -statistic for $k = 4$ and t -statistic was superior for $k = \{8, 15\}$. The COPA statistic performed worse than the OS statistic at all values of k .

The simulation studies in the ORT paper choose a larger and equal sample size of 25. One gene out of 1000 genes was simulated to have an outlier subset. This simulation also used a $\mu = 2$ for the quantity added to the data from a standard normal distribution and examined 6 different levels of the number of samples in the outlier subset, $k = \{1, 5, 10, 15, 20, 25\}$. The ROC curves demonstrated the t -statistic has much poorer detection power when $k = \{1, 5\}$ and the OS and COPA have increasingly poorer performance when $k = \{15, 20, 25\}$. Meanwhile, the ORT method has comparable performance to the best methods at each value of k simulated. The FDR rates examined indicate the ORT has the best FDR at the smaller k values and comparable to the t -statistic at the larger values. The main results taken from simulation study is that the ORT has superior performance at any value of k compared to the other outlier detection methods.

In the MOST simulation study, 20 samples per group and 1000 genes were simulated with an outlier subset and 1000 were simulated without an outlier subset. Several different sets of parameter values were simulated, namely, $\mu = 2$, $k = \{1, 5, 10, 15, 20\}$; $\mu = 1$, $k = \{10, 20\}$

and $\mu = 4$, $k = \{1, 2, 5\}$. The MOST statistic has comparable performance to the ORT in each simulation except for when $\mu = 1$, and $\mu = 4$, $k = 1$.

The final simulation study included in the LSOSS article used similar methodologies for measuring each outlier detection methods' performance. In it, 2000 genes were simulated with half of them exhibiting an outlier subset under the parameter values of μ and k . The main examination was done with $\mu = 2$ and various k values $\{2, 5, 10, 15, 20\}$. The authors noted a slight increase to the AUC over all other methods at the point $\mu = 2$ and $k = 10$. The false discovery rates were not examined in this paper.

The results from the four simulations study focused on a small set of points mainly where $\mu = 2$ while varying k across the entire range, $k = \{0, \dots, n_i\}$, from no outlier subset to the entire group as a subset. The focus where $\mu = 2$ is probably the most reasonable value to inspect, but it very reasonable that other performance features can be uncovered. When applying an outlier detection method to a real data set, the underlying parameters are unknown. Unknown underlying parameters values in a real data set require an understanding of the performance of a reasonably large set of parameter values of μ and k . This provides the motivation for a grand review of the set of two-group outlier detection methods. The lack of FDR analysis for the MOST and LSOSS also provides an extra incentive to review these methods.

The extent of simulation studies in the two-group outlier detection has been limited to a very small set of parameters. The studies determine performance of these methods using ROC curves and indicating superiority based on which method has a higher curve. This is analogous to using the AUC resulting from the ROC curves. The lack of inspection into why some methods do not perform well when a set of parameters are chosen provides motivation for an expanded review. A simulation study that adequately illustrates the strengths and weakness of

each method across a wide range of parameters will allow for better confidence of results, because the underlying parameter values of a real data set is unknown. Therefore the goal of an expanded review of a simulation studies to determine detection power and the false discovery rates (FDR) necessary for proper application.

Two-group outlier detection provides an excellent alternative to standard methods for two group comparison of high-dimensional data. The benefit of using outlier detection occurs when it is reasonable to assume an outlier expression pattern exist and due to the handling of multiple comparisons which is done by controlling for false discovery. When the number of comparison groups is expanded with more abnormal groups, two-group outlier detection methods are no longer suitable. Multi-group comparisons are common in such high-dimensional data; for example, comparing cognitively normal samples to samples with mild cognitive impairment (MCI) and samples with Alzheimer's disease would be a realistic target of multi-group outlier detection. A standard method for comparing multiple groups is the F-statistic which is still hampered by multiple comparisons and estimation with few subjects relative to the number of markers.

The comparison of a disease group to a normal group to identify important markers with an outlier subset which is related to the disease is not suitable for data that is categorized into multiple abnormal groups. This provided the motivation to develop outlier detection suitable for multi-group comparisons. Liu and Wu [6] developed two extension suitable multi-group comparisons: the Outlier Robust F-statistic (ORF) and the Outlier F-statistic (OF). The ORF is a direct extension of the ORT method and the OF is a conversion of the F-statistic with robust measures to consider outliers. The ORF derives this extension by assessing the outlier group in each of the abnormal groups separately and combining the results with robust measures of the

entire data. The OF modifies the F-statistic by ignoring samples in the abnormal groups that are not labeled as outliers for the calculation of the statistic.

The number of parameters to simulate in a simulation studies grows by two for every group analyzed. This makes it more difficult to understand the full performance of multi-group outlier detection. However, the Liu and Wu choose to study their methods by simulating a trend in the magnitude of change across three abnormal groups as $\mu_2 = 1$, $\mu_3 = 1.5$, and $\mu_4 = 2$ while varying k . This implied trend is reasonable basis for the parameters which focuses on moderate magnitude. However this is only one configuration to inspect. The magnitude of the outlier subsets could be mutually larger or smaller or varying differences between the groups. The number of samples in the outlier subset for each group, k_g , could also vary.

The simulation study comparing ORF, OF and the F-statistic simulated 1000 genes from the standard normal distribution with 15 samples in each of the four groups. The first group represents a set of the normal samples while the other groups, $g = \{2, 3, 4\}$ represent several classifications of a disease. The disease samples from the first gene are simulated as an outlier by adding a constant μ_g with a probability of π_g to the standard normal random variable. The simulation examined several activation probability $\pi_g = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and one pattern of outlier magnitude, $\mu_2 = 1$, $\mu_3 = 1.5$, and $\mu_4 = 2$. Thus, the expression levels of the disease samples in the first gene is simulated from the following mixture normal distribution: $(1-\pi_g)N(0,1) + \pi_g N(\mu_g,1)$, $g = 2, \dots, G$, and the remaining 999 genes remain as simulated from a standard normal distribution.

The ROC curves indicate that both multi-group outlier detection methods perform much better than the F-statistic when the activation probability, π_g , is 0.5 or smaller. When activation

probabilities are above 0.5, the ORF and F-statistic are comparable while the OF has a lower AUC.

The FDR was analyzed and plotted as the FDR versus the positive rates, the percentage of simulated genes with a test statistic larger than the threshold. The FDR analysis simulated 200 genes with an outlier subset while the remaining 800 were simulated unadjusted from the standard normal distribution. Similar results are obtained with respect to the activation probabilities in the FDR analysis. The ORF and OF have smaller FDR when $\pi_g \leq 0.5$, while the OF performs worse when $\pi_g > 0.5$ relative to the ORF and F-statistic.

This simulation studies presented a different way to inspect the existence of an outlier subset by assigning a probability that a sample will follow an outlier pattern instead of using a fixed number of samples. While not fixing the number of samples reflects situations that occur in real data, it does prevent the true performance to be known at a set of parameter values. The authors choose to analyze one pattern of outlier magnitude across the abnormal samples and varied the activation probabilities to gain perspective on the performance of the first multi-group outlier detection methods. As with the simulation studies of the two-group methods, the number of sets of parameter values inspected with simulations is quite small and does not inform a user about performance across the full range of strengths and weaknesses with unknown underlying parameter values.

Through the derivation and simulation studies, the outlier subset in each group is approached as an outlier of higher magnitude. It is important to consider that each abnormal group could have an outlier subset in either direction further complicating the thorough inspections of all possible sets of parameter values. The issue of directionality is important even with the two-group scenarios, because each statistic searches for an outlier subset in one

direction. Thus the directionality of the statistic must be resolved before being applied to set up the statistic correctly. Changing the directionality to search for outliers of smaller magnitude is made by altering the formula for the outlier region (COPA, OS, ORT, ORF, OF) or simply changing the order of the data (MOST, LSOSS) most likely to be included in the outlier region. Other modification can be made so the statistics is positive.

Multiple group outlier detection serves a great need to extend the two-group methods beneficial for analysis of multi-group comparisons of high-dimensional data with differentially expressed markers. The development of the ORF method demonstrated the way in which two-groups methods can be extended. This serves as a guideline for this research to extend other two-group methods. Also the simulation study can be expanded to uncover further strengths and weaknesses. The development of outlier detection was motivated by issues with microarray data, with similar statistical issues with structural MRI data, this novel application is of great interest for use in the identification of informative brain regions which may affect cognitive health.

3.0 AIMS

In this document, I present a review of the class of two-group outlier detection methods with the aim to expand the current knowledge known about their classification performance. I will also present an adaptation to the LSOSS two-group outlier detection method. The two-group outlier detection methods are applied to structural MRI data, to explore this novel application to real data example. Three different statistical methods are proposed to handle multiple abnormal group comparisons which are extensions of the MOST, LSOSS and the proposed modified LSOSS methods. A simulation study will compare the proposed and existing multi-group outlier detection methods to the standard F-statistic. Lastly, an application to structural MRI data with 3 group trajectories based on neuropsychological tests is used to examine the relationship of volumetric differences and cognitive health in elderly adults.

4.0 TWO-GROUP OUTLIER DETECTION

4.1 ABSTRACT

Standard statistical methods applied to structural MRI data may not identify the most informative brain regions relating to cognitive decline in elderly adults. The heterogeneity of the volumetric data within groups and the high-dimensional aspect of the data suggest that outlier detection methods may be more suitable for discovering informative brain regions. The class of two-group outlier detection methods, Cancer Outlier Profile Analysis (COPA), Outlier Sum (OS), Outlier Robust t -statistic (ORT), Maximum Ordered Subset t -statistic (MOST) and Least Sum of Ordered Subset Square t -statistic (LSOSS), was developed to detect differentially expressed genes that occur in only a subset of the disease group, but is suitable for other high-dimensional data where differential expression patterns occur within a group such structural MRI data. The extent of simulation studies of two-group outlier detection have been limited and a comprehensive review is critical to understanding these methods. A modification of the LSOSS is presented as a result of this review. These methods are applied to a structural MRI data set comparing to search for brain regions which may attribute to cognitive decline in elderly adults.

4.2 INTRODUCTION

Detecting differentially expressed genes is important to microarray studies comparing samples of a normal group to a disease group to find genes which may be related to the disease. The most common approach is to use the t -statistic. Applying the t -statistic on high-dimensional data is complicated by the need to correct for multiple comparisons, and estimating parameters with p variables much larger than n samples. Correcting for multiple comparisons is too conservative to provide any reasonable results and estimating parameters will have large confidence intervals.

It is now known that genes can be differentially expressed in only a small subset of the disease samples which leads to further complications when using the t -statistic. The t -statistic assumes that all samples within a group follow the same distribution and is violated in this scenario. Recently, several statistical methods have been proposed which focus on discovering subsets which are differentially expressed and have found through simulation studies that these methods perform better than the t -statistic.

The first method, Cancer Outlier Profile Analysis (COPA) by Tomlins et al., selects a value as a cutoff to define outliers as the r^{th} percentile of the data or greater. To compute the COPA statistic, the r^{th} percentile of the data from the abnormal group is centered based on the median from the pooled data and then scaled by the median absolute deviation of the pooled data [1]. Tibshirani and Hastie introduced the outlier-sum statistic (OS) which defines an outlier as the interquartile range (IQR) plus the 75th quartile of the pooled samples [2]. Wu presented the Outlier Robust t -statistic (ORT) with robust measures used in the calculation of the statistic [3]. Lian proposed the Maximum Ordered Subset t -statistic (MOST) which takes a novel approach by searching for the best possible number of outliers for each marker by examining all possible subsets of outliers [4]. Wang & Rekaya developed the Least Sum of Ordered Subset Square t -

statistic (LSOSS) method which finds the best way to split the diseased group to identify the outlier subset [5].

The remaining issue with these methods is how to appropriately use these methods to find informative genes that are related to the disease status. Each method used all prior methods in a simulation study compared to the t -statistic and tested these method with a real dataset where they compare their results to a list of known genes that are related to the cancer. The evaluation of the outlier detection methods were limited to simulations that examined only a few parameters values of k , the number of samples in the outlier subset, and μ , the magnitude of increased expression levels of the outliers. The sample of space of k and μ are theoretically bounded from 0 to size of the disease group for k and μ greater than or equal to 0 for μ which can then be flipped to the negative region for under expression scenarios. The maximum number of points in the k and μ sample space used in the simulation studies was 10 leaving the performance of the method in many other areas of the sample space unexplored. Also, the false discovery rates of all two-group outlier detection methods were not studied.

While these methods have been applied extensively in the area of microarray analysis, they have seen little application in the area of neuroimaging. Neuroimaging data presents some of the same challenges as those present in genomics; however these methods have seen little application in this area. Magnetic Resonance imaging data are collected on several hundred brain regions and report the volume for each of these regions. There is interest in both positive and negative outliers in brain volumes which require modifications to the methods to properly search for the correct directional outliers.

In this paper we present an expansive simulation study investigating a much larger set of parameter values than has been presented before. Through the construction of empirical AUC,

the strengths and weakness of each method are highlighted. Based on these results we also present a modification to the LSOSS method that is included in the simulation study. A small set of parameter values of interest are chosen to display the difference in the false discovery rates. An application to a real data example is conducted to compare the results of previous studies.

4.3 STATISTICAL METHODS

Let x_{ijg} be the measurement of the j^{th} marker out of a total of p regions, $j = \{1, \dots, p\}$, for the i^{th} sample ($i = 1, \dots, n_g$) from the g^{th} group, $g = \{1, 2\}$, and $n = \sum_{g=1}^2 n_g$. Group 1 is considered to be the normal group without loss of generality and group 2 denotes the diseased group. We consider the following statistics for identifying differentially expressed markers between two groups.

4.3.1 T-statistic

The t -statistic is defined as follows:

$$t_j = \frac{\bar{x}_{j2} - \bar{x}_{j1}}{s_j},$$

where \bar{x}_{j1} and \bar{x}_{j2} are the mean marker level in the normal and diseased group respectively. The

pooled standard error is estimated as $s_j^2 = \frac{\sum_{g=1}^2 \sum_{i=1}^{n_g} (x_{ijg} - \bar{x}_{.jg})^2}{n_1 + n_2 - 2}$

4.3.2 Cancer Outlier Profile Analysis

The Cancer Outlier Profile Analysis (COPA) statistic is defined as follows:

$$\text{COPA}_j = \frac{q_r(x_{ij2} : 1 \leq i \leq n_2) - \text{med}_j}{\text{mad}_j},$$

where $q_r(\cdot)$ is the r^{th} percentile of the diseased samples. The med_j is the median of the j^{th} marker, $\text{med}_j = \text{median}(\{x_{ij1}\}_{1 \leq i \leq n_1}, \{x_{ij2}\}_{1 \leq i \leq n_2})$, and mad_j is the median absolute deviation of the pooled samples, $\text{mad}_j = 1.4826 \times \text{median}(\{|x_{ij1} - \text{med}_j|_{1 \leq i \leq n_1}, |x_{ij2} - \text{med}_j|_{1 \leq i \leq n_2}\})$.

4.3.3 Outlier Sum statistic

The Outlier Sum (OS) statistic is defined as follows:

$$\text{OS}_j = \frac{\sum_{x_{ij2} \in R_j} (x_{ij2} - \text{med}_j)}{\text{mad}_j},$$

and uses a fixed threshold to identify outliers. The outlier set, R_j , is defined as

$$R_j = \left\{ x_{ij2} : x_{ij2} > q_{75}(\{x_{ij1}\}_{1 \leq i \leq n_1}, \{x_{ij2}\}_{1 \leq i \leq n_2}) + \text{IQR}(\{x_{ij1}\}_{1 \leq i \leq n_1}, \{x_{ij2}\}_{1 \leq i \leq n_2}) \right\}.$$

4.3.4 Outlier Robust t -statistic

The Outlier Robust t -statistic (ORT) is defined as:

$$\text{ORT}_j = \frac{\sum_{x_{ij2} \in R_j} (x_{ij2} - \text{med}_{j1})}{\text{mad}'_j},$$

where med_{j1} is the median of the normal samples, and mad'_j is given by

$mad'_j = 1.4826 \times \text{med} \left(\left\{ \left| x_{ij1} - med_{j1} \right|_{1 \leq i \leq n_1} \right\}, \left\{ \left| x_{ij2} - med_{j2} \right|_{1 \leq i \leq n_2} \right\} \right)$. Note that the ORT improves upon the OS by choosing a slightly different threshold. The outliers are defined relative to the normal group only,

$$R_j = \left\{ x_{ij2} : x_{ij2} > q_{75} \left(\{x_{ij1}\}_{1 \leq i \leq n_1} \right) + \text{IQR} \left(\{x_{ij1}\}_{1 \leq i \leq n_1} \right) \right\}.$$

4.3.5 Maximum Ordered Subset t -statistic

The Maximum Ordered Subset t -statistic (MOST) is defined as:

$$\text{MOST}_j = \max_{1 \leq k \leq n_2} \left(\frac{\sum_{1 \leq i \leq k} (x_{ij2} - med_{j1})}{mad'_j} - \mu_k \right) / \delta_k$$

where μ_k and δ_k are the expected mean and variance from the largest k order statistics from a standard normal distribution. The OS and ORT statistics use *ad hoc* techniques to define the thresholds for flagging outliers. The MOST differs from this by examining all possible sets of outliers using a technique similar to the ORT.

4.3.6 Least Sum of Ordered Subset Square t -statistic

This Least Sum of Ordered Subset Square t -statistic (LSOSS) method identifies a change point in the expression data that would exist if a subset of the group of interest would exhibit differential expression and searches for all possible number outliers. The steps to define LSOSS are described below.

For each marker, the observed values in the diseased group are sorted in descending order and then divided into two groups. This creates two subsets within the diseased group where the

first subset, S_{jk1} , contains the k samples potentially labeled as outliers, $S_{jk1} = \{x_{ij2}: 1 \leq i \leq k\}$; and the second subset S_{jk2} contains samples that would be labeled as non-outliers, $S_{jk2} = \{x_{ij2}: k+1 \leq i \leq n_2\}$. Denote $SS_{S_{jk1}}$ and $SS_{S_{jk2}}$ as the sum of squares of the two partitions. A value of k is chosen by minimizing the total sums of squares across all possible partitions:

$$\arg \min_{1 \leq k \leq n_2 - 1} (SS_{S_{jk1}} + SS_{S_{jk2}}).$$

The pooled standard error for marker j is estimated by $s_j^2 = \frac{s_{j1}^2 + SS_{S_{jk1}} + SS_{S_{jk2}}}{n_1 + n_2 - 2}$, where

$$s_{j1}^2 = \sum_{i=1}^{n_1} (x_{ij1} - \bar{x}_{j1})^2 \text{ is the sum of the squares of the normal group.}$$

The LSOSS is then defined as:

$$\text{LSOSS}_j = k \frac{\bar{x}_{2S_{jk1}} - \bar{x}_{j1}}{s_j},$$

where $\bar{x}_{2S_{jk1}}$ is the mean of the first k samples in descending order.

4.3.7 Modified Least Sum of Ordered Subset Square t -statistic

A modification to the LSOSS is proposed that strengthens the power to detect differences in some areas of the k and μ sample space and sacrifices in another region at the benefit of better false discovery rates. This method will be used in the simulation study that follows. The steps to finding the best threshold for the outlier subset are the same in the original LSOSS method. The final statistic is constructed in a different form by dropping the k . The modified version of the LSOSS is as follows:

$$\text{Modified LSOSS}_j = \frac{\bar{x}_{2S_{jk1}} - \bar{x}_{j1}}{s_j}.$$

4.4 SIMULATION STUDY

4.4.1 Methods to Evaluate Performance

The receiver operating characteristic (ROC) curves have been used to compare the ability of different outlier detection methods to distinguish the differentially expressed (DE) markers from the non-DE markers in previous studies. The determination of superior effectiveness is made by choosing the method with the higher curve on the ROC plot. This is analogous to using the area under the curve (AUC). In this paper, we will use the AUC to measure the detection power and to compare the different methods. In addition, we will also compare the performances of different methods using the false discovery rate (FDR) for the top rank list of a fixed size. Analyzing the FDRs is a valuable tool to fully understand the strength of these methods.

4.4.2 Simulation Analysis

We conducted a comprehensive simulation study to compare the performance of the existing methods. In each simulated dataset, the values of 2000 markers were generated for n_1 normal subjects and n_2 diseased subjects. Among the 2000 markers, only 1000 were differentially distributed between the first k diseased subjects compared to the normal subjects. We conducted the simulation studies with a moderate number of subjects, $n_1 = n_2 = 20$, comparable to the other simulation studies from the outlier detection methods. Following the previous studies, we generate data from normal distributions:

$$x_{ijg} \sim N(\mu_g, 1)$$

$$\text{where } \mu_g = \begin{cases} \mu, & \text{if } g = 2, i \leq k \\ 0, & \text{otherwise} \end{cases}$$

We evaluate the empirical AUC of the ROC curves resulting from the application of these methods to simulated data. Specifically, the empirical AUC was calculated by evaluating the sensitivity and specificity of the statistics by choosing many thresholds at each 5th percentile from 0th to the 100th percentile. The COPA statistic was evaluated using the 90th percentile in all simulations analogous to the previous simulation studies. The simulation is repeated 100 times to obtain the averaged empirical AUC and FDR for each μ and k .

Compared to previous simulation studies, we covered a greater range of parameters. To better understand where these statistics correctly identify informative markers better than the t -statistic, μ is simulated from 0 to 4 by 0.2 and k by all possible integer values from 0 to n_g samples. Lastly, the FDR was analyzed under all parameter settings in the simulation. We choose to display the FDR in a region of the parameter space that exhibit the most significant differences between the methods and of greater interest to idea that a small number of samples are outliers. The FDR was examined at $\mu = \{1, 2, 3, 4\}$, $k = \{2, 3\}$ for each of the seven statistics in this simulation study.

4.4.3 Simulation Results

Figure 1 presents the contour plots of the AUC for ROC curves of each method with 20 samples per group. Similar results were observed for simulations with 10 samples per group (data not shown). The AUC along each axis, either $k = 0$ or $\mu = 0$, represents a coin-flip decision assigning a marker to be positive or negative, since both groups have the same distribution. As

is expected, a trend for the t -statistic shows that if a majority of the samples exhibit the difference in expression level or the size of the difference is large then the AUC will increase. This property is observed in the ORT, MOST and the modified version of the LSOSS. The COPA, OS and LSOSS statistics, however, do not exhibit a non-decreasing function of the AUC. The COPA and OS statistic utilized the median of the pooled samples to center the outlier information which grows much faster with respect to the number of samples in the outlier samples, which in turn reduce the value of the statistic relative to the statistic obtained from a marker without an outlier subset. The LSOSS statistic uses the number of samples, k , in the outlier subset in the statistic. However, in the absence of an outlier subset the k is estimated to be half the sample and not zero which inflates the statistic. This occurs with large μ and small k when markers with true outlier subset estimate k quite accurately, but the statistic is dominated by markers without an outlier subset. This provides motivation to present the modified version of the LSOSS which drops the k from the calculation of this statistic.

To better visualize the performance of each outlier detection statistic in comparison to the t -statistic, we plotted the contour of the differences of the AUC for each outlier detection statistic and the t -statistic in **Figure 2**. Overall, the performance of the t -statistics is fairly robust when the k is reasonably large (i.e. $k > 10$). As expected, the outlier detection statistic performs better when the k is extremely small and the difference is relatively large (i.e. a small portion of the cases exhibit a large differences). However, it is worth noting that when the group differences are moderate to small $\mu \leq 1$, the t -statistic performs marginally better than the outlier detection statistics. Among the outlier detection statistics, ORT, MOST, LSOSS and the modified LSOSS perform much better than the earlier methods (COPA and OS). Overall, the MOST performs the best when compared to the t -statistic. It outperforms the t -statistic even in the bottom right area

of the plot (large k and large μ) and is comparable to the top left area of the plot (large k and small μ).

In **Figure 3**, we display the comparisons between the ORT, MOST, LSOSS and the modified LSOSS methods. In this comparison, we find that the modified LSOSS corrects the main weaknesses of the LSOSS which has a lower and decreasing AUC at the bottom right area of the plot (small k and large μ). The inclusion of k in the statistic inflates the statistic of the normally simulated markers, because without a true outlier the best split is the one where the sample is split in half. It can also be seen that this is the region in which the MOST performs the best and the modified LSOSS and MOST are comparable by each displaying slightly higher AUC at different regions of the plot.

An extension of this simulation examines the strengths and weakness of the outlier detections methods when the proportion of activated regions is reduced from 1000 to 200. The results obtained from these simulations are similar to those shown in Figures 1 and 2 (data not shown).

The FDR of each statistic at various points for $k = \{2, 4\}$ and $\mu = \{1, 2, 3, 4\}$ in **Figure 4**. This region is chosen for two reasons: this is the region which outlier detection methods were developed (small number of samples with a sizable increased expression levels) and the FDR of the methods tended to differ the most. The COPA statistic is the least reliable outlier detection statistic as it is commonly much higher than the other methods which are often clustered together. The LSOSS method has better FDR when μ is small, but it becomes high when the $\mu \geq 3$ and comparable to the t -statistic. The FDR for OS, MOST, modified LSOSS and ORT statistics are comparable to that for the t -statistic when $\mu = 1$ and is better than the t -statistic at larger μ values. This indicates that these methods have increased the power for detection of

differentially expressed markers in the presence of heterogeneity while avoiding a significant increase in the FDR from the t -statistic.

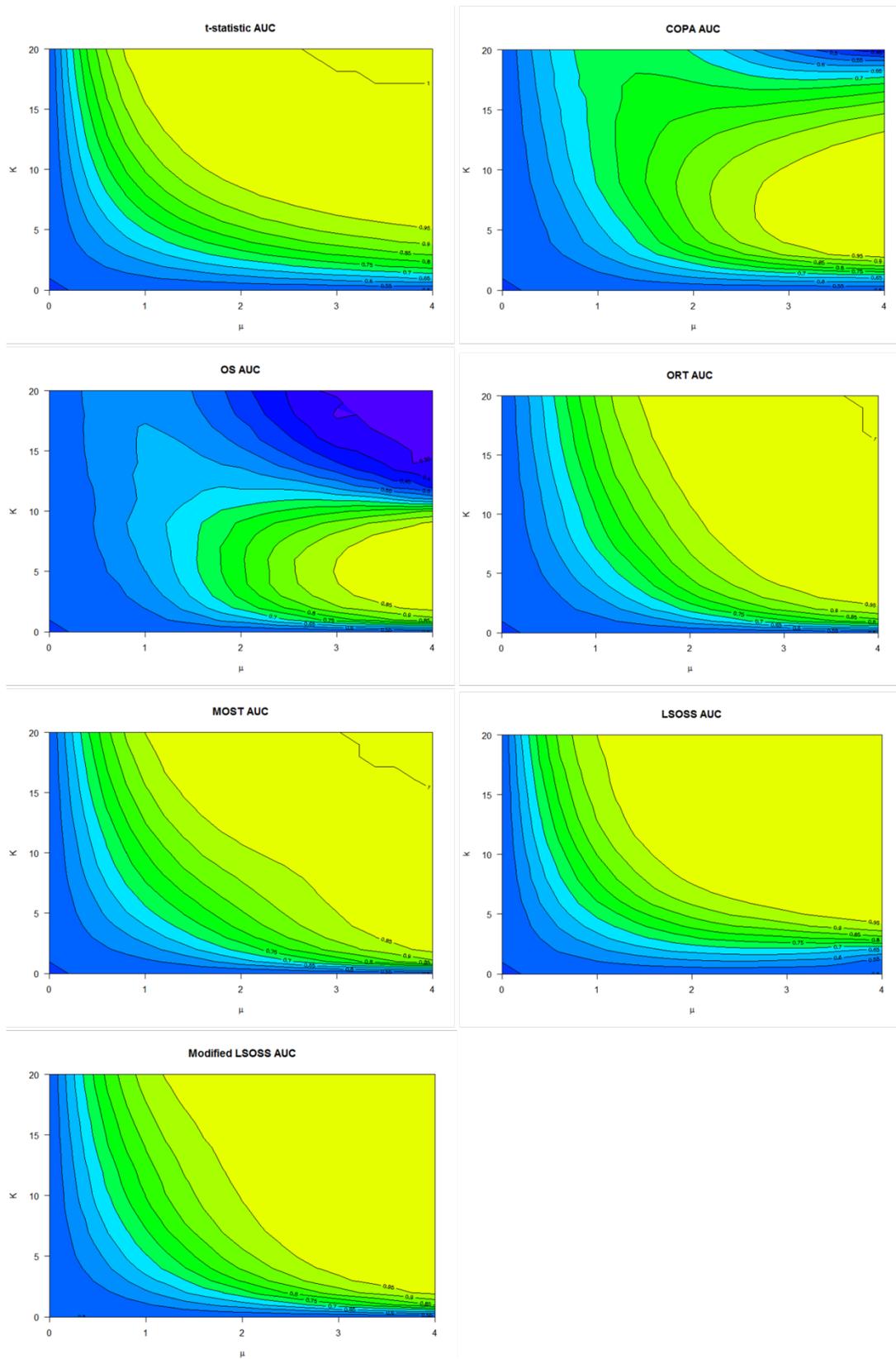


Figure 1. Contour plots of the AUC from each two-group outlier detection method.

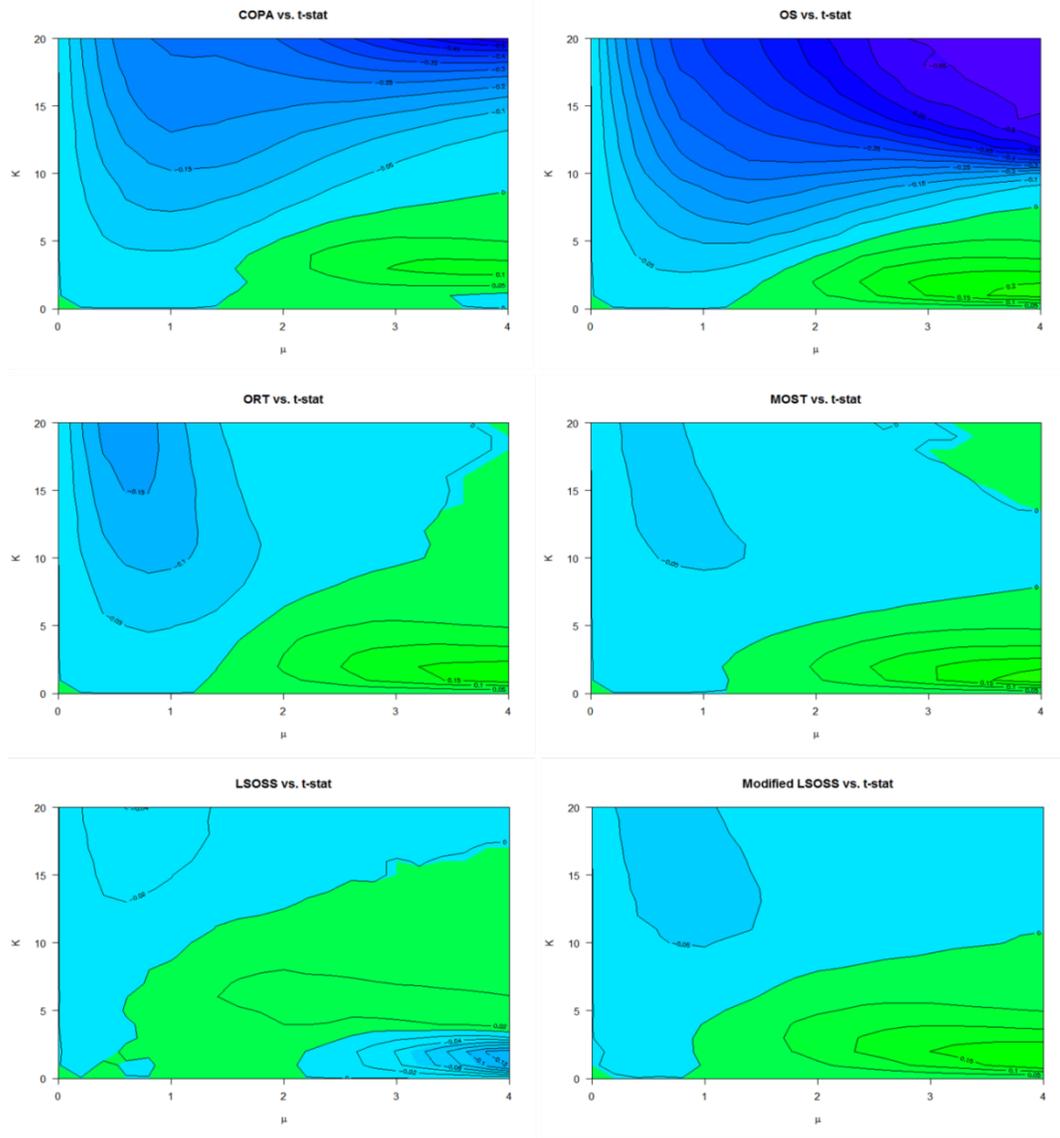


Figure 2. Contour plots of AUC differences between the t -statistic and the outlier detection methods.

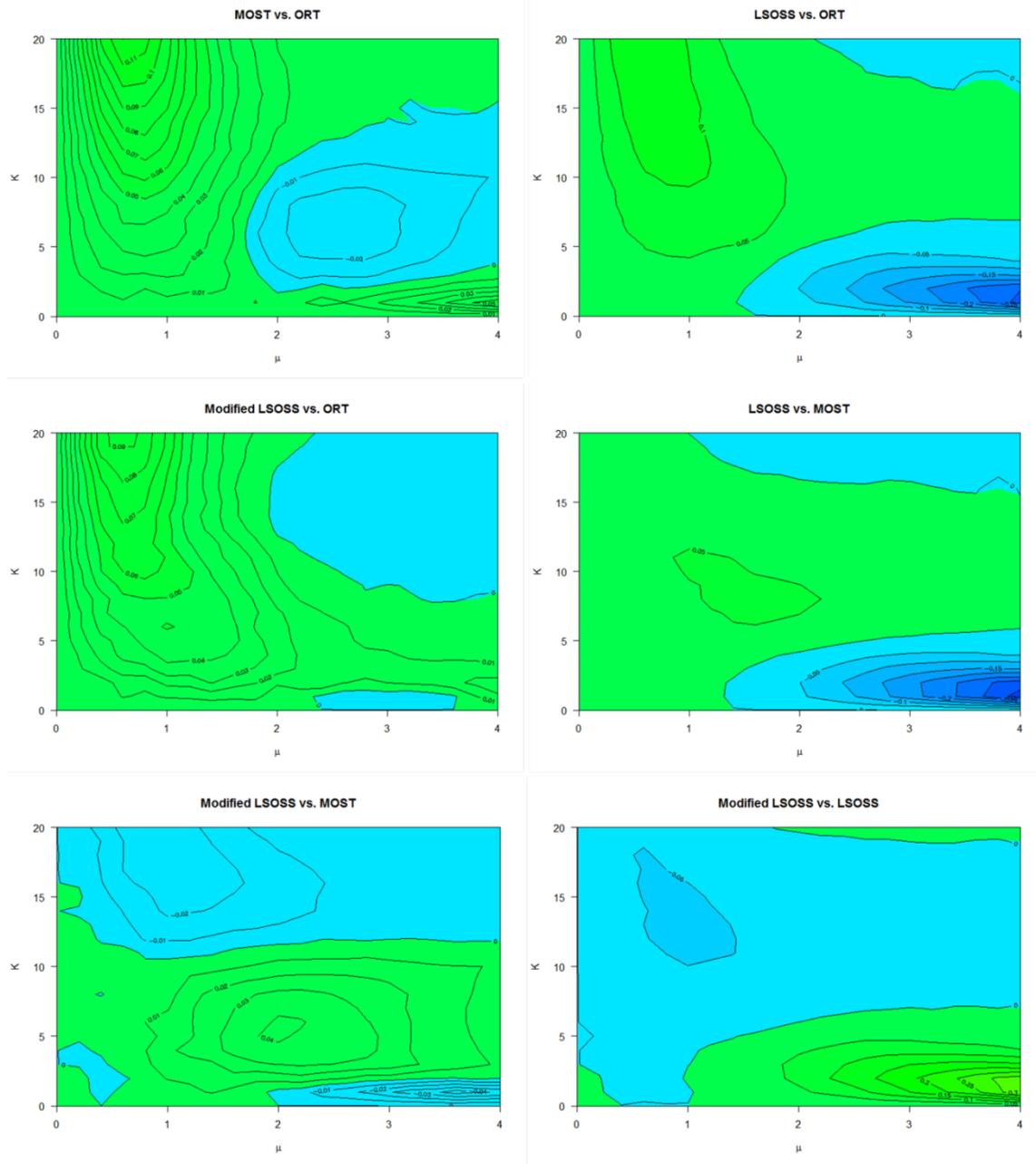


Figure 3. Contour plots of the differences in AUC between the outlier detection methods.

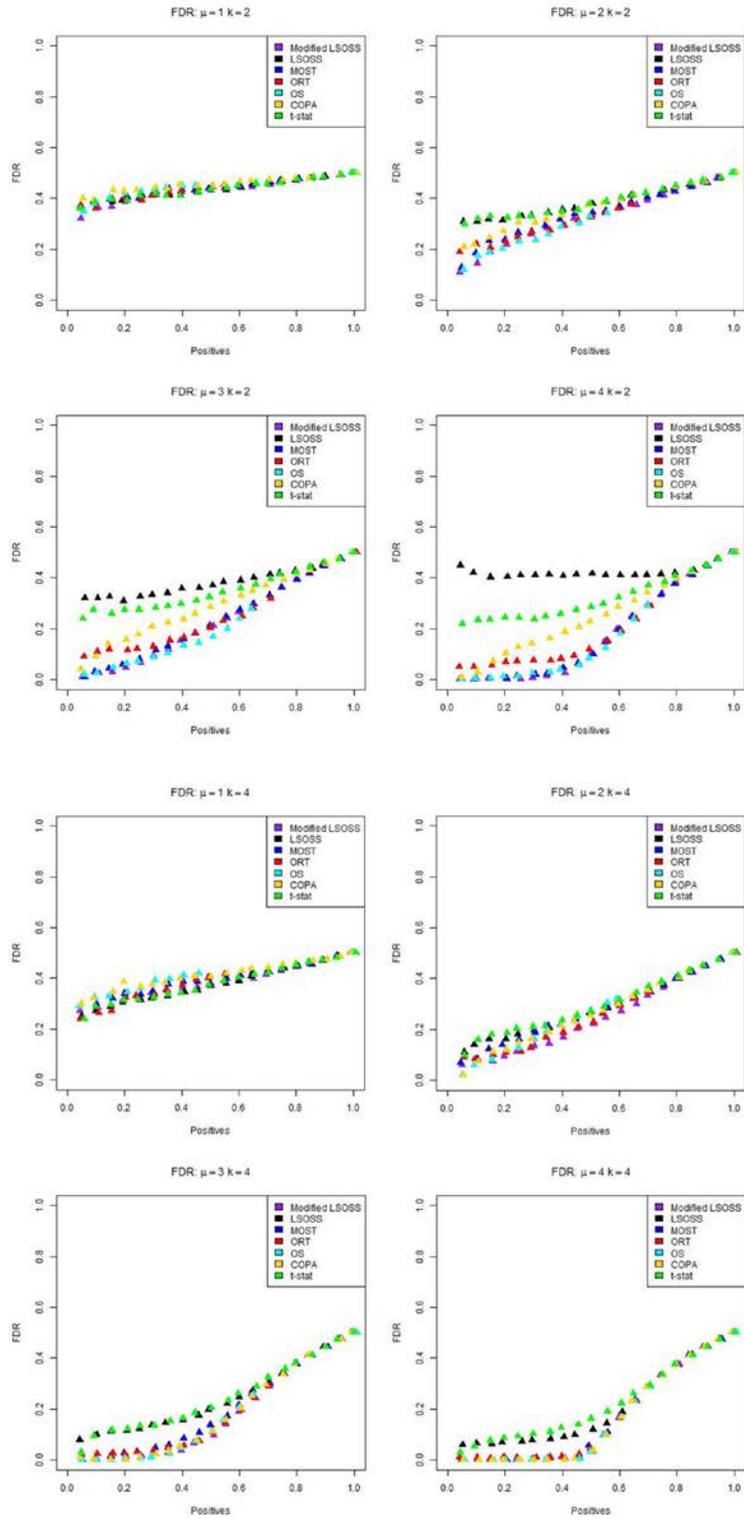


Figure 4. The false discovery rates versus the probability of all regions called significant.

4.5 APPLICATION

Voxel counts of gray matter from nearly 198 structural brain regions were analyzed with the class of multi-group outlier detection methods. A battery of 22 neuropsychological tests was given to 2546 subjects from an elderly adult cohort study at baseline and various time points throughout the multi-year study. A trajectory analysis was applied to the longitudinal data of the neuropsychological test to fit subjects into 3 main groups in which the majority of subjects were classified into a single group denoted as cognitively normal subjects. The remaining subjects were categorized into two other groups which were classified as cognitively superior group and a cognitively declining group. An ancillary biomarker study included a subset of the samples, $N = 372$, of the cohort who underwent MRI scans to measure their structural brain volumes. The data were standardized by dividing the total voxel counts by the intracranial volume (ICV) so that the voxel counts are relative to the volume of the skull. It is thought that structural brain volume may be related to the trajectories of the neuropsychological memory scores in a manner that is similar to the way in which differentially expressed genes are related to disease outcomes. This provides us with the motivation to use outlier detection methods to search for brain regions with outlier subsets resulting in an improvement over standard methods.

We focus on a neuropsychological test of particular interest, a memory test, to narrow the scope of this analysis. Group membership is modeled through the normal model applied to the psychometric data from the data set of 2546 subjects. Three trajectory groups are estimated which allow for a square term to assess the relationship between the times of the test and the test scores. The highest trajectory was compared to the lowest trajectory for this analysis with the lowest trajectory assumed to be the normal group for this comparison. The goal of this analysis

is to identify structural regions of the brain that may be related to a higher trajectory of a neuropsychological test.

The set of two-group outlier detection methods are applied to the data set of 206 subjects who are grouped into either trajectory group. The group with the higher trajectory represents an improving cognitive ability based on improving test scores which is thought to be related to larger brain regions. This requires the outlier detection methods to be structured to identify a positive outlier group. The methods are applied to the MRI data and the regions are sorted from high to low based on the test statistic. The top 15 regions identified by each method are listed in the Table 1 below. Brain regions which were identified in the top 15 regions of at least three methods are highlighted in red. In this analysis there is a large amount of overlap with the ORT, MOST and modified LSOSS have the most common regions with 10, 11 and 10 regions, respectively.

Table 1. Top 10 brain regions identified by each two-group outlier detection method.

T-stat	COPA	OS	ORT	MOST	LSOSS	Modified LSOSS
Caudate R	Brodman Area 48 R	Thalamus L	Thalamus L	Caudate R	Brodman Area 48 R	Brodman Area 41 R
Lingual R	Brodman Area 48 L	Caudate R	Caudate R	Lingual R	Brodman Area 18 R	Thalamus L
Caudate L	Cerebellum Crus2 R	Thalamus R	Thalamus R	Brodman Area 18 R	Brodman Area 37 R	Thalamus R
Brodman Area 18 R	Cerebellum Crus2 L	Brodman Area 48 R	Brodman Area 48 R	Putamen R	Cerebellum 8 L	Caudate R
Calcarine L	Thalamus L	Brodman Area 19 R	Brodman Area 37 R	Thalamus L	Cerebellum 8 R	Brodman Area 48 R
Putamen R	Brodman Area 20 R	Temporal Mid L	Temporal Mid L	Lingual L	Cerebellum Crus2 L	Pallidum R
Lingual L	Brodman Area 7 L	Brodman Area 30 R	Brodman Area 30 R	Calcarine L	Brodman Area 48 L	Brodman Area 18 R
Brodman Area 17 L	Brodman Area 18 R	Brodman Area 21 R	Caudate L	Caudate L	Brodman Area 37 L	Brodman Area 18 L
Paracentral Lobule L	Frontal Mid R	Brodman Area 25 R	Brodman Area 25 R	Brodman Area 30 L	Lingual R	Brodman Area 30 R
Pallidum R	Brodman Area 37 L	Brodman Area 30 L	Brodman Area 30 L	Brodman Area 17 L	Brodman Area 7 L	Brodman Area 11 L
Vermis 7	Temporal Mid L	Caudate L	Brodman Area 38 L	Brodman Area 30 R	Calcarine L	Postcentral R
Cingulum Post L	Brodman Area 18 L	Brodman Area 20 L	Calcarine L	Postcentral R	Cerebellum Crus1 L	Brodman Area 17 L
Pallidum L	Brodman Area 37 R	Brodman Area 11 R	Lingual R	Thalamus R	Brodman Area 19 R	Lingual L
Brodman Area 23 L	Temporal Inf R	Temporal Mid R	Brodman Area 19 R	Brodman Area 25 L	Cerebellum Crus2 R	Lingual R
Cerebellum 4 5 L	Temporal Mid R	Brodman Area 18 R	Cingulum Mid R	Cingulum Mid R	Caudate R	Calcarine L

For further comparisons, the caudate right brain region, which ranked high on a majority of methods, is examined further. The distribution of the voxel counts of the gray matter in the

caudate right brain region is nearly indistinguishable between groups. However, a look into the potential outlier samples that arise from the methods describes how the groups may differ. The manner in which the modified LSOSS and LSOSS define the outlier regions are the same and are represented by LSOSS. A sequence of boxplots in **Figure 5** demonstrates the distribution of the structural MRI data. The first plot examines all samples from each group while two other plots display the distribution of samples identified as outliers by each methods and the non-outliers, respectively, in the high trajectory group.

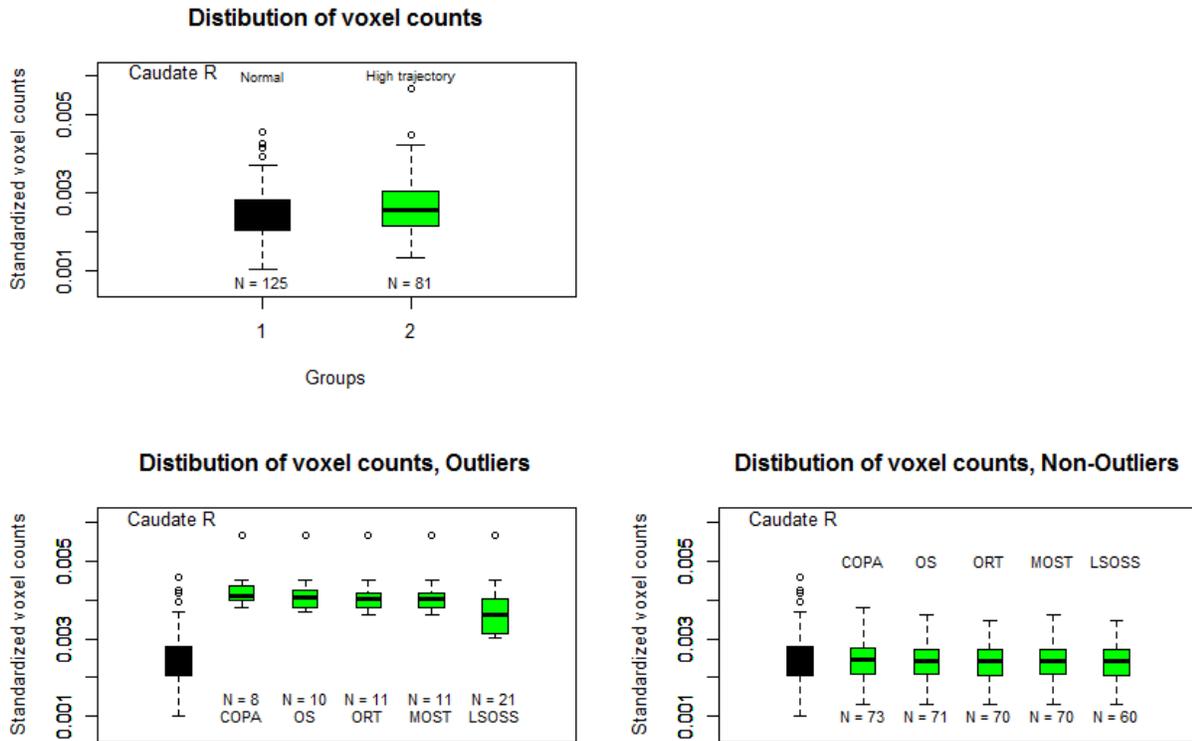


Figure 5. Distribution of voxel counts of caudate right brain region overall and by outlier status.

4.6 DISCUSSION

The detection of genes that are differentially expressed in only a subset of disease samples is of great interest to cancer research. Standard methods for group comparisons do not identify genes well where only a subset of the disease samples are differentially expressed. Outlier detection statistics have been developed that specifically search for these outlier subsets that have proved to be valuable to cancer studies. Although outlier detection was developed for microarray studies, it is suitable for other high dimensional data such as MRI data.

The class of outlier detection statistics was analyzed in simulation studies using different techniques in each simulation. The extent of these simulations only choose a small number of parameter values for the size of the subset, k , and the magnitude of difference, μ , and do not illustrate in any great detail the weaknesses of the methods. This prompted us to perform a comprehensive review of the class of two-group outlier detection so that more known about the statistics performance when being applied to real data. The main goal of our simulations was to simulate many more sets of parameter values to obtain a global understanding of each method. These simulations revealed that the COPA and the OS do not always identify markers with an outlier subset better than a method based on a coin flip, that is, the AUC is below 0.5. Also, the LSOSS has smaller regions where the performance drops which lead to a modification of the LSOSS to have a non-decreasing function of the AUC with respect to the parameter values. This modification produced better overall detection power with minimal loss in other regions. The results of the simulation suggest the MOST and the modified version of the LSOSS has the most favorable detection power while having large advantages over the t -statistic.

The application to structural MRI was a novel approach to the analysis of this high dimensional data with heterogeneous patterns. The goal of this analysis is identify potential

brain regions with volumetric outlier subsets that are related to the performance of the neuropsychological test over time in elderly adults. There was large overlap of the brain regions which scored in the top 15 regions of each of the methods used. However, the ORT, MOST and the modified LSOSS had the most brains that were identified by at least 3 of the methods. This suggests that methods are identifying brain regions with different volumetric patterns between the neuropsychological trajectories. The t -statistic also identified several brain regions in common with the other outlier detection methods which suggest that the number of subjects in the potential outlier subset is large which is more suitable for the t -statistic. The results from outlier detection should provide researchers with a focused set of brain regions to study further.

5.0 MULTI-GROUP OUTLIER DETECTION

5.1 ABSTRACT

Detection of differentially expressed genes has been of great interest in the area of cancer research studies. In recent years researchers have focused on the expression patterns of genes which do not occur homogeneously in all diseased samples. Standard statistical methods do not account for a subset of samples with different distributional expression level within a group well, especially when that subset is small relative to the size of the group. A set of statistical methods have been developed to find these differences in high-dimensional data call outlier detection. The Cancer Outlier Profile Analysis (COPA), Outlier Sum statistic (OS), Outlier Robust t -statistic (ORT), Maximum Outlier Subset t -statistic (MOST), and Least Square of Ordered Subset Square t -statistic (LSOSS) were developed for two-group comparisons between normal and diseased samples. Outlier detection was also expanded to compare differences among multiple disease groups compared to normal samples with the Outlier Robust F-statistic (ORF) and Outlier F-statistic (OF). In this paper, we propose extensions of the MOST and LSOSS methods to the multi-class setting and compare these multi-class methods in a simulation study. This will provide information with respect to the performance of the multi-group outlier detection methods and any advantages that these methods have over standard methods.

5.2 INTRODUCTION

Identifying differentially expressed genes related to disease outcome is a very important topic in microarray studies. Statistical methods have been developed to detect differences in gene expression levels between normal and disease cases based on currently existing two-sample methods. This is achieved by designing a method based on the biological structure of differential expression in a manner similar to that in which neural networks map data via a constructed neural network. Many of these methods rely on the assumption that the expression levels of a gene are expressed homogeneously in all samples of a group. However, it is reasonable to assume that genes related to the disease status are not expressed in the same manner due to heterogeneous patterns and only a subset may be differentially expressed. We expect that standard methods would not perform well considering the fact that the heterogeneous patterns may violate the assumptions that the methods are based on.

A class of statistical methods that has been developed to account for such heterogeneous patterns is referred to as outlier detection. These methods have been shown to better account for heterogeneous patterns. Outlier detection methods are formulated to handle the issue of high-dimensional data, and multiple comparisons while accounting for the within group heterogeneous patterns when comparing across groups. Outlier detection has been extended from two-class comparisons to multi-class comparisons by Liu and Wu with the development of the Outlier Robust F-statistic (ORF) and the Outlier F-statistic (OF) for the detection of outliers in several abnormal groups [6].

These methods have been widely used for the analysis of microarray data, but have seen little use in other applications with similar type of data. We consider the extension and application of these methods to the analysis of neuroimaging data with a focus on MRI data. The

focus is on the comparison across groups for multiple brain regions; however, the outliers may consist of both large and small values with both sets being of interest. The performance of these methods is often examined through a series of simulation studies. In the past, these simulation studies have been limited to only a few parameter values representing the magnitude of the outliers and the number of subjects exhibiting the outlier pattern. As a result of these limitations, the simulations do not give enough information to guide how to best use these methods. Our simulation studies for the two-group outlier detection methods have revealed several weaknesses with each method that should be considered before application.

In this paper we will propose three extensions of the two group outlier detection methods to multiple group comparison scenarios. A simulation study is conducted to effectively measure the strengths and weakness of these proposed methods. In this study the extensions of the ORF and OF multi-group outlier detection methods are compared to the standard multi-group comparative method, the F -statistic.

5.3 STATISTICAL METHODS

The development of outlier detection has been extended to the multiple group comparison, where multiple abnormal groups are compared to the normal group. The Outlier Robust F-statistic and Outlier F-statistic are two recent examples of these methods. The OF statistic is a direct manipulation of the F-statistic while the ORF is a direct extension from the 2-group ORT method. Each of these methods was compared to the F-test. These statistics were tested in simulation study and applied to breast cancer microarray data.

Consider a structural magnetic resonance imaging dataset measuring a large number of structural brain regions comparing a group of cognitively normal subjects to a group with mild cognitive impairment and another group with dementia status. Let x_{ijg} be the measurement of the j^{th} brain region out of p regions, $j = \{1, \dots, p\}$ for the i^{th} sample, $i = \{1, \dots, n_g\}$ from the g^{th} group, $g = \{1, \dots, G\}$, and $n = \sum_{g=1}^G n_g$. Group 1 is considered to be the normal group without loss of generality and all other groups ($g = 2, \dots, G$) are considered diseased or abnormal groups. The following statistics for identifying differentially expressed markers between multiple groups are used in this analysis. The most commonly used multi-group comparison statistic is the F -statistic.

5.3.1 F-statistic

The F-Statistic is defined as follows:

$$F = \frac{\sum_{g=1}^G n_g (\bar{x}_g - \bar{x})^2 / (G-1)}{\sum_{g=1}^G \sum_{i=1}^{n_g} (x_{ig} - \bar{x}_g)^2 / (n-G)},$$

where $\bar{x} = \frac{\sum_{g=1}^G \sum_{i=1}^{n_g} x_{ig}}{n}$, and $\bar{x}_g = \frac{\sum_{i=1}^{n_g} x_{ig}}{n_g}$, and F follows an F distribution with $(G-1)$ and $(n-G)$

degrees of freedom.

5.3.2 Outlier Robust F-statistic

The Outlier Robust F-statistic (ORF) is defined as follows:

$$ORF = \frac{\sum_{g=2}^G \sum_{i \in R_g} (x_{ig} - m_1)}{\text{median} \left\{ |x_{i1} - m_1|_{1 \leq i \leq n_1}, \dots, |x_{iG} - m_G|_{1 \leq i \leq n_G} \right\}},$$

where the region R_g is defined on the normal samples and the group median is used for a robust measure of standard deviation. In addition,

$$R_g = \left\{ i \leq n_g : x_{ig} > q_{75}(x_{i1} : i = 1, \dots, n_1) + IQR(x_{i1} : i = 1, \dots, n_1) \right\},$$

$$m_g = \text{median}(x_{ig} : i = 1, \dots, n_g), \quad g = 1, \dots, G.$$

Note that this statistic reduces to the ORT method when $G = 2$.

5.3.3 Outlier F-statistic

The Outlier F statistic (OF) is an extension of the F -statistic that ignores samples in the abnormal group that were not labeled an outliers based on the region defined above. It is defined as follows:

$$OF = \frac{(\tilde{n} - \tilde{G})}{(\tilde{G} - 1)} \cdot \frac{n_1 (\bar{x}_1 - \bar{x})^2 + \sum_{g=2}^G \tilde{n}_g (\tilde{x}_g - \tilde{x})^2}{\sum_{i=1}^{n_1} (x_{i1} - \bar{x}_1)^2 + \sum_{g=2}^G \sum_{i \in R_g} (\tilde{x}_{ig} - \tilde{x}_g)^2},$$

where $\tilde{n} = n_1 + \sum_{g=2}^G \tilde{n}_g$, and $\tilde{x}_g = \frac{\sum_{i \in R_g} x_{ig}}{\tilde{n}_g}$, $\bar{x} = \frac{n_1 \bar{x}_1 + \sum_{g=2}^G \tilde{n}_g \tilde{x}_g}{n_1 + \sum_{g=2}^G \tilde{n}_g}$, and $\tilde{n}_g = \sum_{i=1}^{n_g} I(x_{ig} > R_g)$.

The formula for the region, R_g , in ORF and OF identifies positive outliers compared to the normal group. This inherently describes directionality of the outlier detection methods that must be specified prior to the analysis for the discovery of outliers. These regions can be flipped to coincide with a hypothesized deviation of a subset of the abnormal group.

5.4 PROPOSED METHODS

In this section we propose the multi-group LSOSS and multi-group MOST methods which are novel extensions of the two-group comparisons counterpart. These extensions are achieved by applying the same steps from the two-group methods which identify outliers in the abnormal groups and bring together the results to form the final statistic.

5.4.1 Multi-Group MOST

The proposed MG-MOST method involves the comparison of multiple values derived from the use of different values of k in each group. This method requires that the data be sorted in descending order to identify higher magnitude outliers. The step to calculate the statistic is described in the following steps:

$$M_{kg} = \left(\frac{\sum_{1 \leq i \leq k_g} (x_{ig} - med_1)}{1.4286 \cdot med \left[|x_{i1} - med_1|_{1 \leq i \leq n_1}, |x_{ig} - med_g|_{1 \leq i \leq n_G} \right]} - \mu_{kg} \right) / \sigma_{kg},$$

where μ_{k_g} and σ_{k_g} are the expected mean and variance from the largest k order statistics from a standard normal distribution from the g^{th} group, $g = \{2, \dots, G\}$. To search for outliers with smaller values μ_{k_g} and σ_{k_g} , change to the expected mean and variance use the k smallest order statistics from a standard normal distribution.

The solution for the best choice of k_g is obtained from each group separately by the following argument:

$$M_g = \arg \max_{1 \leq k_g \leq n_g} M_{k_g} .$$

The final step is to aggregate the group statistics as follows:

$$\text{MG-MOST} = \sum_{g=2}^G M_g .$$

5.4.2 Multi-Group LSOSS

The proposed multi-group LSOSS (MG-LSOSS) method also extends the LSOSS two-group method by first finding the best outlier threshold in each group. The data need to be sorted in descending order to identify higher magnitude outliers within each group from $g = \{2, \dots, G\}$.

The following steps describe how the method is calculated. First define

$$S_{gk_1} = \{x_{ig} : 1 < i < k\} ,$$

$$S_{gk_2} = \{x_{ig} : k+1 < i < n_g\} .$$

Thus, each abnormal group produces two subsets with the first subset, S_{gk_1} , containing samples labeled as outliers. The second subset, S_{gk_2} , contains samples that are not considered outliers and are similar to the distribution of the normal samples. Sorting the voxel counts in

ascending order is required to search for outliers of smaller magnitude. The mean and the sum of squares are calculated for each subset as follows:

$$\bar{x}_{S_{gk_1}} = \sum_{i=1}^k x_{ig} / k,$$

$$\bar{x}_{S_{gk_2}} = \sum_{i=k+1}^{n_g} x_{ig} / (n_g - k),$$

$$SS_{S_{gk_1}} = \sum_{1 \leq i \leq k} (x_{ig} - \bar{x}_{S_{gk_1}})^2,$$

$$SS_{S_{gk_2}} = \sum_{k+1 \leq i \leq n_g} (x_{ig} - \bar{x}_{S_{gk_2}})^2.$$

Next, the optimal value of k is chosen for each group by obtaining the solution to the following argument within each group:

$$\arg \min_{1 \leq k \leq n_g} (SS_{S_{gk_1}} + SS_{S_{gk_2}}).$$

Let S_1^2 be the sum of squares from the normal group. The estimate of the pooled standard error is defined as follows:

$$s^2 = \frac{s_1^2 + \sum_{g=2}^G (SS_{S_{gk_1}} + SS_{S_{gk_2}})}{n - G}.$$

The multi-group LSOSS is defined by:

$$\text{MG-LSOSS} = \frac{\sum_{g=2}^G k_g [\bar{x}_{S_{gk_1}} - \bar{x}_1]}{s}.$$

In this way, a search for an outlier subset in each group is completed independent of one another. If a particular group is hypothesized to have outliers of smaller magnitude, the quantity added in the final step can be changed to the following:

$$(n_g - k_g) [\bar{x}_{S_{gk_2}} - \bar{x}_1].$$

This allows the best choice of k_g to define outliers for every group independent of each other. A summation of the resulting terms from each group is used for the calculation of the statistic.

5.4.3 Modified Multi-Group LSOSS

A modification to the original LSOSS method is found to be important to maintain a non-decreasing function of the AUC with respect to μ_g and k_g . The LSOSS statistic utilizes the estimated size of the outlier subset in the statistic. A simulation tallied the estimated k_g from simulated markers with no outlier subset and found that k_g was centered around half of the sample size of the group and not zero. This negatively affects the AUC of this method in the region where k_g is very small and μ_g is very large. This happens because in the presence of a single large outlier sample, the method accurately estimates a small k_g , but estimates a much larger k_g when no outlier exists. The end result is the reversal in relative rank of the statistic. This modified version of the MG-LSOSS is presented below.

$$\text{Modified MG-LSOSS} = \frac{\sum_{g=2}^G [\bar{x}_{S_{1k_g}} - \bar{x}_1]}{S}.$$

5.5 SIMULATION STUDY

5.5.1 Simulation Setup

A simulation study was conducted to compare the power and false discovery rates of the proposed multi-group LSOSS, the multi-group MOST and the existing methods, ORF, OF and the F-statistic. The performance of these methods is determined by the AUC resulting from the detection of the heterogeneous patterns across the abnormal groups. The FDR will be examined at various points of interest to understand how the probabilities of false discoveries differ between methods. This simulation will greatly expand the information known about these methods from previous simulation studies by examining many more points across the (μ_g, k_g) sample space.

The simulation was conducted with three groups with equal samples sizes of 20 samples per group. Without loss of generality, the first group will be considered the normal group and all other groups are defined by a varying degree of abnormality. The standard normal distribution was used to simulate 2000 voxels regions of the brain for all samples. In each abnormal group, $g = \{2, 3\}$, a constant μ_g is added to k_g samples in the first 1000 simulated voxel regions to resemble outliers of larger magnitude. The remaining 1000 regions are not altered and resemble a homogeneous voxel region that does not contribute to the disease status of the sample. For simplicity, the search space is restricted to identical patterns of outliers across the abnormal groups. Thus the size of the subset, k_g , and the size of volumetric differences, μ_g , are fixed to be equal for each group. These parameters are iterated over a large range of μ and k , $\mu = \{0, 0.2, 0.4, \dots, 4\}$ and $k = \{0, 1, 2, \dots, 20\}$. Each statistical method is applied to each simulated region and the sensitivity and specificity are measured at each 5th percentile of the range of the pooled

statistics from the 0th to 100th percentile. The empirical AUC is calculated and plotted in contour plots. The FDR were analyzed at selected points of μ and k to display how the FDR changes as the percentage of regions declared to be positive for an outlier group is increased.

5.5.2 Simulation Results

Figure 6 shows contour plots of the AUC for each multiple group outlier detection method and the F-statistic. The deficiency of the F-statistic is easily seen where the number of samples with volumetric differences is low, $k > 5$, or when the magnitude of volumetric difference is small, $\mu < 1$. The ORF and OF display a large improvement over the F-statistic mainly in the region where k is small. The ORF has better AUC than the OF where $k = 1$. The MG-MOST performs equally well at low k values with high μ , but this statistic is slightly better when μ is small, $\mu < 1$. The MG-LSOSS also performs better than previous methods when μ is small. However, this method fails to correctly classify samples when a small number of samples exhibit a very large volumetric difference. This is likely due to the manner in which the value of k is chosen which is based on the sum of squares.

Figure 7 displays the differences in the AUC between the multi-group outlier detection methods versus the F-statistic. This highlights that each method has a better detection power than the F-statistic when the size of the outlier subset is about half the size of the group or less in each group. The noted exception is the area where the detection power of the MG-LSOSS method begins to decrease. This occurs around where $k = 1$ or 2 and μ is larger than 3 . The modified MG-LSOSS was constructed to correct the overestimation of the statistic for simulated data with no outliers which results in poorer detection power for this region of the μ and k sample space. **Figure 8** demonstrates the differences among the proposed methods and the ORF.

The modified MG-LSOSS is shown to correct the poor detection power of the MG-LSOSS in the problematic region. The MG-MOST has the best AUC in areas of the μ and k sample space where the signal of outliers is weak, that is, regions where μ or k is very small.

The false discovery rates were analyzed at each point included in the simulation and a few parameter values for μ and k were chosen in a region of interest where the FDR differ. In **Figure 9**, the false discovery rates by percent of simulated regions declared positive is shown where $k = 2, 4$ and $\mu = 1, 2, 3, 4$. The F-statistic and the MG-LSOSS have consistently larger false discovery rates when large μ 's are simulated compared to the other methods. The ORF, MG-MOST and modified MG-LSOSS have comparable and the lowest FDR among the region of the sample space shown. The differences in the FDR is negligible for the majority of the other simulated parameter values of the samples space in this simulation.

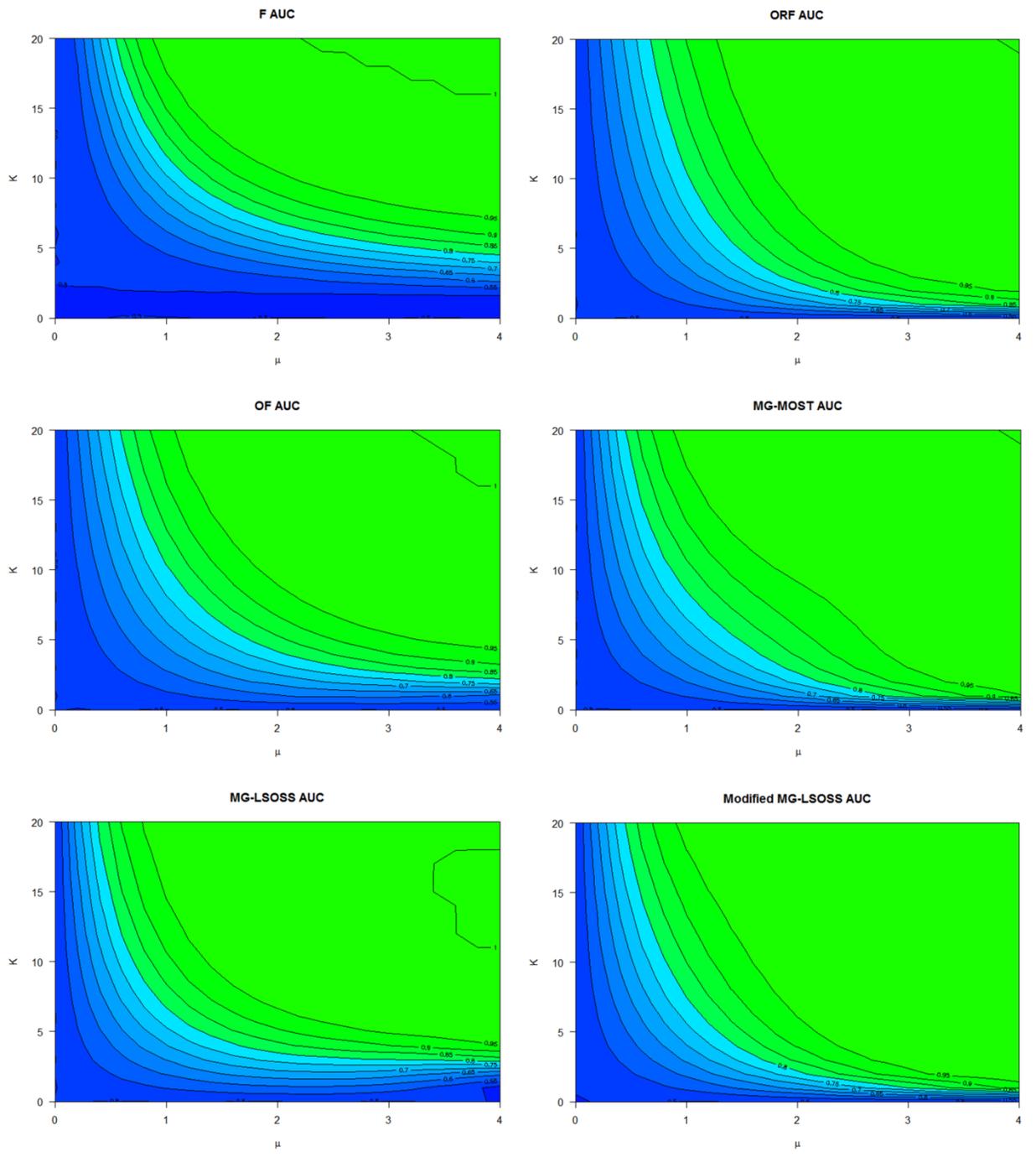


Figure 6. Contour plots of the AUC from each multi-group outlier detection method.

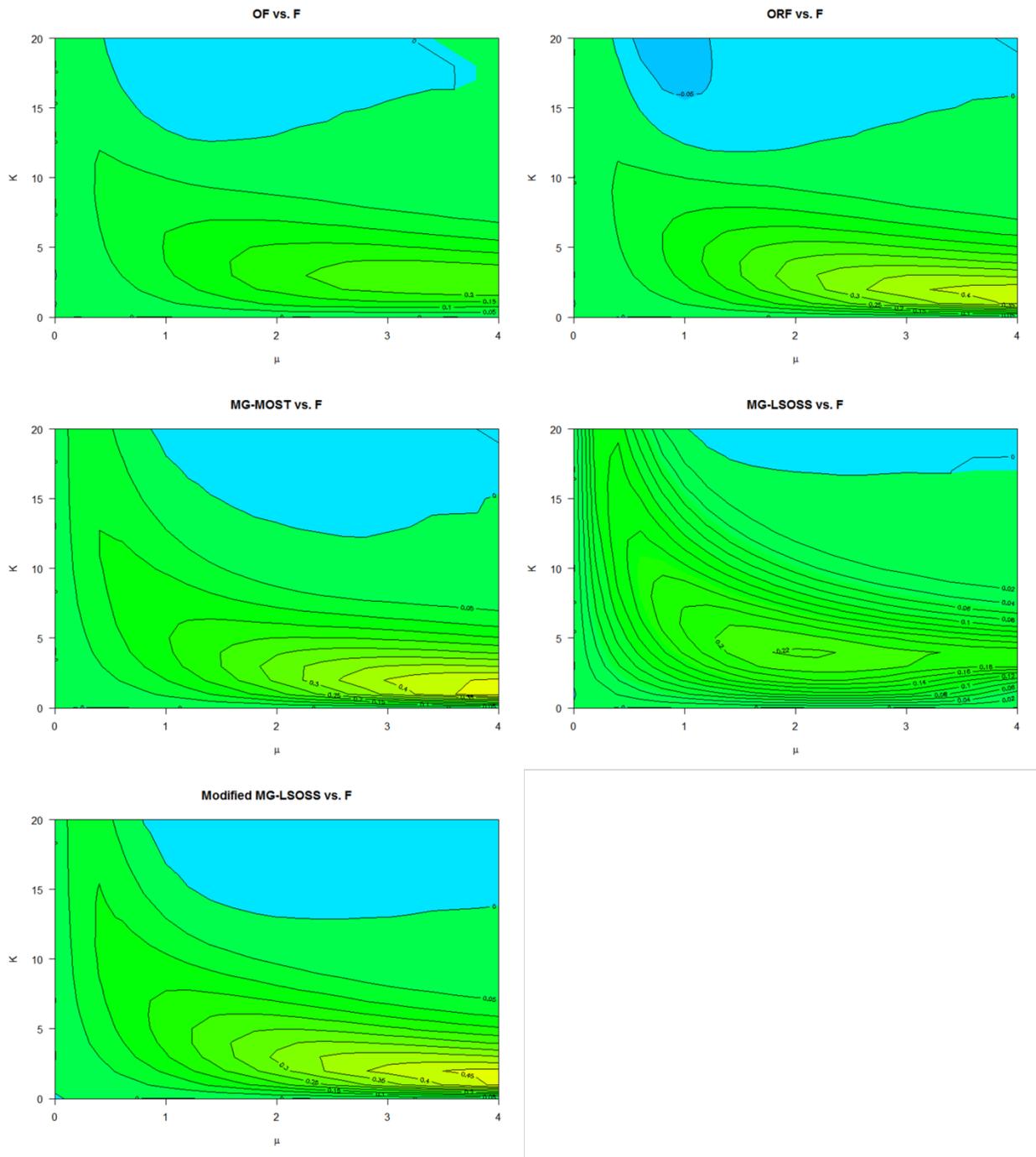


Figure 7. Contour plots of the differences in AUC between the outlier detection methods and the F -statistic.

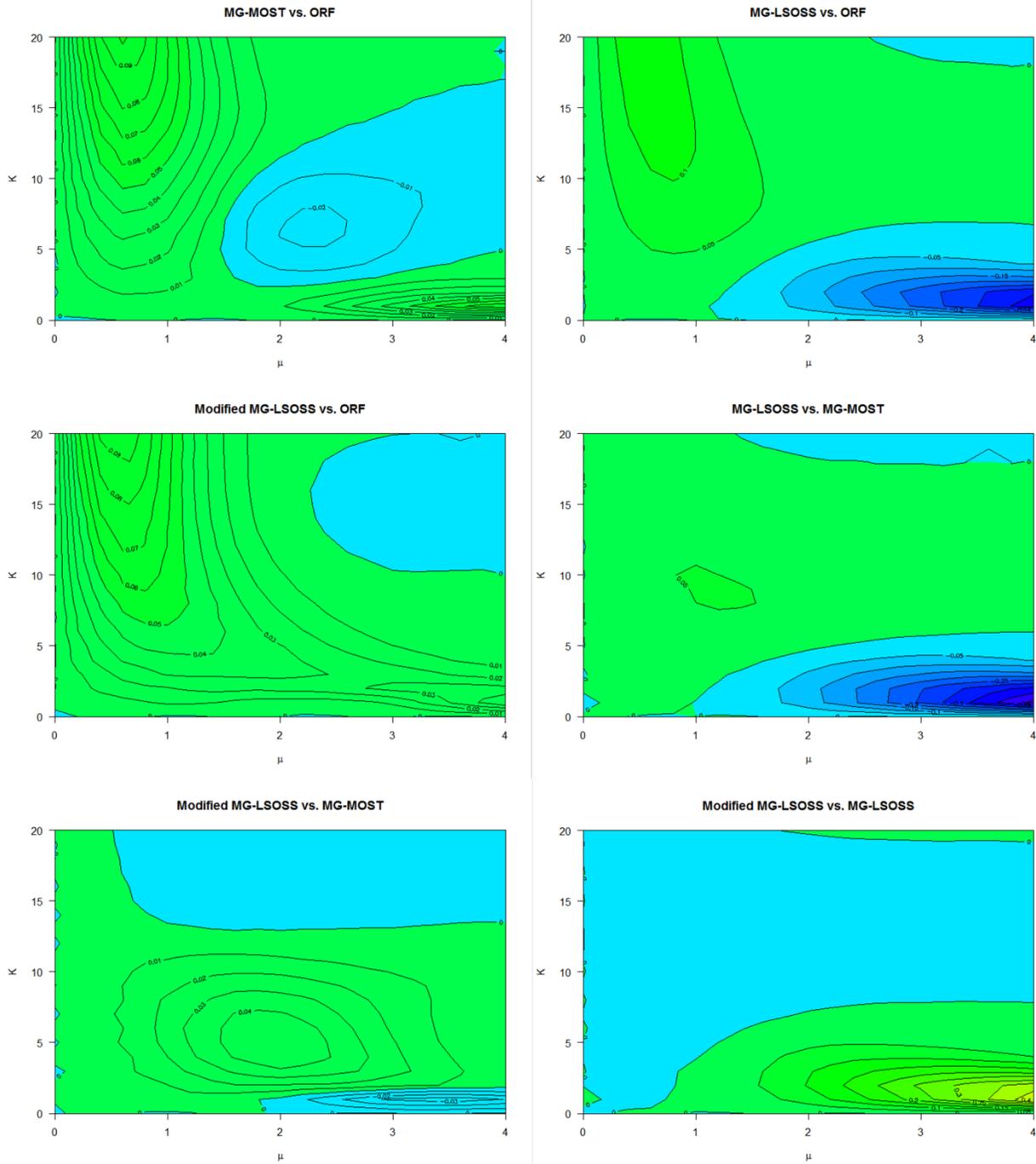


Figure 8. Contour plots displaying the difference in AUC in comparisons between the multi-group outlier detection methods.

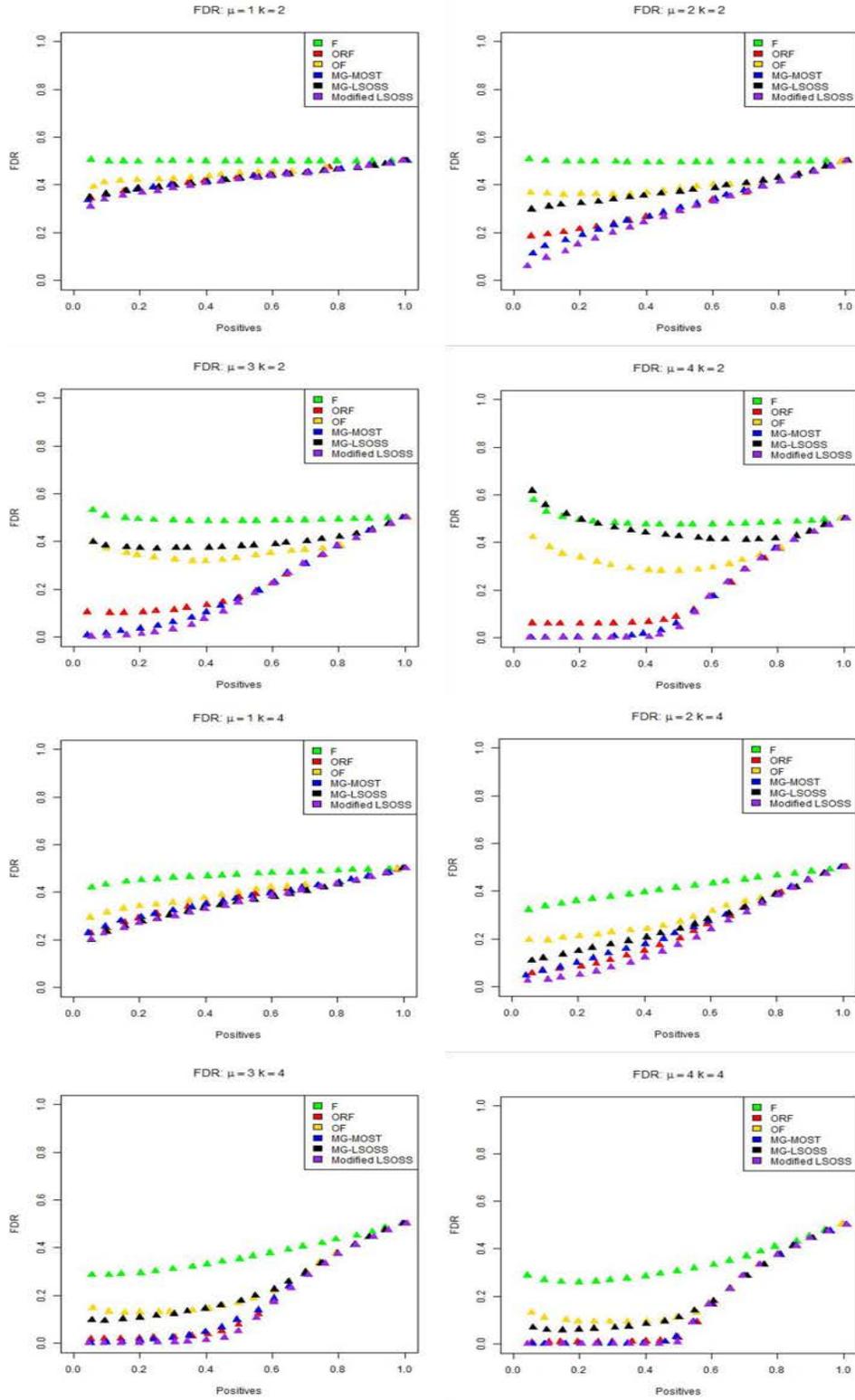


Figure 9. False discovery rates of the two-group outlier detection methods.

5.6 APPLICATION

Voxel counts of gray matter from 198 structural brain regions were analyzed with the class of multi-group outlier detection methods. A battery of 22 neuropsychological tests was given to 2546 subjects from an elderly adult cohort study at baseline and annually after 3 to 4 years after entrance into the study. A trajectory analysis was applied to the longitudinal data to group subjects into 3 main groups in which the majority of subjects were classified into a single group denoted as cognitively normal subjects. The remaining subjects were classified into two separate groups which were classified as a cognitively superior group and a cognitively inferior group. These three groups are based on their trajectories of the repeated neuropsychological test and will form the multiple groups for comparisons. A biomarker sub-study, $N = 372$, was conducted which subjects underwent MRI scans to measure their structural brain volumes. The data were standardized by the dividing the voxel counts by the intracranial volume (ICV) so that the voxel counts are relative to the volume of the skull. The hypothesis is that the structural brain size may reveal similar patterns to differentially expressed genes. If such a pattern is observed, it would provide a possible mechanism to understand how volumetric brain volume may impact the results of neurological test. The advantage of using outlier detection that it is better suited to identify these regions than standard methods for group comparison if outlier subsets exist in an abnormal group.

We focus on a memory neuropsychological test that is of particular interest to narrow the scope of this analysis. Group membership is determined by modeling the trajectories of the longitudinal psychometric data from the 2546 subjects in this study. Three groups are estimated which includes a square term to assess the relationship between time and the test scores shown in Figure 10. The normal group is chosen to be the group with a flat trajectory which happens to

contain the most subjects. The other group trajectories have higher and lower scores relative to baseline. The goal of this analysis is to identify structural regions of the brain that may impact the trajectories of a neuropsychological test.

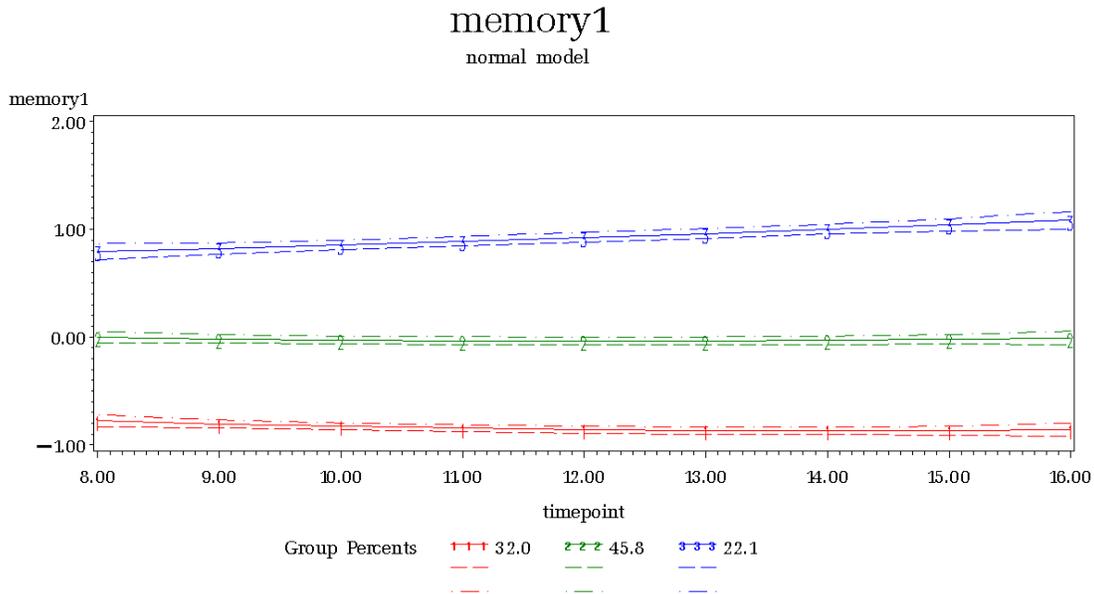


Figure 10. Trajectory plot of 3 groups modeled using the neuropsychological memory test scores.

The set of multi-group outlier detection methods are applied to subject with MRI data of 372 subjects. The group with the higher trajectory represents an improving cognitive ability which is thought to be a result of larger brain regions and the lower trajectory to represent a declining cognitive ability. We would expect positive outliers in the higher trajectory group and negative outliers in the lower trajectory groups in accordance to the hypothesis that larger volumes would indicate better cognitive ability and smaller volumes a reflection of poorer cognitive ability. This requires the outlier detection methods to be structured to search outliers in the expected direction. The brain regions are then sorted from high to low based on the test scores with higher ranked regions to be more likely to contain outlier groups in the abnormal groups. The top 15 regions identified by each method are listed in Table 2 below. Any region which appears in the top 15 of any of the methods at least three times is highlighted in red to

illustrate brain regions which are brain regions are consistently identified as having a significant outlier pattern.

Table 2. Top 15 brain regions identified by each multi-group outlier detection method

F	OF	ORF	MG-MOST	MG-LSOSS	Modified MG-LSOSS
Brodman Area 28 R	Cerebellum 10 R	Cerebellum 10 R	Cerebellum 10 R	Caudate L	Vermis 10
Amygdala R	Thalamus L	Caudate R	Pallidum L	Cerebellum Crus2 L	Brodman Area 41 R
Brodman Area 34 R	Thalamus R	Brodman Area 25 L	Brodman Area 41 R	Putamen R	Pallidum R
Frontal Latx Orb R	Frontal Sup Orb L	Lingual L	Cerebellum 10 L	Cerebellum 8 L	Cerebellum 10 R
Hippocampus R	ROI1	Brodman Area 27 L	Pallidum R	Cerebellum 8 R	Pallidum L
ROI1	Brodman Area 29 L	Brodman Area 25 R	Olfactory R	Brodman Area 7 L	Cerebellum 10 L
Temporal Pole Mid R	Cingulum Mid R	Cerebellum 10 L	Vermis 10	Caudate R	Thalamus L
Brodman Area 28 L	Brodman Area 11 L	Thalamus L	Thalamus R	Cingulum Post L	Thalamus R
Hippocampus L	Pallidum L	Thalamus R	Brodman Area 27 L	Brodman Area 18 R	Brodman Area 27 R
Frontal Sup Orb R	Vermis 1 2	Olfactory R	Brodman Area 25 R	Lingual R	Caudate R
Brodman Area 36 R	Pallidum R	Brodman Area 47 R	Thalamus L	Brodman Area 7 R	Brodman Area 29 L
Brodman Area 38 R	Brodman Area 41 R	Brodman Area 27 R	Precuneus L	Parietal Sup L	Brodman Area 27 L
Temporal Pole Sup R	Brodman Area 30 L	Pallidum L	Olfactory L	Angular L	Brodman Area 25 R
Parahippocampal R	Putamen R	Brodman Area 30 R	Brodman Area 27 R	Cerebellum Crus2 R	Olfactory R
Frontal Sup Orb L	Brodman Area 32 R	Brodman Area 18 L	Caudate R	Parietal Sup R	Olfactory L

For further comparisons, the Brodman Area 28 right brain region is a region which is located the memory area of the brain which would likely affect the neuropsychology test which the trajectories are modeled from. The distribution of the voxel counts of the gray matter in the Brodman Area 28 in the right hemisphere is nearly undistinguishable between groups. However, a closer look into the potential outlier samples that arise from the methods describes how the groups differ. The manner which the outlier region is determined is not different for each outlier detection method. The ORF and OF method define outliers by the formula and are characterized by ORF in the figures below. The MG-LSOSS and the modified MG-LSOSS each determine the number of samples in the outlier subset in the same manner and are represented by the MG-LSOSS below.

A series of boxplots in Figure 11 demonstrate the distribution of the structural MRI. The first plot examines all samples from each group while two other plots display the distribution of predicted outlier samples and non-outlier samples, respectively, within the abnormal groups.

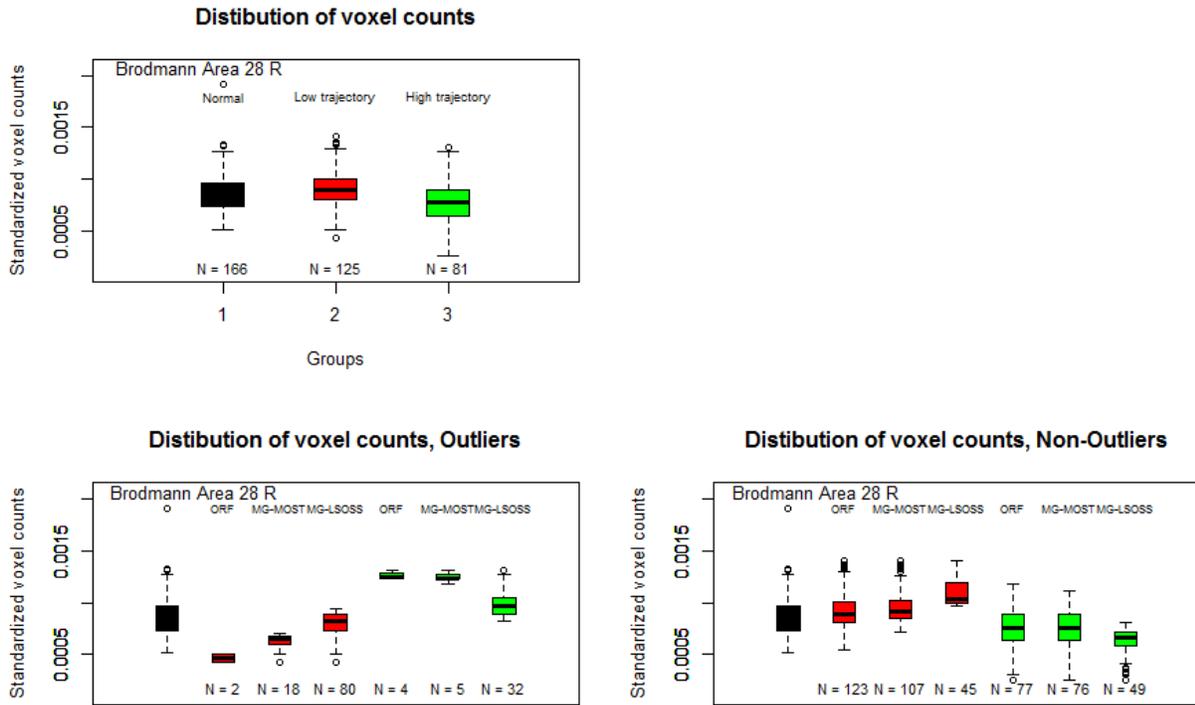


Figure 11. Distribution of voxel counts of Brodmann Area 28 right brain region overall, and by outlier status.

5.7 DISCUSSION

The detection of heterogeneous patterns in cancer samples is of particular interest to researchers. The heterogeneous patterns are sometimes due to differentially expressed genes in which a subset of the cancer samples has higher expression levels. The detection of a subset of outlier

samples provides the motivation to develop methods that would specifically search for the outliers in a group of disease samples. Through simulations, outlier detection has provided several reasonable methods which can identify simulated outlier patterns better than standard methods when comparing normal samples to disease samples.

There is also a need to detect the outlier patterns that can exist across several groups. The obvious progression was to extend the existing two-group outlier detection methods to handle multiple abnormal groups. The ORT method was the first method extended to handle multiple groups. The MOST and LSOSS and the proposed modified LSOSS methods have better detection power than ORT at various parameter values in simulation and are a clear choice to develop extensions. The extended version of these methods provided similar benefits that are observed in the two-group simulations and are superior to the F-statistic.

The simulation that was performed is limited by the set of parameters that were chosen to analyze the detection power. The parameter values chosen for this simulation was optimal for these methods to detect the correct marker with outlier subsets. However, there is a many other outlier scenarios where the methods may not be able to detect the simulated outlier groups with as good of precision.

The application to structural MRI data is a novel application that is suitable for outlier detection. There exist heterogeneous patterns in the volumetric data that present a challenge to researchers to find meaningful relationship between brain region size and the development of cognitive decline. The OF, ORF and MG-MOST each identified the cerebellum 10 right brain region as the top ranked brain region. The OF, ORF, MG-MOST and the modified MG-LSOSS had consistency with the top 15 brain regions identified. Examining the top 15 brain regions from each method, any region identified by at least three methods was highlighted of which there

was a total 12 brain regions highlighted. The OF, ORF, MG-MOST and the modified MG-LSOSS identified 6, 10, 12 & 12 of the highlighted regions respectively. The consistency between the methods is unlikely to be observed if an outlier pattern did not exist. The F statistic did not have any of the 12 highlighted regions while the MG-LSOSS only identified one region of these particular regions. There was no region which was in the top 15 brain in each of the 6 statistics in this applied example.

The F-statistic ranked the Brodmann Area 28 (BA 28) right as the top brain region, which is related to declarative memory. However, the F-statistic does address directionality of the outliers. This result indicates the BA 28 right region had the most difference between the groups and not an indication of the specified direction of outlier samples.

6.0 DISCUSSION

Outlier detection is an excellent method to identify biomarkers where a subset of the samples have larger or smaller expression pattern compared to normal samples. The methods are suitable to analyze high dimensional data for normal versus disease sample comparisons, i.e. two-group outlier detection, or comparisons involving multiple disease groups, i.e. multi-group outlier detection. The methods serve as an initial search for informative markers when standard methods of group comparisons do not provide significant results.

The review of the class of two-group outlier detection methods revealed much about the methods including regions of the parameter space of μ and k which do not work as desired. Notably, the COPA and the OS statistics fail to identify makers when more than half of samples in the group of interest are outliers. As a result of this analysis we presented a simple modification to the LSOSS method that corrects for an overestimation of outlier information when no outliers are present. The modified LSOSS, MOST and ORT each have the desirable quality of a non-decreasing function of the AUC with respect to μ and k while providing improved detection power across the range parameters values. These three methods are the best choice for an effective outlier detection method to apply to real data where the underlying parameters are unknown.

The desire to identify outlier subsets across multiple abnormal groups led to extending several two-group methods for this scenario. The MG-MOST, MG-LSOSS and the modified

MG-LSOSS were presented after the ORT has been extended to multi-group scenario in ORF statistic. Each of these methods reduces to their respective two-group statistic when there is only one abnormal group. The concept of differing direction of outlier subsets was introduced in the application to the MRI data. The hypothesized direction of an informative outlier subset would have to correspond to the performance of the trajectory group which required manipulation from how the methods were presented. The manipulation made for each method to detect negative outliers was made positive so that the outlier information between positive outliers and negative outliers did not cancel each other out.

The OF, ORF, MG-MOST and the modified MG-LSOSS each displayed a non-decreasing function of the AUC with respect to the μ_g and k_g , however, the ORF, MG-MOST and the modified MG-LSOSS had much lower FDR in the region where the FDR varied the most. These three methods should provide the best detection power with the smallest FDR when applied to real data where the underlying parameters are unknown.

The application of outlier detection to structural MRI data was very appealing application area because of the lack of meaningful results using standard methods and the outlier detection is a reasonable fit for this type of data. With groups determined by neuropsychological trajectories, potential brain regions with outlier subsets were ranked to evaluate candidate brain regions. The two-group analysis showed 11 regions that were ranked in the top 15 of at least three methods, although none were identified by all methods. Several of these regions were highly ranked with the t -statistic which may suggest the number of samples in a potential subset is large or the magnitude difference between groups was large for those regions. A particularly interesting point is the number of subject that each method estimates as the size of the outlier subset. The LSOSS generally estimates a large size than that of all preceding methods. This was observed in

simulation when there were no outliers. The LSOSS method tends to estimate the number of outlier samples to be about half the group size which lead to the development of the modified LSOSS.

The multi-group application to the structural MRI data had 12 brain regions identified in the top 15 of at least three methods. It is particularly noteworthy that these brain regions were identified almost entirely in the following methods: OF, ORF, MOST and modified LSOSS. This set of methods excludes the MG-LSOSS, which does not have a non-decreasing function of the AUC, and the F-statistic, which does not incorporate the direction search for outliers. The 12 brains regions that were repeated identified with outlier detection and not in the F-statistic suggest that these methods have identified brain regions with significant outlier subsets that the F-statistic does not detect.

The multi-group outlier detection methods have been analyzed in this simulation study using many sets of parameter values. However, there are many more configurations of outlier patterns can be tested such as groups without outliers or unequal outlier magnitude between groups. Assessing the performance of multi-group outlier detection under other configurations can be a one direction for future research. Another unexamined factor is the underlying distribution of the normal samples which have restricted to only normally distributed data. Determining the effectiveness under alternative distributions could be more important to real data applications. Another approach to outlier detection is to simultaneously search for outliers in either direction. This two-sided approach was presented alongside the outlier sum statistic but was not analyzed in a simulation study. Studying the effect of correlation between markers is an additional scenario that has yet to be examined.

BIBLIOGRAPHY

1. Tomlins, S.A., et al., *Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer*. *Science*, 2005. **310**(5748): p. 644-8.
2. Tibshirani, R. and T. Hastie, *Outlier sums for differential gene expression analysis*. *Biostatistics*, 2007. **8**(1): p. 2-8.
3. Wu, B., *Cancer outlier differential gene expression detection*. *Biostatistics*, 2007. **8**(3): p. 566-75.
4. Lian, H., *MOST: detecting cancer differential gene expression*. *Biostatistics*, 2008. **9**(3): p. 411-8.
5. Wang, Y. and R. Rekaya, *LSOSS: Detection of Cancer Outlier Differential Gene Expression*. *Biomark Insights*, 2010. **5**: p. 69-78.
6. Liu, F. and B. Wu, *Multi-group cancer outlier differential gene expression detection*. *Computational Biology and Chemistry*, 2007. **31**(2): p. 65-71.