DATA ANALYSIS FOR MICROARRAY EXPERIMENT AND DNA BARCODE OF LIFE

BY CHING-RAY YU

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Statistics

Written under the direction of

Professor Javier Cabrera

and approved by

New Brunswick, New Jersey

Jan, 2008

ABSTRACT OF THE DISSERTATION

Data Analysis for Microarray experiment and DNA Barcode of Life

by Ching-Ray Yu

Dissertation Director: Professor Javier Cabrera

DNA microarray experiment, a well-established experimental technique, aims understanding the function of genes in some biological functions and cellular processes. One of the most common experiments in functional genomic research is to compare two groups of microarray data to determine which genes are differentially expressed. In this dissertation, we propose (1) a methodology to estimate the proportion of differentially expressed genes in microarray experiments, (2) parametric and non-parametric methods to estimate error distribution of microarray data, and (3) an optimal scoring method and LDA on HLdata on the DNA barcoding data to cluster the species using COI sequence. We study the performance of our methods using simulation studies where we compare it to other standard methods and apply it on real data sets to show the advantage of our method.

Acknowledgements

I would like to express my sincerest gratitude and appreciation to my adviser, Professor Javier Cabrera. He has been actively interested in my work and has always been available to advise me. I am very grateful his patience, motivation, and immense knowledge in cDNA microarray analysis and DNA barcode, that make him a great mentor. His invaluable and constant guidance and encouragement are indispensable for the completion of this dissertation. I would like to thank my other committee members, Professor William Strawderman, Professor Kesar Singh and Dr. Dhammika Aramatunga for their precious advise and suggestion to this dissertation.

I also like to thank Professor Cun-Hui Zhang who gave me great helps when I came to Rutgers university. I would especially thank graduate director Professor John Kolassa, whose kind helps during my study in Rutgers have been generous and invaluable. I also grateful to every members of the department of Statistics and Biostatistics.

Dedication

To my family both in America and Taiwan

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction to Genomic Data

In 1953, Dr. James Watson and Dr. Francis Crick proposed the double helical structure of Deoxyribose Nucleic Acid(DNA) in Nature [100] indicating that the chemical structure of two parallel strands is made of 4 types of nucleotides, which are adenine(A), thymine(T), guanine(G), and cytosine(C). The bases on one strand are paired with the bases on the other strand according to the complementary base pairing rules: adenine only pairs with thymine, guanine only pairs with cytosine. The pairs formed are called base pairs. Genes are pieces of DNA sequence and each piece is about 2000 base pairs on average. Genes consist of DNA, which is the hereditary material that passes from one generation to the next, indicates the inherent properties of a species. There are two types of regions in DNA sequences. One is protein-coding region, exon, which is a segment of DNA sequence that can produce proteins and contain hereditary messages. The other one is protein-noncoding region, intron, which can not produce proteins. Only about 5% of DNA sequence in chromosomes are exons. Dr. Francis Crick also proposed the Central Dogma of Molecular Biology (Figure 1.1) in Nature [25] showing that the production of proteins is controlled by genes, which are coded in DNA. Protein production from genes involves two principal stages, known as transcription and translation. During transcription, the double strands DNA sequence that corresponds to the gene is separated and one of two strands is copied into a single strand sequence which we called mRNA, messenger ribonucleic acid. After transcription, mRNA is used as a template to assemble a chain of amino acids to form a protein. The process is activated by the protein called the transcription factor. There are transcription fatcor protiens that activate the transcription factor binding site to start the process of transcriptions. Some transcription factor proteins trigger transcription process of one gene, but some can trigger the process of several genes at the same time.

After the great discovery, many researchers started to focus on the DNA sequence in order to realize how DNA affects our life. Gene expression studies investigate the amount of transcribed message RNA (mRNA) for a specific biological system. Several techniques are available for measuring gene expression, including serial analysis of gene expression(SAGE), complementary DNA (cDNA) library sequencing, differential display, multiplex quantitative Reverse Transcription Polymerase Chain Reaction (RT-PCR), which is a process of rapid generating multiple copies of any fragments of DNA, and gene expression microarrays.

Genomics is the branch of biology that studies the structure and function of genes. Structure genomics is the application of sequencing technologies to establish representative genome sequences for different organisms, particular in humans. Functional genomics is the study of the function of genes. It is important to realize how genes operate in pathways that are as part of biological processes also called biological pathways that are involved in the biological function. It can be said to have appeared in the 1980s, and took off in the 1990s with the initiation of genome projects for several biological species. A major branch of genomics is still concerned with sequencing the genomes of various organisms, but the knowledge of full genomes has created the possibility for the field of functional genomics, mainly concerned with patterns of gene expression during various conditions. DNA is found in the nucleo of cells, but also found in the mitochondrion called mitochondrial DNA. Microchondrial DNA is very useful not only for the discovery of gene mutations but also classification of the world's species. Scientific researchers ask questions about genome of interest and try to solve them by developing new experiments and new methods of data analysis. The most important question from the scientists point of view is to identify genes and cellular pathways that are difficult to study directly, studying protein coding and gene expression, providing large databases that are amenable to statistical methods, identifying variant sequences that may have subtle phenotypes or new species, and studying evolution of the organism and genome. It generates the set of techniques, analytical methodologies, and scientific questions to the study of complete genomes.

Many scientists from statistics, bioinformatics, computer science $\cdots$ focus on the data analysis of genomics data sets. Huge and variant data sets were collected from many different aspect of experiments, such as 1. Spotted cDNA microarrays, 2. Oligonucleotide(Affymetrix) arrays, 3. Single nucleotide polymorphisms(SNPs) , 4. Protein arrays, 5. Chromatin immunoprecipitation chip(ChIP-chip), 6. DNA barcoding of life, and 7. The Y chromosome. The components of our approach are the data analysis for the gene expression microarray experiments and the DNA sequence or SNPs arrays.

1.1 Spotted cDNA microarray

A microarray is a glass slide whose surface has been divided into series of imaginary square cells to form a rectangular grid. A probe is the single-stranded DNA molecules, whose sequence is known, is prepared and labeled with a reporter chemical, usually a radioactive or fluorescent substance. Onto each square cell, stick a tiny amount of liquid that contains DNA corresponding to a gene of known sequence(probe). Separately prepare a solution that contains a mixture of mRNA whose sequences are unknown which is called targets; add to this solution a substance that fluoresces when excited by light; pour the solution onto the slide. The mRNA molecules will diffuse over the slide and find a matching (complementary) DNA sequence, such as hybridization to each other and the solution will stick to the slide. Without a match, the solution will not stick to the slide and can be washed away. Use a laser scanner to detect and measure the fluorescent signal being emitted at each cell of the slide.

The spotted cDNA microarrays (one or several channel microarrays) are cDNA, pieces of genes that we like to identify, or small fragments of PCR products that are reverse transcribed using RT-PCR from mRNAs and are printed onto the microarray chip surface. This type of array is typically hybridized with cDNA (Figure 1.2), which are reverse transcribed from mRNA of two samples to be compared (e.g. diseased tissue versus healthy tissue). For the typical two channel experiments, cDNA are labeled with two different fluorophores (e.g. Cy 5 (red) and Cy 3 (green)), and two samples are mixed and hybridized to a single microarray. During hybridization, samples from different tissues compete to hybridize with the particular cDNA that they are complementary matched on the array. After hybridization, cDNA microarray is then scanned in a microarray scanner to visualize fluorescence of the two fluorophores. The relative intensities of each fluorophore are then used to identify up-regulated and down-regulated genes in analysis. The advantage of microarrays is that it can monitor the expression levels of tens of thousands of genes simultaneously. Microarray techniques have therefore dramatically accelerated many types of investigations. Microarray data was obtained through several processes (Figure 1.3), we will discuss them more detail in the section 1.1.1 - 1.1.3, and chapter 2.

1.1.1 Experimental design

Before the experiment, the researchers have to decide which genes are to be printed on the arrays, which sources of RNA are to be hybridized to the arrays and on how many arrays the hybridizations will be replicated to obtain reliable data. Unlike classical statistical problems, microarray data structure has about ten thousand genes i.e. ten thousand dimensions but replication is very small. Researches have to decide how many replicates we need in an experiment. In Ting et al.[94], they pointed out that 3 replicates will be the minimal requirement on detecting differentially expressed gene with false positive rate less than 0.7% under the assumption of the normality of microarray data. Kerr and Churchill [55] and Glonek and Solomon [43] suggested an optimal design for replicated microarray experiment. Pan, Lin and Le [74] considered the sample size of replicates that can detect the differentially expressed genes. Speed and Yang [91] provided the efficiency of using a reference sample as against direct comparison. After experiment, we suggested that the quality control (QC) was essential for microarray experiment. Low quality will affect the results of statistical analysis [99].

The choice of experiment design depends not only on the number of different samples to be compared but on the aim of the experiment and on the comparisons which are primary interest. For example, suppose the primary focus of an experiment involving a large series of tumor and normal tissues is on finding genes that are differently expressed between the tumor and normal samples. Then direct tumor-normal comparisons on the same slide may be the best approach. By contrast, if the focus of the analysis is to determine tumor subtypes as in Alizadeh et. al.[2], then the use of a common reference RNA on each array may be better. Here the choice follows from the aim of the study, although statistical efficiency considerations also play a role. In the first case, tumor-normal comparisons could be made indirectly, via a common reference RNA, but precision would be lost in so doing.

1.1.2 Image analysis

The primary purpose of the image analysis step is to extract numerical foreground and background intensities for the red and green channels for each spot on the microarray. The background intensities are used to correct the foreground intensities for local variation on the array surface, resulting in corrected red and green intensities for each spot which become the primary data for analysis. A secondary purpose of the image step is to collect quality measures for each spot that might be used to detect unreliable spots or arrays or to assess the reproducibility of each spot value.

The first step is to image the array using an optical scanner. The array is physically scanned to produce a digital record of the red and green fluorescence emissions at each point on the array. This digital record typically takes the form of a pair of 16-bit tiff images, one for each channel, which records the intensities at each of a large number of pixels covering the array. Depending on the scanner, a number of settings can be varied to improve the sensitivity of the resulting image, one of the most common being the photomultiplier tube (PMT) voltage. The PMT voltage is usually adjusted so that the brightest pixels are just below the level of saturation (2^{16}), thus increasing the sensitivity of the image analysis for the less bright pixels.

The next step after scanning is to locate each spot on the slide. This is done mostly automatically by the image analysis software, using the known number and basic layout of spots on the slide, with some user interventions to increase reliability. Once a region containing a spot itself (the foreground) and those in the background. There are a number of methods for doing this. The oldest method is the histogram method. A mask is chosen surrounding each spot and a histogram is formed from the intensities of the pixels within the mask. Pixels are classified as foreground ground if their value is greater than a threshold and as background otherwise. Variations on this method are implemented in QuantArray software [111] for the GSI Lumonics scanner and in DeArray [113] by Scanalytics. The main advantage of this method is simplicity. The resulting foreground pixels are not necessary connected though and the foreground and background intensities may be over and under-estimated respectively.

Chen. et. al.[22] proposed a nonparametric method to detect the spot microarray intensities. The intensity of each spot would be measured by Mann-Whitney rank-sum test. The Mann-Whitney rank-sum test as employed here. Assume that $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_m$ are independent samples coming from two distributions F and G with median M_X and M_Y, respectively. The rank-sum test statistic W, which is the sum of the ranks of all X samples in the combined ordered sequence of the X and Y samples, is to test the null hypothesis,

$$H_0: \ M_X = M_Y \ \text{ vs. } H_1: \quad M_X > M_Y.$$

Rejection of H_0 occurs, when $W \geq w_{\alpha,n,m}$, the critical value at level α. For each spot, they define a target mask which is a region containing all of the signal pixels. 8 i.i.d. sample pixels outside the target mask as $Y_1, \cdots, Y_8$ and pick lowest 8 i.i.d. sample pixels within the target mask as $X_1, \cdots, X_8$. The rank-sum statistics W is calculated and compares with $w_{\alpha,8,8}$, which has been tabulated (e.g., see Hollander and Wolfe [48]). If the null hypothesis is not rejected, then some predetermined number (perhaps 1) of the 8 samples is discarded from inside of the target mask and selected the lowest 8 remaining samples from region. This procedure is repeated until the null hypothesis is rejected. It says that the distribution of the pixels inside the target mask is different from that outside the target mask. After determining the pixels of this spot, it is usually estimated the average of the pixels within the mask as the foreground intensity of this spot and the median of the pixels without the mask as the background intensity of this spot, but there is a first decision should be made regarding which pixels to include in the local background.

Other methods are designed to find spots as connected groups of foreground pixels. The simplest method is to fit a circle of constant diameter to all spots in the image.

This is easy to implement and works nicely when all spots are circular and of the same size. In practice, this is not always the case. A generalization is to allow the circle's diameter to be estimated separately for each spot. GenePix [110] for the Axon scanner and Dapple [18] are two software programs which implement such algorithms. Dapple calculates the second differences (Laplacian) between the pixels in each small square and finds the brightest ring (circle) in the Laplacian images. Adaptive circle segmentation often works well, but spots are rarely perfectly circular, especially from non-commercial arrayers.

Two methods for segmentation which do not assume circularity of the spot are the watershed method [9] and seeded region growing [1]. Both methods require the specification of starting pixels or seeds. Pixels adjoining is progressively added to the spot until adjacent spots appear to be distinctly less intense. Seeded region growing is implemented in the software Spot [17] and AlphaArray [99]. Both the watershed method and seeded region growing allow for spots of general shapes.

One choice for the local background is to consider all pixels that are outside the spot mask but within the bounding box. Such a method is implemented by ScanAlyze [36]. An alternative method used by QuantArray [111] and ArrayVision [109] is to consider a disk between two concentric circles outside the spot mask. This methods is in principle less sensitive to the the performance of the segmentation procedure because the pixels immediately surrounding the spot are not used.

Another method is to consider the valleys of the array which are the background regions farthest from the nearest spot. The method is used by GenePix [110]. It is also used by Spot as a quality control measure, although no for background correction. Since the valleys are further from any previous definitions to corruption by bright pixels affected by printed cDNA. Any of the local background methods can result in background estimates which are higher than the foreground values either because of corruption by mis-segregated pixels or local artifacts or simply because of local variation.

The Spot software estimates the background using a non-linear filter called morphological opening [90]. The filter has the effect of smoothing the entire slide image so that all local peaks, including artifacts such as dust particles as well as the spots themselves, are removed leaving only the background intensities. Technically, the filter consists of a local minimum filter followed by a local maximum filter. This method of background estimation has several advantages over the use of local background regions. Firstly, it is less variable because the background estimates are based on a large window of pixels values and are yet not corrupted by bright pixels belonging to the actual spots. Secondly, it yields background intensity estimates at the actual spot location rather than merely nearby. Another characteristic is that the morphological background estimates are usually lower than the local background estimates and very rarely yield background estimates which are greater than the foreground values. Yang et al. [104] compared various segmentation and background estimation methods. They found that the choice of background method has a larger impact on the log-ratios of intensities than the choice of segmentation method and that morphological opening provides a more reliable estimate of background than other methods.

Having estimated the background intensities, it is almost universal practice to correct the foreground intensities by subtracting the background, and the adjusted intensities then form the primary data for all subsequent analysis. The motivation for background adjustment is the belief that a spot's measured intensity includes a contribution not specifically due to the hybridization of the target to the probe, for example non-specific hybridization and fluorescence emitted from other chemicals on the glass. If such a contribution is present, we would like to measure and remove it to obtain a more accurate quantification of hybridization. An undesirable side-effect of background correction is that negative intensities may be produced for some spots and hence missing values if log-intensities are computed, resulting in loss of information associated with low channel intensities. Researcher has begun on more sophisticated methods of background adjustment which will produce positive adjusted intensities even when the background estimate happens to be larger than the foreground [57]. Empirical experience suggests that local background estimates often over-estimate the true background while the morphological method may under-estimate and these differences have a marked impact on the mean of red and green channels for less intense spots. There is a need for further research on adaptive background correction methodologies which can produce intensities with consistent behavior regardless of background estimator used.

1.1.3 Normalization

The purpose of normalization is to adjust for any bias which arises from variation in the microarray technology rather than from biological differences between the RNA samples or the printed probes. Most common is red-green bias due to differences between labeling efficiencies and scanning properties of the two fluorophores complicated perhaps by the use of different scanner settings. Other biases may arises from variation between spatial position on a slide or between slides. Positions on a slide may vary because of differences between the print-tips on the array printer, variation over the course of the print-run or non-uniformity in the hybridization. Differences between arrays may arise from differences in print quality or from differences in ambient conditions when the plates were processed. It is necessary to normalize the intensities before any subsequent analysis is carried out.

The general method for normalization of red-green bias is the following:

1. Global normalization: In order to correct the bias, we let $R = k \times G$, where R and G are the intensity of the red and green channel, respectively, and k is a constant. Then
$$log_2 \frac{R}{G} \longrightarrow log_2 \frac{R}{G} - c = log_2 \frac{R}{k \times G},$$
where the location parameter $c = log_2 k$.

2. Intensity dependent normalization:
$$log_2 \frac{R}{G} \longrightarrow log_2 \frac{R}{G} - c(A) = log_2 \frac{R}{k(A) \times G},$$
where the location parameter $c(A) = log_2 k(A),\ A = \frac{1}{2}(log_2 G + log_2 R)$.

3. Tip normalization:
$$log_2 \frac{R}{G} \longrightarrow log_2 \frac{R}{G} - c_i(A) = log_2 \frac{R}{k_i(A) \times G},$$

where $i = 1, \cdots, I$ and I presents of the number of print-tips and the location parameter $c(A_i) = log_2 k_i(A)$. On the same spot, the intensity of the red and green channels are different (Figure 1.4). This is chemical mechanical bias. So we must remove this bias $(c, c(A), c_i(A))$ using dye-swap.

4. Paired-slides normalization (dye-swap):

   Paired-slides normalization applies to dye-swap experiment: two hybridizations for two mRNA samples, with dye assignment reversed in the second hybridization. Denote the normalized log-ratio for the first slide by $log_2\frac{R}{G} - c$ and those for the second slide by $log_2\frac{R'}{G'} - c'$. Here R' and G' are the red and green intensities of the second slide and c and c' indicate the normalization functions for the two slides; these could be obtained by any of the within-slide normalization methods 1 - 3. If $c \approx c'$, then

   $$\frac{1}{2}[log_2\frac{R}{G} - c - (log_2\frac{R'}{G'} - c')] \approx \frac{1}{2}(\frac{R}{G} + log_2\frac{G'}{R'}) = \frac{1}{2}log_2\frac{RG'}{R'G} = \frac{1}{2}(M - M').$$

   where M and M' are $log_2\frac{R}{G}$ and $log_2\frac{R'}{G'}$ respectively. The main assumption here is that $c \approx c'$ and this method can be applied to a set of genes expected to have constant expression levels (such as housekeeping genes), if such genes are available. In dye-swap experiment, we expect that

   $$log_2\frac{R}{G} - c \approx -(log_2\frac{R'}{G'} - c').$$

   We can estimate the normalization function c(red-green bias) by

   $$c \approx c' = \frac{1}{2}(log_2\frac{R}{G} + log_2\frac{R'}{G'}) = \frac{1}{2}(M + M').$$

   Similarly, we can estimate $c(A), c_i(A)$ using loess or lowess normalization method with polynomials fitted locally using iterated weighted least squares [24].

5. Quantile normalization

   The quantile normalization method for microarray data is proposed by Amaratunga and Cabrera in 2001[4]. The idea is that if X_{gi} denotes the transformed spot intensity for the g^{th} gene $(g = 1, \cdots, G)$ in the i^{th} microarray $(i = 1, \cdots, I)$, the median mock array for i^{th} gene will define as:

$$M_g = median\{X_{g1}, \cdots, X_{gI}\}.$$

Then both percentiles $(Q_{i0}, \cdots, Q_{i100})$ of the i^{th} array and $(Q_{M0}, \cdots, Q_{M100})$ of the median mock array are calculated. For any value X_{gi}, find a interval, $[Q_{ih}, Q_{i(h+1)}]$, such that $X_{gi} \in [Q_{ih}, Q_{i(h+1)}]$ and obtain its normalized value, X'_{gi}, by linearly interpolating between the pair of points (Q_{Mh}, Q_{ih}) and $(Q_{M(h+1)}, Q_{i(h+1)})$. Quantile normalization is useful for normalizing across a series of conditions where it is believed that a small but indeterminate number of genes may be differentially expressed, and it can be assumed that the distribution of spot intensities does not vary too much.

In all of the above normalization methods, it is usual to use all or most of the genes on the array. It can be useful to modify the normalization methods if a suitable set of control spots is available. A traditional method is to use housekeeping genes for normalization. However housekeeping genes often do show sample specific bias. Housekeeping genes are also typically highly expressed so they will not allow the estimation of dye-biases for less expressed genes when the dye-bias is intensity dependent. Housekeeping genes may also not be well represented on all parts of the plate so that spatial effects may not be well estimated. The most satisfactory set of controls is a specially designed microarray sample pool (MSP) titration series. MSP is analogous to genomic DNA as control with the exception that non-coding regions are removed. Typically a concentration titration is done to span as wide an intensity range as possible. Theoretically all labeled cDNA sequences could hybridize to this mixed probe sample, so it could be minimally subject to any sample specific biases. On the other hand, the use of all genes for normalization offers the most stability in terms of estimating spatial and intensity dependent trends in the data. In some cases it may be beneficial to use a compromise between the sub-array loess curves and the global titration series curve[97].

An alternative method is to select an invariant set of genes as described for oligonucleotide arrays by Schadt et. al.[83] and Tseng et al. [95]. A set of genes is said to be invariant it their ranks are the same for both red and green intensities. In practice, the set of invariant or approximately invariant genes is too small for comprehensive normalization. When there are sufficient invariant genes, the use of invariant genes is similar to global intensity-dependent normalization as describe above.

1.1.4 Data analysis

Statistical Analysis of Microarray [96], SAM, is a penalized t-statistic to detect differently expressed genes. We will discuss more in section 2.2. The other useful software, LIMMA, LInear Model for MicroArray data by [88], is a package for the analysis of gene expression microarray data, especially the use of linear models for analyzing designed experiments and the assessment of differential expression. The package includes preprocessing capabilities for two-color spotted arrays. The differential expression methods apply to all array platforms and two channel experiments in a unified way. LIMMA is available from the R Project CRAN site [112] or as part of Bioconductor project. Both parametric and nonparametric methods are applied to microarray data analysis. More statistical methods will be discussed in chapter 2.

1.2 Affymetrix Oligonucleotide array

In oligonucleotide microarrays (or single-channel microarrays), the probes are designed to match parts of the sequence of known or predicted mRNAs. Oligonucleotide array technology [64] has recently been adopted in many areas of biochemical research. In [63], 16-20 probe pairs are used to interrogate each gene; each probe pair has Perfect match(PM) and Mismatch(MM) signal, and the average of the PM-MM differences for all probe set (called ”average difference”) is used as an expression index for the target gene (Figure 1.5). Researchers rely on the average differences as the starting point for ”high-level analysis” such as SOM analysis [93] or two way clustering [3]. Besides the original publication by Affymetrix scientists [102], there have been studies on important ”low-level” analysis issues such as feature extraction, normalization, and computation of expression indexes [82].

Suppose that a number ($I > 1$) of samples have been profiled in an experiment. The expression-level estimates are constructed from the $2 \times I \times 20$ (assuming a probe set has 20 probe pairs) intensity values for the PM and MM probes corresponding to this gene. The estimation procedure is based on a model of how the probe intensity values response to changes of the expression levels of the gene. Let θ_i denote an expression level for the gene in the i^{th} sample. The intensity value of a probe will increasing linearly as θ_i increases, but the rate of increase will be different for different probes. It is also assume that within the same probe pair, the PM intensity will increase at a higher rate than the MM intensity. Then the statistical model for oligonuceotide array can be:

$$MM_{ij} = \nu_j + \theta_i \alpha_j + \epsilon \tag{1.1}$$

$$PM_{ij} = \nu_j + \theta_i \alpha_j + \theta_i \phi_j + \epsilon, \tag{1.2}$$

here PM_{ij} and MM_{ij} denote the PM and MM intensity values for the $i^{th}, i = 1, \cdots, I$ array and the $j^{th}, j = 1, \cdots, 20$ probe pair for this gene, ν_j is the baseline response of the j^{th} probe pair due to non-specific hybridization, α_j is the rate of increase of the MM response of the j^{th} probe pair, ϕ_j is the additional rate of increase in the corresponding PM response, and ϵ is a generic symbol for a random error(usually assumed normally distributed). The rate of increase are assumed to be non-negative. A simple model for the PM-MM differences is :

$$y_{ij} = PM_{ij} - MM_{ij} = \theta_i \phi_j + \epsilon_{ij}.$$

Recently, many researchers focused on the PM-MM difference model to select differentially expressed gene from oligo array experiments.

1.3 Protein microarrays

A protein microarray is a highly ordered pattern of proteins immobilized on a pretreated surface of a small and planar metal, plastic, or glass support. Protein array experiments display similarities to their DNA microarray counterparts. Protein microarray technology enables high throughput analysis of protein function, such as interactions between protein, catalysis, binding to drugs and other biochemical reactions. This can be inferred from more than 100 protein array-oriented scientific publications in the past two year. Ultimately, a single microarray containing the complete set of 20,000 - 40,000 proteins expressed in the cells would allow comprehensive assessment of a given protein function. Putting diverse protein repertoires on a microarray requires the simultaneous and quality-assured production of many recombinant proteins of high purity. Protein can be extracted from flood, fluid, cell lines, or fresh tissue by the use of various cell analysis buffers. Sample preparation is critically important because it may affect the reproducibility and thus comparability of a given set of proteins. Variability in protein expression between samples may result from the heterogeneity of cell populations in a sample. The data of protein microarray is the intensity of the protein spot scanned from image machine(e.g. Kodak CCD camera), which is very similar to cDNA microarray data. The statistical methods in protein array is similar to methods in cDNA microarray, but statistical models are more complicated than that of cDNA microarray because of heterogeneity.

Proteins carry out all kinds of housekeeping activities, they are catalysts of chemical reactions, they act as channels and pumps, and they perform motor functions. Some of the proteins involved in protein array experiments are as follows:

- Antibodies

  Antibodies are proteins produced by B-lymphocyte cells, which are a certain type of white blood cell. As part of the immune system, the function of an antibody is to bind with a specific protein lying on the surface of a foreign call. This protein-binding property plays an important role in the technology for the realization of protein array experiments. There are five classes of antibodies that are also called immunoglobulins.

- Antigens

  Antigens are proteins that lie on the surface of foreign cells and are detected by specific antibodies. Antigens will bind with antigens in order to neutralize them and to help other parts of an organism's immune system recognize foreign cell such as bacteria or viruses.

- Enzymes

  These are proteins that perform catalytic functions; that is, they accelerate a chemical reaction without been consumed by it. In particular, enzymes are involved in the synthesis of DNA and proteins. Enzymes are involved in the synthesis of proteins from RNA code by translation. The RNA code is subdivided into triplets of ordered nucleotides that are called codons. Proteins are formed of chains of amino acid molecules. There are 20 possible amino acids, and each codon codes for one specific amino acid, but more than one codon may code for the same amino acid. The process of protein formation consists of translating the RNA code into a chain of amino acids bonded together to form the protein molecule. The enzyme's role in the protein formation is similar to the role of an assembly line in the making of a product.

Although there are many similarities between the images scanned from protein arrays and the images scanned from DNA microarrays, the processes that generated them are quite different. Some of the issues that differentiate protein arrays from their DNA sibling that affect the data analysis [5]. The data analysis part is essentially similar to data from cDNA microarray experiments.

1.4 ChIP-chip array

ChIP-chip(or ChIP-on-chip), known as genome-wide location analysis, is a technology for isolating genomic sites occupied by specific DNA binding proteins in living cells. This strategy may be used to annotate functional elements, such as promoters, motif region, enhancers, represser elements, and insulators, in genomes by mapping the locations of protein markers associated with these sites. "ChIP" refers to "chromatin immunoprecipitation", which is a method for isolating DNA fragments that are bound by specific DNA binding proteins. "Chip" indicates the DNA microarray technology for measuring the concentrations of these DNA fragments. The models for ChIP-chip data are the probabilities calculated for one random genome sequence manifest themselves as frequencies among the large number of genome sequences in an experiment. Some researchers [82][83] used protein arrays to analyze the active promoter in human genome. They used antibodies specially recognizing components of the transcription pre-initiation complex to obtain a high-resolution map of active promoters in human genome. Using this approach, they were able to annotate transcriptional start sites and discover novel genes. Recent ChIP-chip studies have begun to address the question of how the cell is able to use single transcription factor to elicit multiple downstream transcriptional responses.

1.5 Single Nucleotide Polymorphisms(SNPs)

A SNP is a specific location in our DNA where different people have different DNA bases. 99.9% of one individuals DNA sequences are identical to that of another person. Over 80% of this 0.1% difference will be SNPs. A SNP is a single base substitution of one nucleotide with another. Both substitutions have to be observed in the general population at a frequency greater than 1%. An example of a SNP is individual ”A” has a sequence GAACCT, while individual ”B” has sequence GAGCCT, the polymorphism is a A/G. Current estimates are that SNPs occur as frequently as every 100-300 bases. This implies that in an entire human genome there are approximately 10 to 30 million potential SNPs. More than 4 million SNPs have been identified and the information has been made publicly available. Unfortunately, many of these SNPs have unknown associations.

Recent work has suggested that SNPs in human population are not inherited independently; rather, sets of adjacent SNPs are present on alleles in a block pattern, so called haplotype. Many haplotype blocks in human have been transmitted through many generations without recombination. This means although a block may contain many SNPs, it takes only a few SNPs to identify or tag each haplotype in the block Many common diseases in humans are not caused by one genetic variation within a single gene, but are determined by complex interactions among multiple genes, environmental and lifestyle factors. Genetic factors confer susceptibility or resistance to a disease and influence the severity or progression of disease. Researchers may begin to reveal relevant genes associated with a disease, by studying SNP profiles or haplotypes associated with a disease trait. Association study can detect and indicate which pattern is most likely associated with the disease-causing genes. Eventually, SNP profiles that are characteristic of a variety of diseases, will be established. Then, it will only be a matter of time before physicians can screen individuals for susceptibility to a disease just by analyzing their DNA samples for specific SNP patterns.

The race among pharmaceutical companies today, is to apply new system genomics approach to identify novel targets and validate these targets in the most efficient fashion. SNP research will provide fundamental understanding of many polygenic diseases, thus providing new therapeutic targets. Another significant goal is to identify those SNPs which are associated with significant biological effects in response to chemical drugs. A large percentage of people given a drug respond in the intended medically beneficial way, however some smaller percentage might either have no response or have a life threatening response and death. This adverse drug response (ADR) is believed to cause thousands of deaths annually. The SNP effort will serve as the bedrock of pharmacogenomics, the emerging field of personalized medicine: the right drug, in the right dose, to the right person, at the right time.

SNP study is also extremely important in organisms other than humans. Within agriculture, genetic modification of the agriculturally important crops (corn, wheat, rice, soybeans, etc.) could lead to improve crop yields at lower cost by reducing the amounts of fertilizer, insecticides, herbicides required. Within microorganisms and viruses, SNPs are known to cause increased drug resistance. Some of the recent E. Coli outbreaks are due to new evolving strains of the bacterium. HIV, the causative agent of AIDS, has historically been so difficult to treat with drugs due to very high mutation frequency primarily in the form of SNPs.

Recently, SNPs are found to be very useful for a complete different purpose. SNPs turned out to be valuable genetic markers for revealing the evolutionary history of populations. Their occurrence throughout the genome also makes them ideal for analysis of specification and historical demography, especially in light of recent theory suggesting that many unlinked nuclear loci are needed to estimate population genetic parameters with statistical confidence. In spite of having lower variation compared with microsatellites, SNPs should make the comparison of genomic diversities and histories of different species (the core goal of comparative biogeography) more straightforward than has been possible with microsatellites. The most pervasive, but correctable, complication to SNP analysis is a bias toward analyzing only the most variable loci, an artifact that is usually introduced by the limited number of individuals used to screen initially for polymorphisms.

1.6 DNA Barcoding of Life(BoL)

In the past two years, a series of studies [46][47] have been published in which "DNA barcoding" was proposed as a tool for differentiating biological species. Barcoding is based on the assumption that short gene regions evolve at a rate that produces clear interspecific sequence divergence while retaining low intraspecific sequence variability. With million of species and their life-stage transformations, the animal kingdom provides a challenging target for taxonomy. Recent work has suggested that a DNA-based identification system, founded on the mitochondrial gene, cytochrome *c* oxidase subunit 1 (COI) with 648 base pairs long, can aid the resolution of this diversity. COI has emerged as a suitable barcode region for most taxonomic groups of animals. Some articles [46] [47] showed that the sequence divergences at COI sequence regularly enable the discrimination of closely allied species in most animal phyla. This success in species diagnosis reflects both the high rates of sequence change at COI in most animal groups and constrains on intraspecific mitochondrial DNA divergence arising, at least in part, through selective sweeps mediated via interactions with the nuclear genome. There is no compelling a priori reason to focus analysis on a specific gene, but COI sequence does have two important advantages. First, this gene is very robust . Second, COI appears to possess a greater range of phylogenetic signal than any other mitochondrial gene. So the species-level diagnoses can routinely be obtained through COI analysis.

1.7 The Y Chromosome

The Y chromosome, with the genes to make a man, has been sequenced. It is often regarded as a genetic wasteland, but the sequence of the Y chromosome reveals that we may have underestimated its powers. Because of its distinctive role in sex determination, the Y chromosome has long attracted special attention from geneticists, evolutionary biologists and even the lay public. It is known to consist of regions of DNA that show quite distinctive genetic behavior and genomic characteristics. The two human sex chromosomes, X and Y (Figure 1.6), originated a few hundred million years ago from the same ancestral autosome, a non-sex chromosome, during the evolution of sex determination [72]. They then diverged in sequence over the succeeding aeons. Nowadays, there are relatively short regions at either end of the Y chromosome that are still identical to the corresponding regions of the X chromosome, reflecting the frequent exchange of DNA between these regions ('recombination') that occurs during sperm production [19]. But more than 95% of the modern-day Y chromosome is male-specific, consisting of some 23 million base pairs (Mb) of euchromatin, the part of our genome containing most of the genes, and a variable amount of heterochromatin, consisting of highly repetitive DNA and often dismissed as non-functional. Skaletsky et al. [87] reported the complete sequence of the 23-Mb euchromatic segment, which they designated the male-specific region of the Y (MSY). As Skaletsky et al. reported, the MSY is a mosaic of complex and interrelated sequences that made this one of the most problematic regions of the human genome thus far to be successfully sequenced and assembled. How much we can learn from Y-chromosome analysis depends on:

1. Choosing the right loci to look for mutations on the Y-chromosome. Loci mutate at different rates, so choosing the right loci is important. Kayser et. al. [51] established direct experimental evidence to support mutation rates at Y-chromosome loci in father-son pairs. They used the binomial probability model to estimate the mutation rate at microsatellite loci in human Y-chromosome.

2. Choosing how many loci to compare between Y-chromosomes Bruce Walsh [97] at the University of Arizona used the binomial probability model to determine how many markers are needed to group Y-chromosomes into biological family groups. The minimum number of markers appears to be 10 by Maximum Likelihood Estimation from Y-chromosome microsatellite markers.

3. The number of researchers participating in a Y-chromosome analysis is important to establish the halpotype of a family group. The more scientists that participate, the more confident we can be of grouping individuals together as family groups.

1.8 Discussion

Completed in 2003, the Human Genome Project (HGP) was a 13-year project coordinated by the U.S. Department of Energy and the National Institutes of Health. The goals of Human Genome Project are (1) identify all the approximately 20,000-25,000 genes in human DNA, (2) determine the sequences of the 3 billion chemical base pairs that make up human DNA, (3) store this information in databases, (4) improve tools for data analysis, (5) transfer related technologies to the private sector, and (6) address the ethical, legal, and social issues (ELSI) that may arise from the project. These goals are all very important to us. In addition, after HGP, many types of genomic data other than human's are generated from researchers based on different research purposes. I only list some of them that are more interested by researchers. In this thesis, I only focus on the analysis of cDNA microarray and barcode of life.

In this thesis, I propose a group of methodologies that try to answer some of the biological questions that are pored from the problems and data. First, in chapter 2, we develop an algorithm to estimate the proportion of differentially expressed genes in the microarray experiments. Second, in chapter 3, we develop non-parametric and parametric methods to estimate error distributions. Third, in chapter 4, we use Linear Discriminants Analysis (LDA) to classify the species in the NDA barcode data which is a data with high dimension (many variables) and low samples (few observations).

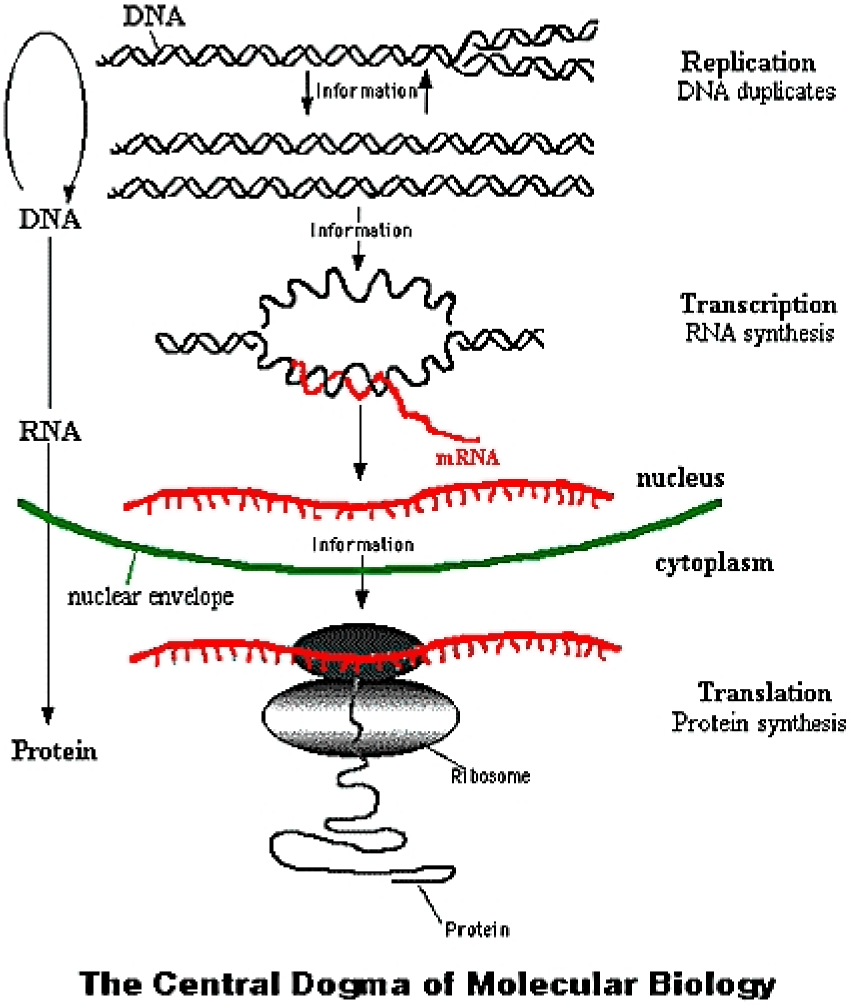
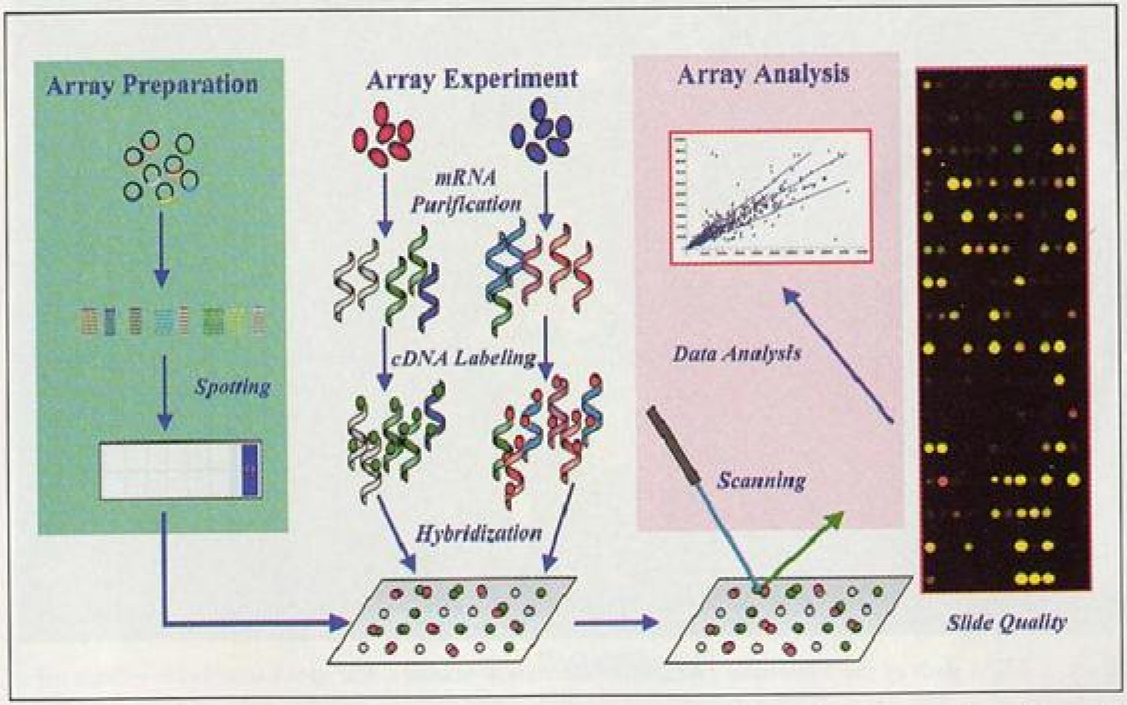


Figure 1.1: The Central Dogma of Molecular Biology(Source: Access Excellence)



Flow chart of two-color microarray experiments. The major phases of array preparation, differential gene-expression experiment, and analysis are listed, with the topics discussed in this chapter labeled in italics.

Yang et al. 2001

Figure 1.2: Flow chart of microarray experiments

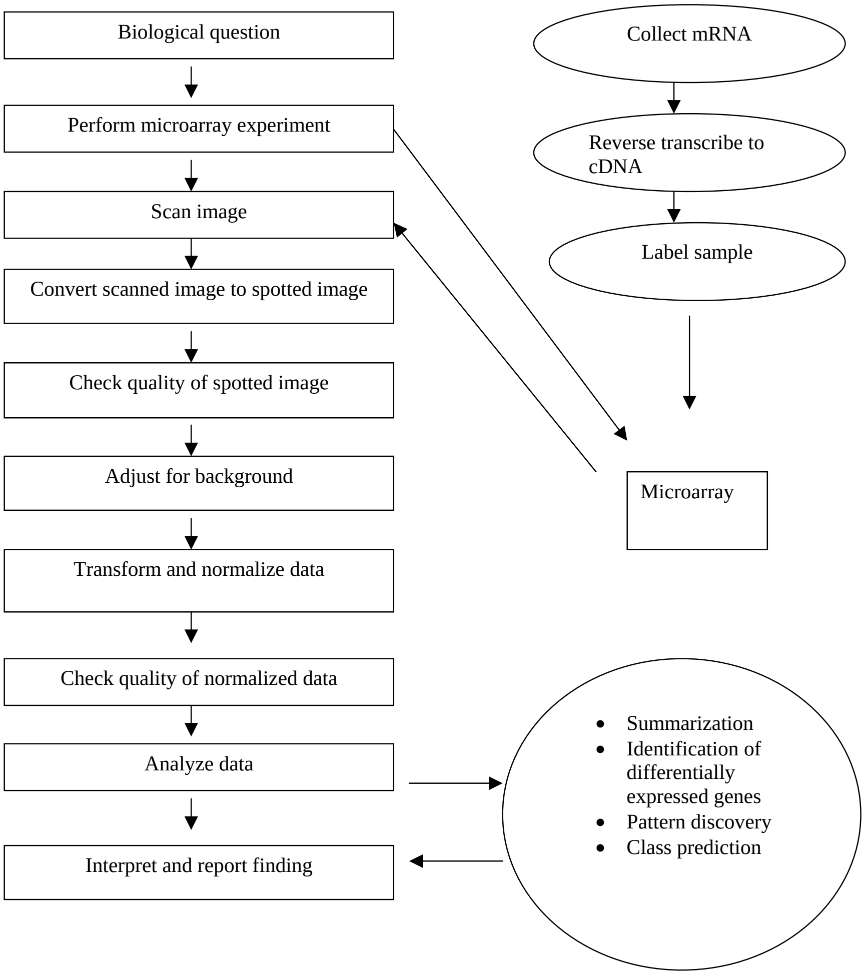
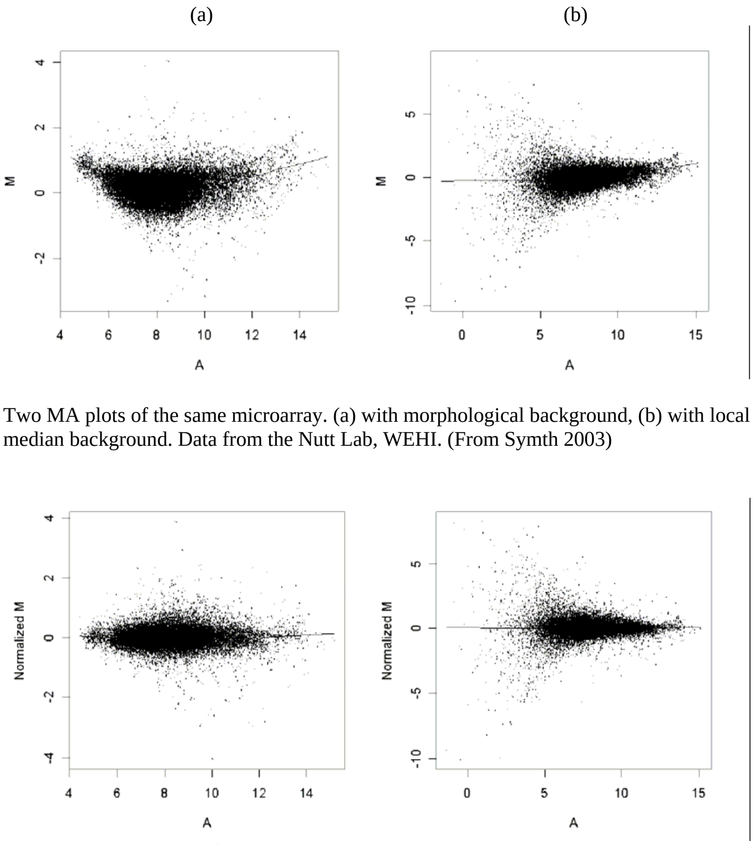

Biological question
Perform microarray experiment
Scan image
Convert scanned image to spotted image
Check quality of spotted image
Adjust for background
Transform and normalize data
Check quality of normalized data
Analyze data
Interpret and report finding
Collect mRNA
Reverse transcribe to cDNA
Label sample
Microarray
Summarization
Identification of differentially expressed genes
Pattern discovery
Class prediction

Figure 1.3: A flow chart of a typical microarray data analysis (Source: D. Amaratunga and J. Cabrera(2004). *Exploration and Analysis of DNA Microarray and Protein Array data*)



Two MA plots of the same microarray. (a) with morphological background, (b) with local median background. Data from the Nutt Lab, WEHI. (From Symth 2003)

The same two MA-plot after tip-normalization.

Figure 1.4: Tip normalization

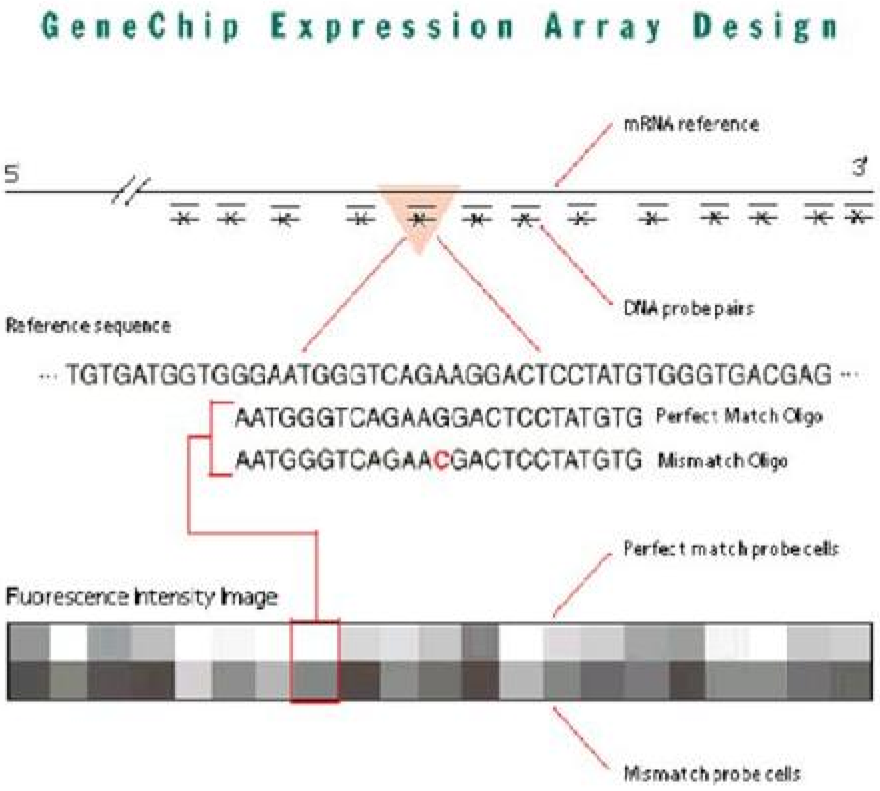

GeneChip Expression Array Design
mRNA reference
5
3'
DNA probe pairs
Reference sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAG ···
AATGGGTCAGAAGGACTCCTATGTG Perfect Match Oligo
AATGGGTCAGAACGACTCCTATGTG Mismatch Oligo
Perfect match probe cells
Fluorescence Intensity Image
Mismatch probe cells

Figure 1.5: Affymetix oligonuceotide microarray(From Affymetrix website)

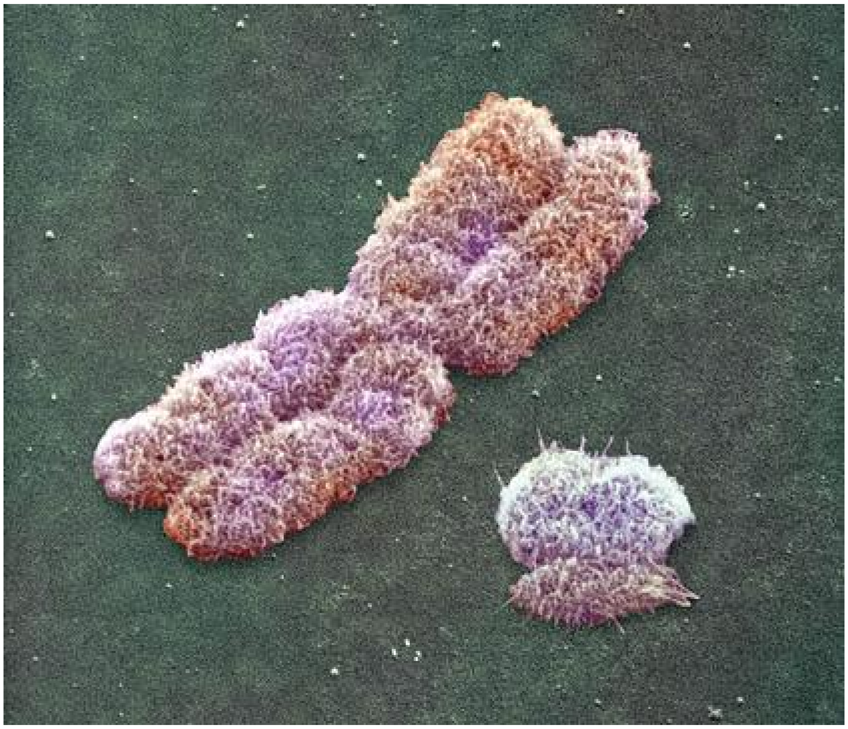


Figure 1.6: X-chromosome(left) versus Y-chromosome(right) Source: Willard, H.(2003), Nature

Chapter 2

cDNA microarray experiments and Data analysis

2.1 Introduction to the cDNA microarray experiments

The human genome and a number of other genomes have been almost fully sequenced, but the functions of most genes are still unknown. One of the difficulties is to understand gene functionality since gene expression is only one of the pieces of cellular processes sometimes called biological pathways or networks, and it is not yet possible to observe these pathways directly. The technology of cDNA microarray is now becoming widespread for measuring the simultaneous expression levels of thousands to tens of thousands of genes in a given cell type. It provides a powerful tool for genetic research and has been used to monitor changes in gene expression during important biological processes (e.g., cellular replication and the response to changes in the environment), and to study variation in gene expression across collections of related samples(e.g., tumor samples from patients with cancer). Statistical considerations are frequently to address the analysis of microarray data, as researchers sift through massive amounts of data and adjust for various sources of variability in order to identify the important genes among the many which are measured.

Any microarray experiment involves a number of distinct stages. First there is the design of the experiment(section 1.1.1). The researchers must decide which genes are to be printed on the array, which sources of RNA are to be hybridized to the arrays and on how many arrays the hybridizations will be replicated. Secondly (section 1.1.2), after hybridization, there follows a number of data-cleaning process of the microarray data. The microarray images must be processed to acquire red and green foreground and background intensities for each spot. The intensities have to be normalized to adjust for dye-bias and for any systematic variation other than due to the difference between the RNA samples being studies. Thirdly (section 1.1.3), researchers tried to select the genes that are differentially expressed or group of genes whose expression profiles can reliably classify the different RNA sources into meaningful groups. Then in the next section, we will focus on standard statistical methods and models to select differentially expressed genes.

2.2 A review of standard statistical techniques for selecting differentially expressed genes

2.2.1 t-statistic method

One of the core goals of microarray data analysis is to identify which of the genes show good evidence of being differentially expressed. This goal has two part. The first is select a statistic which will rank the genes in order of evidence for differential expression, from strongest to weakest evidence. The second is to choose a critical-value for the ranking statistic above which any value is considered to be significant. The first goal is more important than the second and, as it turns out, also easier. The primary importance of ranking arises from the fact that only a limited number of genes can be followed up in a typical biological study. In many microarray studies the aim is to identify a number of candidate genes for confirmation and further study. It will usually be practical to follow-up only a limited number of genes, 100 say, so it is most important to identify the 100 most likely candidates. The complete list of all genes which can be considered statistically significant may be of less interest if this list is too large to be followed up.

For simplicity, we will assume in this section that we have data from the simplest possible experiment. We will assume that we have a series of n replicate arrays on which samples A and B have been hybridized and we wish to identify which genes are differentially expressed. Many data analysis programs sort the genes according to the absolute level of $\bar{M}$, where $\bar{M}$ is the mean of the M-values for any particular gene across the replicate arrays. This is known to be a poor choice as it does not take account of the variability of the M-values over replicated is not constant across genes and genes with larger variances have a good chance of giving a large $\bar{M}$ statistic even if they are not differentially expressed. A better choice is to rank genes according to the absolute value of the t-statistic

$$t = \frac{\bar{M}}{s/\sqrt{n}},$$

where s is standard deviation of the M-values across the replicates for the gene in question, as this incorporates a different variability estimate for each gene. An added advantage of the t-statistic is that it introduces some conservative projection against outliers M-values and poor quality spots. Any M-value which is an outliers will give rise to a large standard deviation s which will usually prevent the gene in question from being spuriously identified as differentially expressed.

The ordinary t-statistic is still not ideal because a large t-statistic can be driven by an unrealistically small value for s. The shortcoming of the t-statistic is the opposite of that of $\bar{M}$. Genes with small sample variances have a good chance of giving a large t-statistic even if they are not differentially expressed. A suitable compromise between the $\bar{M}$ and t-statistics is therefore desirable. Efron et al. [34] have used penalized t-statistics of the form

$$t = \frac{\bar{M}}{(a+s)/\sqrt{n}},$$

when assessing differentially expressed for oligonucleotide microarrays. Lönnstedt and Speed [64] adopt a parametric empirical Bayes approach to the problem of identifying differentially expressed genes. The proposed a B-statistic which is an estimate of the posterior log-odds that each gene is differentially expressed. Subject to the parametric assumptions being valid for the data, values for the B-statistics greater than 0 correspond to a greater than 50-50 chance that the gene in question in differentially expressed. The B-statistic is equivalent for the purpose of ranking genes to the penalized t-statistic

$$t = \frac{\bar{M}}{\sqrt{(a+s^2)/n}},$$

where the penalty a is estimated from the mean and standard deviation of the sample variances s^2. Tusher [96] choose a to minimize the coefficient of variation of the absolute t-values while Efron [35] choose a to be the 90^{th} percentile of the s values. These choices are driven by empirical rather than theoretical considerations. Efron et al. uses the above t-value as the basis for a non-parametric empirical Bayes method leading to an estimated log-odds that each gene is differentially expressed. Lönnstedt and Speed [64] show in a simulation that both forms of penalized t-statistic are far superior to the mean $\bar{M}$ or to ordinary t-statistic for ranking differentially expressed genes.

The penalized t-statistics can be extended in several natural ways to apply to more general experimental situations. If there are missing values for some arrays, perhaps because low quality spots have been flagged for removal, then the value n in the denominator will reflect the actual number of observations for each gene rather than the total number of arrays.

The t-statistic also extends to more complicated experiment designs. For example we might use a penalized two-sample t-statistic if we are comparing samples A and B through a reference rather than directly on the same arrays. In that case there will be n_A replicate arrays comparing sample A with reference RNA and n_B replicate arrays comparing B with the same reference and a two-sample t-statistic,

$$t = \frac{\bar{M}_A - \bar{M}_B}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}},$$

where $s_p = \sqrt{a + s^2}$ is the penalized pooled sample standard deviation, and s is the pooled sample standard deviation from sample A and B, might be used. Here $\bar{M}_A$ and $\bar{M}_B$ are the average of the M-values for the two groups of arrays. Tusher et al. [96] suggested another panelized test statistics called SAM *t statistics*,

$$T_g(c) = \frac{\bar{M}_A - \bar{M}_B}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}} + c},$$

where SAM is for ' significant analysis of microarray'. An implement of this SAM by Tusher et al.[96] is as follows:
Let s^α be the α^{th} percentile of the $\{s_g\}$ values, and

$$T_g(s^\alpha) = \frac{\bar{M}_A - \bar{M}_B}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}} + s^\alpha}.$$

Compute the percentiles, $q_1 < q_2 < \cdots < q_{100}$, of the s_g values. For $\alpha \in \{0, 5, 10, \cdots, 100\}$, compute the MAD(median absolute deviation from median), $v_j(\alpha)$ of the $T_g(s^\alpha)$ values within the interval $[q_j, q_{j+1}]$ for $j = 1, \cdots, 100$. They then compute $cv(\alpha)$, the coefficient of variation of the $v_j(\alpha)$ values and then choose as $\hat{\alpha}$ the value of α that minimizes $cv(\alpha)$ and fixed as $\hat{\alpha}$ the value $s^{\hat{\alpha}}$.

In the general case, differential expression can be judged using a penalized t-statistic of the form

$$t = \frac{b}{s_r \times se},$$

where b is a regression coefficient estimated by the multiple regression which discriminates between the RNA samples of interest, se is the unscaled standard error for b returned by the multiple regression and $s_r = \sqrt{a + s^2}$ where s is the residual standard deviation returned by the multiple regression, and a is the penalized term. L$\ddot{o}$nnstedt and Speed indicated the extension of the empirical Bayes B-statistic to general experimental designs.

Another direction in which the t-statistic can be generalized is to replace the sample mean $\bar{M}$ and sample standard deviation s with location and scale estimators which are robust against outliers. This extension is very useful for microarray data because it is impossible to guarantee or adjust for the data quality of every individual spot. The general idea of robust estimation is to replace $\bar{M}$ and s with values which behave very much like $\bar{M}$ and s when the data actually are normally distributed but which are insensitive to a small proportion of aberrant observations [49] [66]. For general microarray experiments, a robust multiple regression can be computed for each gene and a penalized t-statistic formed from the robust versions of b, s and se.

2.2.2 Statistical models for microarray experiments

The statistical model for t-statistics assumes that the preprocessed intensities are approximately normally distributed with variances homogeneous across the groups, i.e.

$$X_{gij} \sim N(\mu_{gj}, \sigma^2),$$

where $g(g = 1, \cdots, G)$ indexes the genes on the array, $j, (j = 1, 2)$ indicates the groups, and $i(i = 1, \cdots, n_j)$ is the samples. For each gene g, performed a t statistic to test $H_0 : \mu_{g1} = \mu_{g2}$. Then Γ^* is a set of genes that are statistically significant at some pre-set α level for the test, or is a set of h genes with the smallest p-values for some pre-set number h. Multiple comparison problem is a very important when determining Γ^*. In 1995, Benjamini and Hochberg [8] proposed controlling the false discovery rate (FDR) :

$$FDR = E[\frac{V}{R}|R > 0]P(R > 0),$$

where V is the number of hypotheses rejected while null is true, and R is the total number of hypotheses rejected. Storey and Tibshirani [92] proposed a modified version of the FDR, called the positive false discovery rate (pFDR):

$$pFDR = E[\frac{V}{R}|R > 0].$$

pFDR is especially appropriate for exploratory analysis in which one is interested in finding several significant results among many tests and at least one test is rejected. So in microarray data analysis, we use pFDR to ensure that false positive rate is not too high.

The variation of microarray data is so significant that we should not ignore when replications are small, but the manufacturers of microarray equipment do not stress the need for replication of studies. As a result, most current molecular genetic studies that use microarray technology are sometimes done without replication or with little replications $(3, 4, \cdots)$. The difficulty is that if replications are small the distribution of residuals of the intensities is often not normally distributed. That leads to many statistical models losing power. (e.g., Two-sample t test or ANOVA model) However, a great improvement of t statistics is a conditional t (Ct) suite of tests proposed by D. Amaratunga and J. Cabrera [5][6] to identify differentially expressed genes while the microarray experiment is with little replication. The statistical model of Ct is not assumed to be a normal distribution for normalized data X_{gij}. They found that when comparing two groups of microarray data to determine which genes are differentially expressed, t-statistics are usually used, but it has been observed that if the sample size per group is small (as it often is), the dependence between the t test statistic and the pooled standard error estimate leads to an excessively high false positive rate for low variance genes and an excessively high false negative rate for high variance genes. The problem is that the data are longer tailed than a normal distribution, the variability of a gene depends on its expression level, and the genes are co-dependent in clumps. The statistical model for Ct is followed:

Suppose X_{gij} are the log transformed and suitably normalized intensities . The intensities can be modeled as :

$$X_{gij} = \mu_{gj} + \sigma_g \epsilon_{gij}, \tag{2.1}$$

where μ_{gj} is the mean of the g^{th} gene in the j^{th} group, and σ_g^2 is the variance of the g^{th} gene. The treatment effect for the g^{th} gene is

$$\tau_g = \mu_{g2} - \mu_{g1}.$$

The random errors, $\epsilon_{gij} \sim F_g$, unspecified distribution with 0 mean and unit variance, with some correlation across genes. The t test statistic for testing $H_0 :\ \tau_g = 0$ for gene g is :

$$T_g = \frac{\bar{X}_{g2} - \bar{X}_{g1}}{s_g(1/n_1 + 1/n_2)^{1/2}}$$

where s_g, the pooled standard error, is :

$$s_g = \sqrt{\frac{(n_1 - 1)s_{1g}^2 + (n_2 - 1)s_{2g}^2}{n_1 + n_2 - 2}}.$$

The conventional approach would designate any gene whose $|T_g| > t_\alpha$, for some pre-set value of $\alpha(0 < \alpha < 0.5)$, as statistically significant at level α, with the critical value t_α, defined by

$$P(|T| > t_\alpha; H_0) = \alpha. \tag{2.2}$$

This would select a set of genes. For example, *mice* and *mice* 2: Two mouse data sets from toxicology experiments (Amaratunga and Cabrera (2004) [5]). These datasets correspond to typical toxicology experiments where a group of mice is treated with a toxic compound and the objective is to find genes that are differentially expressed against samples from untreated mice. The *mice* and *mice*2 are two of the data sets that consist $n_1 = n_2 = 4$ mice in the control and treatment groups and total number of genes are $G = 4077$ from *mice* and $G = 3434$ for *mice*2, respectively. They represent two examples of cDNA chips, the first one *mice* has a high proportion π of differentially expressed genes whereas *mice*2 has a much smaller π.

In both data sets, it is observed (Figure 2.1) that there is a dependence between $|T_g|$ and s_g substantial enough to seriously disrupt the performance of the t test, where $s_{1g},\ s_{2g}$ are the standard error group 1 and group 2, respectively. This happens even in the simple null case where F_g is a standard normal distribution, $\mu_{g1} = \mu_{g2}$ and $\sigma_g = \sigma$ for all genes g; then

$$T_g|s_g \sim N(0, \frac{\sigma^2}{s_g^2})$$

indicating that the variance of $T_g|s_g$ is inversely proportional to s_g^2. This explains the wedge effect which is the fact that the variance of $T_g|s_g$ is a strictly decreasing function of s_g. From the real data, the wedge effect exists even when σ_g^2 is not constant across genes and is in fact greater than when it is constant. The wedge effect is more significant when n_1 and n_2 are small. It seems that when replicates are small, the wedge effect is especially a cause for concern. Not only is the wedge effect very significant, but also the s_g estimates are quite unreliable. This leads to a high false positive rate for genes with low variability and a high false negative rate for genes with high variability. For number of replications is small, it would be preferable to examine the distribution of $|T|$ conditional on s. D. Amaratunga and J. Cabrera proposed a critical envelope, $t_\alpha(s_g)$, rather than a constant critical value, as

$$P(|T| > t_\alpha(s_g)|s_g; H_0) = \alpha$$

to assess significance. This is the idea of Ct approach. They developed an R-package DNAMR to detect the differentially expressed genes.

2.3 The Shrinkage Algorithm on Microarray data

2.3.1 Introduction to a statistical model for cDNA microarray

In this section, we generalized the statistical model in (2.1) by adding gene effects. Then the normalized intensities $\{X_{gij}\}$ of microarray data can be modeled as follows:

$$X_{gij} = \mu_g + \tau_{gi} + \sigma_g \epsilon_{gij}, \tag{2.3}$$

where μ_g and $\sigma_g^2, g = 1, \cdots, G$, are the effect and variance of the g-th gene respectively, τ_{gi} is the effect of the g-th gene in the i-th group $(i = 1, 2)$, and $j = 1, \cdots, n_i$ indexes the samples. This is the same model in Amaratunga and Cabrera (2004)[5](2006)[6]. The treatment effect of the g-th gene is :

$$\tau_g = |\tau_{g2} - \tau_{g1}|.$$

We assume that $\{\epsilon_{gij}\}$ are iid observations from an unknown distribution F and we assume that σ_g and τ_g are iid observations from unknown distributions F_σ and F_τ, respectively. F_σ is the distribution of the gene variances and F_τ is like to have mass zero with probability π representing the proportion of the genes that are not differentially expressed. In microarray data analysis, when the number of samples per group is very small $(3, 4, 5, \cdots)$ and residuals are subject to two constraints (sample mean $\bar{X} = 0$, sample standard deviation $s = 1$), then if we pool the residuals together, the estimated distribution $\hat{F}_\epsilon$ in (2.4) that is obtained gives a very poor estimate of the error distribution in the sense of Q-Q normal plot, even though the errors come from normal distribution Figure 2.2 (a).

$$\hat{F}_\epsilon = \text{ Empirical CDF}\{\hat{\epsilon}_{gij} = \frac{X_{gij} - \bar{X}_g}{s_g}, g \in S_G, i = 1, 2, j = 1, \cdots, n_i\}, \tag{2.4}$$

where $\bar{X}$ is the sample mean and s_g is the sample variance for gene g. We proposed a method to address this problem in section 2.3.2.

To illustrate the estimation of π, we apply our procedure for the mouse data sets *mice* and *mice*2. They represent two examples of cDNA chips, the first one *mice* has a high proportion π of differentially expressed genes whereas *mice*2 has a much smaller π.

2.3.2 Procedures to estimate F_ϵ, F_σ, and F_τ

The data from cDNA microarray experiments consists of suitably normalized intensities: X_{gij}, where $g(g = 1, \cdots, G)$ indicates the genes on the microarray, $i(i = 1, 2)$ indexes the groups, and $j(j = 1, \cdots, n_i)$ is the i-th mouse in the j-th group. The model we used in this section is (2.3), which are

$$X_{gij} = \mu_g + \tau_{gi} + \sigma_g \epsilon_{gij}.$$

If the sample sizes were bigger the unknown distributions could be readily estimated by their respective cdf's but for small sample sizes the cdf's would produce very biased estimators. In the remainder of this section we will provide three procedures to estimate the three distributions F_ϵ , F_τ , and F_σ , which try to overcome the biases induced by small sample size.

In the model step:

1. Estimating F_ϵ :

   In (2.3) when the number of samples per group is very small (3, 4, 5) and after residuals are subject to two constraints (sample mean $\bar{X} = 0$, sample standard deviation $s = 1$) then if we pool the residuals together, the empirical distribution that is obtained gives a very poor estimator of the error distribution F_ϵ.

   For example: Suppose we sample 1000 genes from N(0,1) with two groups of subjects of sizes 4 and 4. The empirical distribution of the residuals is not so close to the true error distribution (which is standard normal) which is shown in the Figure 2.2 (a). We also simulated the t-distribution with degrees of freedom 4 and the QQ-plot of the residual distribution is not so good which is shown in the Figure 2.2(c).

   One simple way to avoid this problem is to select a subset of genes S_G that have small absolute t-values (say below 1 or some threshold that gives a large set of numbers). For each gene in S_G, both samples are pooled together and normalized by subtracting the mean of genes and dividing by the standard deviation of genes. If the sample size per group is very small (3, 4, 5) instead of the sample mean and standard deviation it is much better to use bi-square for location and Huber proposal 2 for scale [49]. This will result in a table of residuals $\hat{\epsilon}_{gij},\ g \in S_G$. The error distribution F_ϵ is estimated by

   $$\hat{F}_\epsilon = \text{Empirical CDF}\{\hat{\epsilon}_{gij} = \frac{X_{gij} - B_g}{H_g} : g \in S_G, i = 1, 2, j = 1, \cdots, n_i\}, \quad (2.5)$$

   where B_g and H_g are the bi-square location estimator for gene g and scale estimator of Huber robust M-estimator (proposal 2) for gene g, respectively. The Figure

2.2 (b,d) shows the QQ-plot for the estimated error distribution on t-distribution. The improvement is very clear.

2. Estimating F_σ:

   We follow the method described in Amaratunga and Cabrera [5] [6]. They pointed out that the empirical distribution, $\hat{F}_\sigma$, of s_g is a very poor estimator of the distribution F_σ, because on average $\hat{F}_\sigma$ is much more scattered than F_σ. They proposed an estimate $\tilde{F}_\sigma$ of F_σ that shrinks $\hat{F}_\sigma$ toward its center and the bias can be corrected using a method initially proposed in Amaratunga and Cabrera [4]. It is also a version of the target estimation procedure of Cabrera and Fernhols [20]. The key concept is to estimate the function $h : [0,1] \longrightarrow [0,1]$ defined by $h(F_\sigma(x)) = \hat{F}_\sigma(x)$. Since h is strictly monotonic, it can be inverted in order to obtain an estimate $F_\sigma(x)$ are follows:

   (a) Assume that $\hat{F}_\sigma(x)$ is the true distribution of σ and draw a random sample, s^{*2}, from $\hat{F}_\sigma$.

   (b) Take a random sample (with replacement) of size N from $\hat{F}_\epsilon \ : \ r^*_{ij} \sim \hat{F}_\epsilon$ for $i = 1, \cdots, n_j, \ j = 1, 2$.

   (c) Combine these to form pseudo-data: $X^*_{ij} = s^* r^*_{ij}$.

   (d) Calculate the pooled standard error s^{**} for the pseudo-data $\{X^*_{ij}\}$.

   (e) Repeat steps (a)-(d) a large number (say 100,000) of times and record, for each iteration, the pair of values $\{s^{*2}, s^{**2}\}$.

   (f) Let $\hat{F}_{\sigma^*}(x)$ be the empirical distribution $\hat{F}_\sigma$ onto $\hat{F}_{\sigma^*}$. i.e.

   $$\hat{h}(y = \hat{F}_\sigma(x)) = \hat{F}_{\sigma^*}(\hat{F}_\sigma^{-1}(y))$$

   and then

   $$\hat{h}^{-1}(y) = \hat{F}_\sigma(\hat{F}_{\sigma^*}^{-1}(y)),$$

   where $\hat{F}^*_\sigma$ is the empirical cdf of s^{**2}. Hence based on this procedure, the bias-corrected estimator of F_σ can be obtained as:

   $$\tilde{F}_\sigma(x) = \hat{F}_\sigma(\hat{F}_{\sigma^*}^{-1}(\hat{F}_\sigma(x))). \tag{2.6}$$

Once bias-corrected edf, $\tilde{F}(x)$ is estimated, it can be used to generate the pooled standard errors in **1.1)** of the algorithm below.

3. Estimating F_τ:(determine the proportion of differential expressed genes)

We said earlier that τ_g is drawn from some distribution F_τ . We expect that F_τ has a mass at zero of probability $F_\tau(0) \geq 0$, which represents the genes that are not differentially expressed. In order to estimate the probability $P(\tau_g = 0)$ we apply an algorithm that will produce an estimator $\tilde{F}_\tau$ such the the

$$E_{\tilde{F}_\tau}(\hat{F}_\tau^*(t)) = \hat{F}_\tau(t),$$

where $\hat{F}_\tau^*(t)$ is the random variable representing the empirical cdf of τ^{**} at value t, which is constructed in following algorithm and $\hat{F}_\tau(t)$ represents the empirical cdf of actual observed value.

The algorithm is as follows:

Algorithm:

Step 1:

1.1) Draw a random sample, s^* , from $\tilde{F}_\sigma$, which is the bias-corrected edf estimate of $\hat{F}_\sigma$ in (2.6).

1.2) Estimate the error distribution F_ϵ with the empirical distribution $\hat{F}_\epsilon$ defined in (2.5).

1.3) Take a random sample (with replacement): $r_{gij} \sim \hat{F}_\epsilon$ for $i = 1, 2, j = 1, \cdots, n_i, g = 1, \cdots, N$.

1.4) Draw a sample τ_g^* from $\hat{F}_\tau(t) = I_{\{t \geq 0\}}$, where $I_{\{t \geq 0\}} = 1$ if $t \geq 0$ and $I_{\{t \geq 0\}} = 0$ if $t < 0$.

1.5) Construct the pseudo-data:$X_{g1j}^* = s_g * r_{g1j}, X_{g2j}^* = \tau_g^* + s_g * r_{g2j}$.

1.6) Reconstruct the distribution $F_{\hat{F}_\tau}^* = E(\hat{F}_\tau^* | \hat{F}_\tau)$, where $\hat{F}_\tau^*$ is the distribution of τ^{**} by pseudo-data: $\tau_g^{**} = |\bar{X}_{g2}^* - \bar{X}_{g1}^*|$.

1.7) Start by setting $\hat{F}_\tau^{(old)} = \hat{F}_\tau$.

1.8) Let $\hat{F}_\tau^{(new)} = \hat{F}_\tau(F^{*-1}_{\hat{F}_\tau^{(old)}}(\hat{F}_\tau))$.

1.9) Set $\hat{F}_\tau^{(old)} = \hat{F}_\tau^{(new)}$ and go to **1.3)**.

1.10) Iterate this procedure until convergence (approximately 100 iterations). At convergence we get our final estimate $\tilde{F}_\tau = \hat{F}_\tau^{(new)}$.

1.11) Give a cutoff point, say η, which is a 95% quantile of the final $\tilde{F}_\tau(t)$.

Step 2:

2.1) Repeat **1.4)-1.8)** using all original data X_{gij} and the estimated $\hat{F}_\tau$.

2.2) Get the estimated percentage of τ_g^{**} which is greater than $\eta \times 95\%$ quantile of standard normal.

Theorem 2.3.1. *At convergence the estimator $\tilde{F}_\tau$ is a fix point of the step in* **1.8)** *of the algorithm. That is $\tilde{F}_\tau = \hat{F}_\tau(F^{*-1}_{\tilde{F}_\tau}(\hat{F}_\tau))$, then we have*

$$E_{\tilde{F}_\tau}(\hat{F}_\tau^*) = \hat{F}_\tau. \tag{2.7}$$

Proof. If the algorithm converges, then $\tilde{F}_\tau = \hat{F}_\tau(F^{*-1}_{\tilde{F}_\tau}(\hat{F}_\tau))$. Thus

$$\begin{aligned}
& \hat{F}_\tau \circ \tilde{F}_\tau^{-1} \circ \hat{F}_\tau = F^*_{\tilde{F}_\tau} = E(\tilde{F}_\tau | \tilde{F}_\tau) = \tilde{F}_\tau \\
\Rightarrow \quad & \hat{F}_\tau \circ \tilde{F}_\tau^{-1} = \tilde{F}_\tau \circ \hat{F}_\tau^{-1} \\
\Rightarrow \quad & (\hat{F}_\tau \circ \tilde{F}_\tau^{-1})^2 = I \\
\Rightarrow \quad & \hat{F}_\tau \circ \tilde{F}_\tau^{-1} = I \\
\text{or} \quad & \hat{F}_\tau \circ \tilde{F}_\tau^{-1} = -I(\text{ impossible, since } \hat{F}_\tau, \tilde{F}_\tau \geq 0) \\
\text{Hence,} \quad E_{\tilde{F}_\tau}(\hat{F}_\tau^*) \quad & = E_{\tilde{F}_\tau}(\tilde{F}_\tau) = \tilde{F}_\tau = \hat{F}_\tau.
\end{aligned}$$

□

Remark 2.3.2. *Base on our simulations, the algorithm converges to a fix point distribution $\tilde{F}_\tau$ in at most 100 iterations and very fast.*

Remark 2.3.3. *At convergence, $\tilde{F}_\tau$ is approximately the same as $\hat{F}_\tau$ and $\hat{F}_\tau^*$ is also approximately the same as $\tilde{F}_\tau$, such that we have a fix point result $E_{\tilde{F}_\tau}(\hat{F}_\tau^*) = \hat{F}_\tau$.*

Remark 2.3.4. *This is a two-stage estimation method. We split data into two pieces. One is non-informative data whose t statistics are less than some threshold and it produces a good estimation of the error distribution. The other is the informative data, we use our algorithm to estimate the distribution of* τ_g *to get a limit convergent distribution, which is our target distribution* $\tilde{F}_\tau$.

Remark 2.3.5. *Amaratunga and Cabrera(2004) [5] used "target estimation" techniques to obtain a bias-corrected estimate* $\tilde{F}_\sigma$ *of* $\hat{F}_\sigma$ *and we then generate this concept to obtain a limit distribution* $\tilde{F}_\tau$ *and show this limit distribution is the distribution we try to estimate. This is a very useful extension.*

2.3.3 Performance assessment

To assess the performance of this method, we simulate data points, which are identically and independently distributed.

1. $X_{gij} \sim F(\tau_g, \sigma^2)$, where $G = 10000, n_1 = n_2 = 4$ and we assume that $G_{sig} = 1000, \cdots, 9000$ of G genes were differentially expressed between two groups and their difference was δ, i.e. $\tau_g = \delta(\delta = 1, 2)$ for all $g = 1, \cdots, G_{sig}$, and $\tau_g = 0$ otherwise.

2. $X_{gij} \sim F(\tau_g, \sigma_g^2)$, where $G = 10000, n_1 = n_2 = 4$ and we assume that $G_{sig} = 1000, \cdots, 9000$ of G genes were differentially expressed between two groups and their difference was $\delta = 1, 2$, for all $g = 1, \cdots, G_{sig}$, and $\tau_g = 0$ otherwise and σ_g^2 are chi-square distributed with degrees of freedom 3. We calibrate the mean of σ_g^2 to 1. i.e. $\sigma_g^2/3$.

We simulated many distributions F, which could be normal, t, gamma, or lognormal with mean μ_g and variance σ or σ_g.

Base on simulation studies, we compare our method to permutation tests and t-tests using a threshold of 0.05 to determine significance. These two methods are standard in biological applications. Our method is much more accurate than other two methods (Table 2.1-2.4, Fig 2.3-2.11). Each cell in the table is the mean (standard deviation) based on 10 times simulations on each condition. In Figure 2, the straight line represents the true values and the red line is obtained by the smooth spline function. Because, the pFDR emphasizes the fact that an adjustment is only necessary when there are positive finding, we calculate the pFDR to our method in different values of $\lambda \in \{0.1, \cdots, 0.9\}$ (Table 2.5-2.6). We find that the pFDR decreases when the true λ increases and it belows 40% when difference $\delta = 2$ and 60% when difference $\delta = 1$. But when $\lambda > 0.5$, the pFDR is less than 10% when difference $\delta = 2$, the pFDR is less than 20% when difference $\delta = 1$. That is an acceptalbe result, becasue we get a very accurate estimates and the pFDR is not too high.

2.4 Discussion and extensions

In this section, the standard analysis of microarray data are introduced and we proposed a statistical model for two-channel cDNA microarray data with little replicates and an algorithm for estimating the proportion of differentially expressed genes in microarray experiments. We also show that the estimator of the distribution converges to a fix point which is a limit distribution. We perform a simulation study to check the performance of our estimate and it is shown to be "satisfactory" and we show that our method has better performance than other alternatives such as permutation tests and standard two-sample t-test. The simulations are performed under normal and gamma error distribution and with constant variances and chi-square variances. In addition we illustrate the method with real data examples on *mice* and *mice*2 (Table 2.7). In the real data examples we obtain estimates of the proportion of significant genes that are more realistic than those produced by the other methods. Hence, this algorithm gives us more accurate prediction to detect differential genes. This same method is generally expendable to other more complicated modeling procedures such as the one-way ANOVA F-test and other linear models. The same model is used and the same ideas are easily expendable into the GO issues by modeling the p-values and getting a null distribution that will be used to detect differentially expressed gene network and subsets.

δ	true λ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	t-test	0.066 (0.002)	0.085 (0.003)	0.103 (0.002)	0.119 (0.003)	0.136 (0.003)	0.154 (0.005)	0.171 (0.004)	0.186 (0.004)	0.207 (0.004)
1	Permutation test	0.039 (0.002)	0.051 (0.002)	0.063 (0.002)	0.073 (0.003)	0.085 (0.003)	0.096 (0.003)	0.107 (0.003)	0.116 (0.003)	0.130 (0.003)
1	New method	0.071 (0.058)	0.163 (0.091)	0.226 (0.084)	0.282 (0.072)	0.304 (0.049)	0.422 (0.081)	0.473 (0.105)	0.479 (0.145)	0.518 (0.120)
2	t-test	0.109 (0.002)	0.171 (0.002)	0.234 (0.003)	0.294 (0.003)	0.354 (0.003)	0.415 (0.004)	0.474 (0.005)	0.534 (0.004)	0.595 (0.004)
2	Permutation test	0.074 (0.003)	0.120 (0.002)	0.168 (0.002)	0.214 (0.003)	0.259 (0.004)	0.305 (0.003)	0.350 (0.005)	0.397 (0.005)	0.442 (0.004)
2	New method	0.087 (0.020)	0.196 (0.022)	0.321 (0.034)	0.431 (0.033)	0.522 (0.030)	0.635 (0.045)	0.720 (0.034)	0.823 (0.022)	0.923 (0.021)

Table 2.1: Normal(0,1)

δ	true λ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	t-test	0.066 (0.001)	0.087 (0.004)	0.110 (0.002)	0.134 (0.004)	0.157 (0.003)	0.180 (0.004)	0.204 (0.003)	0.227 (0.004)	0.252 (0.003)
1	Permutation test	0.045 (0.002)	0.060 (0.002)	0.077 (0.002)	0.095 (0.003)	0.112 (0.002)	0.129 (0.003)	0.148 (0.004)	0.163 (0.003)	0.182 (0.002)
1	New method	0.079 (0.072)	0.145 (0.096)	0.153 (0.040)	0.301 (0.069)	0.327 (0.062)	0.436 (0.119)	0.513 (0.138)	0.576 (0.138)	0.577 (0.116)
2	t-test	0.105 (0.002)	0.172 (0.002)	0.237 (0.003)	0.303 (0.003)	0.370 (0.004)	0.435 (0.003)	0.498 (0.003)	0.565 (0.006)	0.630 (0.003)
2	Permutation test	0.080 (0.003)	0.134 (0.002)	0.186 (0.003)	0.241 (0.004)	0.295 (0.003)	0.347 (0.004)	0.400 (0.004)	0.451 (0.005)	0.508 (0.005)
2	New method	0.111 (0.027)	0.207 (0.034)	0.311 (0.032)	0.413 (0.030)	0.514 (0.025)	0.609 (0.022)	0.712 (0.017)	0.811 (0.018)	0.914 (0.015)

Table 2.2: N(0,a),$a \sim \chi^2_{(3)}/3$

δ	true λ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	t-test	0.067 (0.002)	0.094 (0.004)	0.123 (0.003)	0.150 (0.004)	0.178 (0.004)	0.207 (0.004)	0.233 (0.004)	0.264 (0.004)	0.292 (0.003)
1	Permutation test	0.053 (0.001)	0.075 (0.002)	0.099 (0.003)	0.120 (0.003)	0.143 (0.004)	0.168 (0.003)	0.190 (0.003)	0.213 (0.006)	0.237 (0.003)
1	New method	0.059 (0.043)	0.151 (0.035)	0.225 (0.075)	0.310 (0.062)	0.321 (0.099)	0.377 (0.110)	0.482 (0.094)	0.504 (0.119)	0.626 (0.107)
2	t-test	0.108 (0.002)	0.177 (0.002)	0.246 (0.003)	0.313 (0.003)	0.381 (0.005)	0.450 (0.003)	0.521 (0.004)	0.588 (0.005)	0.657 (0.005)
2	Permutation test	0.090 (0.003)	0.151 (0.002)	0.212 (0.002)	0.272 (0.003)	0.330 (0.004)	0.391 (0.004)	0.454 (0.003)	0.514 (0.005)	0.576 (0.004)
2	New method	0.126 (0.048)	0.232 (0.045)	0.310 (0.024)	0.417 (0.020)	0.515 (0.023)	0.613 (0.010)	0.712 (0.015)	0.802 (0.014)	0.912 (0.013)

Table 2.3: $Gamma(1, 1)$

δ	true λ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	t-test	0.065 (0.003)	0.086 (0.003)	0.109 (0.005)	0.130 (0.004)	0.153 (0.003)	0.174 (0.004)	0.197 (0.003)	0.218 (0.004)	0.243 (0.002)
1	Permutation test	0.043 (0.002)	0.058 (0.002)	0.075 (0.004)	0.090 (0.003)	0.106 (0.003)	0.122 (0.003)	0.138 (0.004)	0.153 (0.003)	0.170 (0.003)
1	New method	0.074 (0.060)	0.141 (0.100)	0.208 (0.065)	0.212 (0.074)	0.319 (0.080)	0.368 (0.091)	0.490 (0.133)	0.530 (0.128)	0.641 (0.084)
2	t-test	0.112 (0.002)	0.177 (0.002)	0.241 (0.002)	0.309 (0.005)	0.373 (0.003)	0.440 (0.003)	0.507 (0.004)	0.575 (0.005)	0.639 (0.005)
2	Permutation test	0.083 (0.002)	0.136 (0.003)	0.190 (0.002)	0.246 (0.004)	0.298 (0.003)	0.352 (0.004)	0.408 (0.004)	0.461 (0.006)	0.517 (0.006)
2	New method	0.113 (0.030)	0.205 (0.013)	0.309 (0.028)	0.411 (0.027)	0.516 (0.022)	0.610 (0.016)	0.718 (0.027)	0.811 (0.017)	0.918 (0.013)

Table 2.4: t_5

true λ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\delta = 1$	0.5471 (0.1679)	0.3768 (0.1371)	0.3823 (0.0537)	0.2372 (0.0535)	0.2372 (0.0535)	0.1924 (0.0354)	0.1486 (0.0363)	0.0860 (0.0209)	0.0482 (0.0131)
$\delta = 2$	0.1963 (0.0753)	0.1924 (0.0741)	0.2416 (0.0876)	0.1533 (0.0393)	0.1215 (0.0406)	0.0965 (0.0242)	0.0841 (0.0255)	0.0601 (0.0093)	0.0465 (0.0112)

Table 2.5: pFDR for our method with Normal(0,1) error distribution

true λ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\delta = 1$	0.634 (0.069)	0.480 (0.060)	0.375 (0.060)	0.323 (0.040)	0.233 (0.053)	0.185 (0.048)	0.102 (0.018)	0.094 (0.017)	0.047 (0.0135)
$\delta = 2$	0.325 (0.099)	0.226 (0.054)	0.167 (0.042)	0.139 (0.022)	0.119 (0.020)	0.107 (0.016)	0.074 (0.017)	0.063 (0.014)	0.037 (0.0047)

Table 2.6: pFDR for our method with $Normal(0, \sigma^2), \sigma^2 \sim \chi^2_{(3)}/3$ error distribution

Estimated π	$Mice$	$Mice2$
$t - test$	0.245	0.499
$Permutation\ test$	0.220	0.443
$New\ method$	0.107	0.363

Table 2.7: Results for the three methods applied to two real examples from toxicology

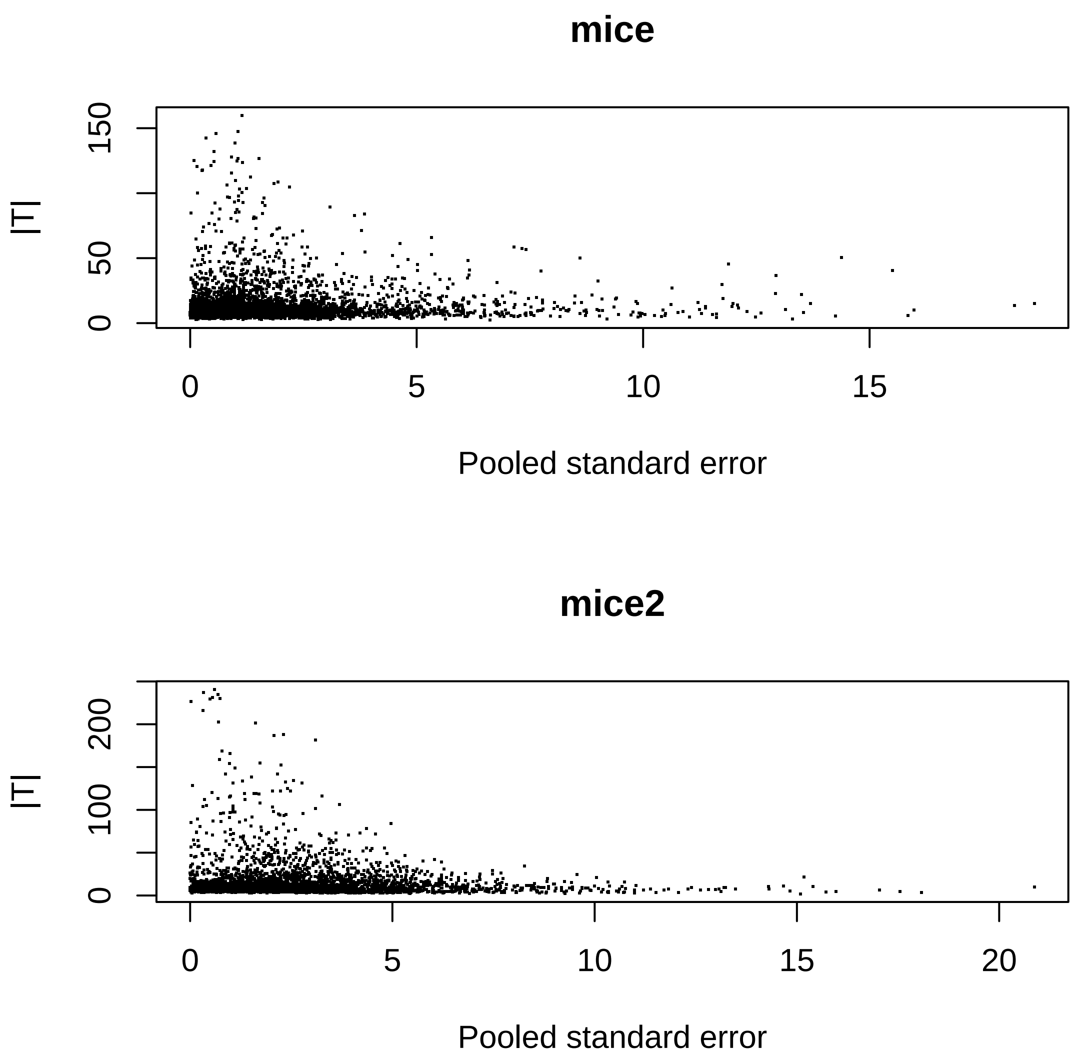


Figure 2.1: Pooled standard error depends on t statistic

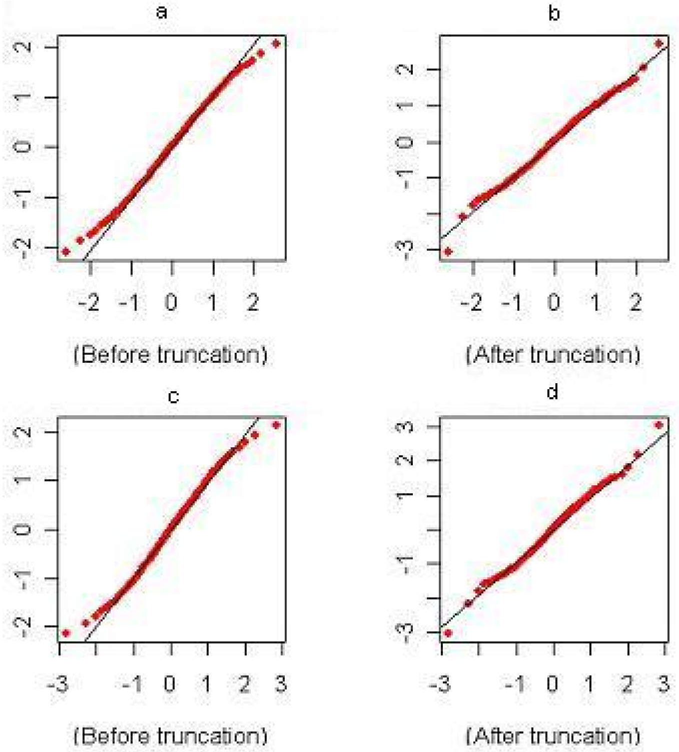


Figure 2.2: Figure a: the QQ-plot of residuals from N(0,1) before truncated; figure b: the QQ-plot of residuals from N(0,1) after truncated; figure c: the QQ-plot of residuals from t_4 before truncated; figure d: the QQ-plot of residuals from t_4 after truncated.

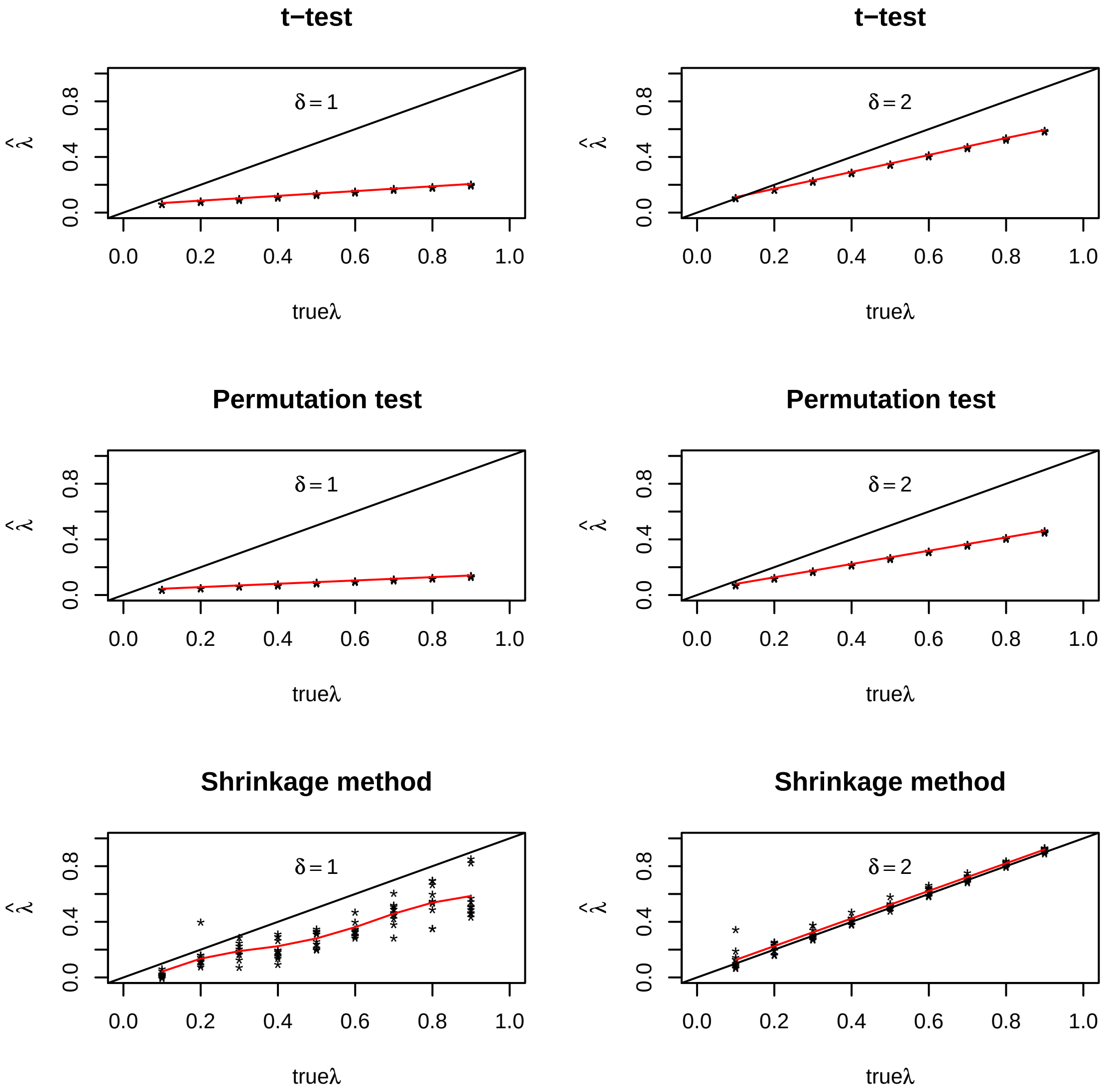


Figure 2.3: Standard normal distribution

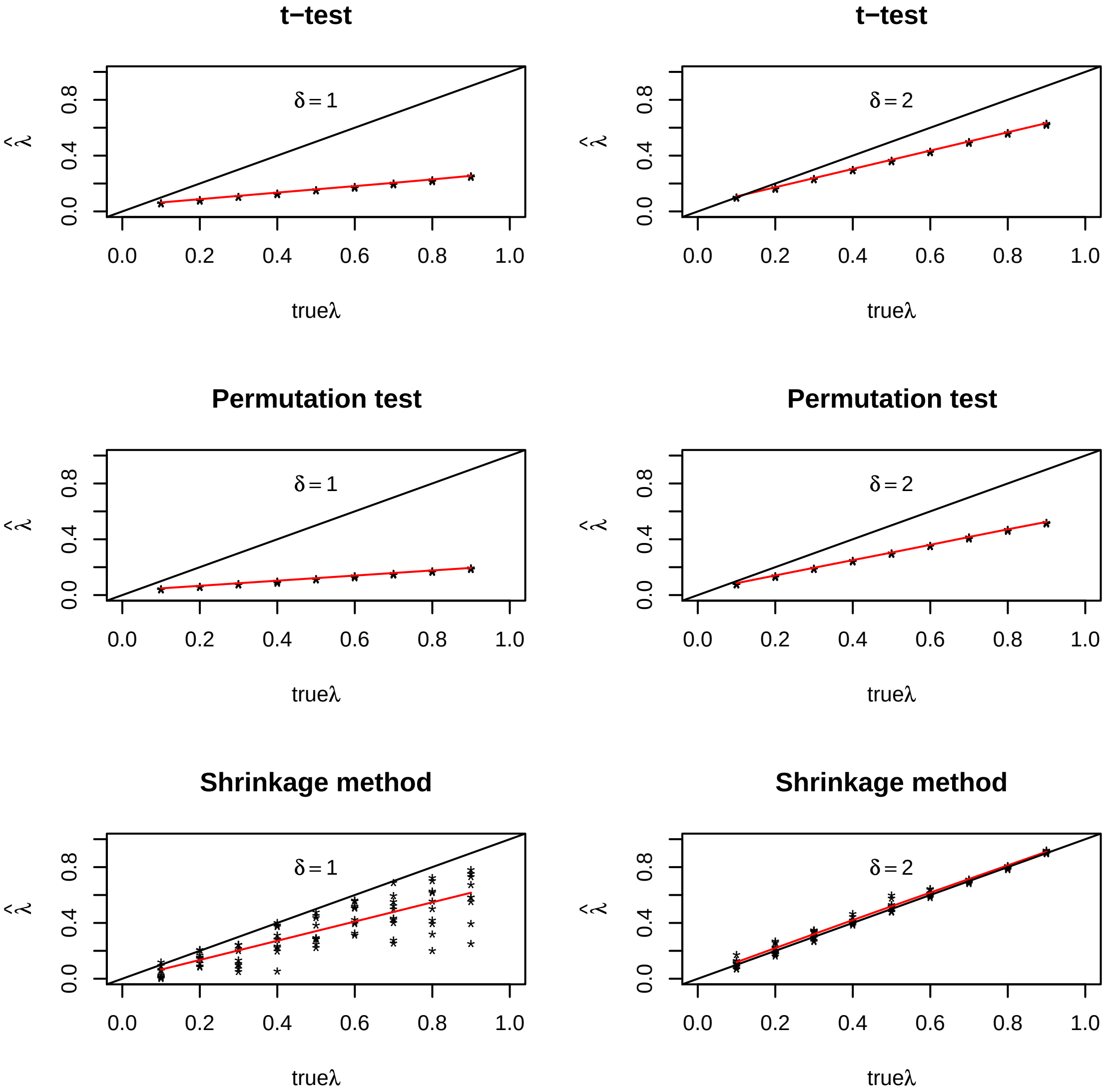


Figure 2.4: N(0,a) with $a \sim \chi^2_{(3)}/3$

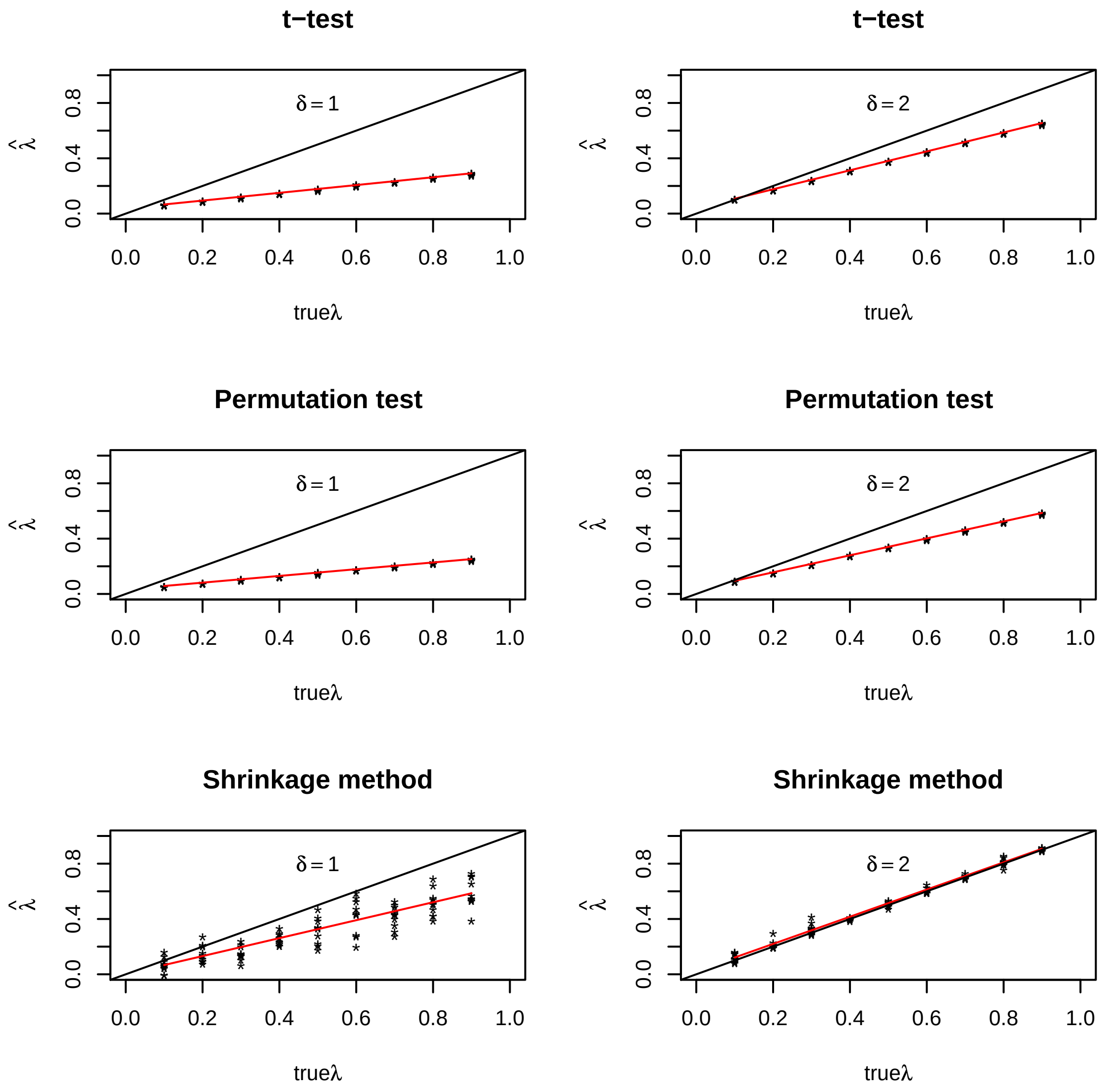


Figure 2.5: Gamma(1,4)

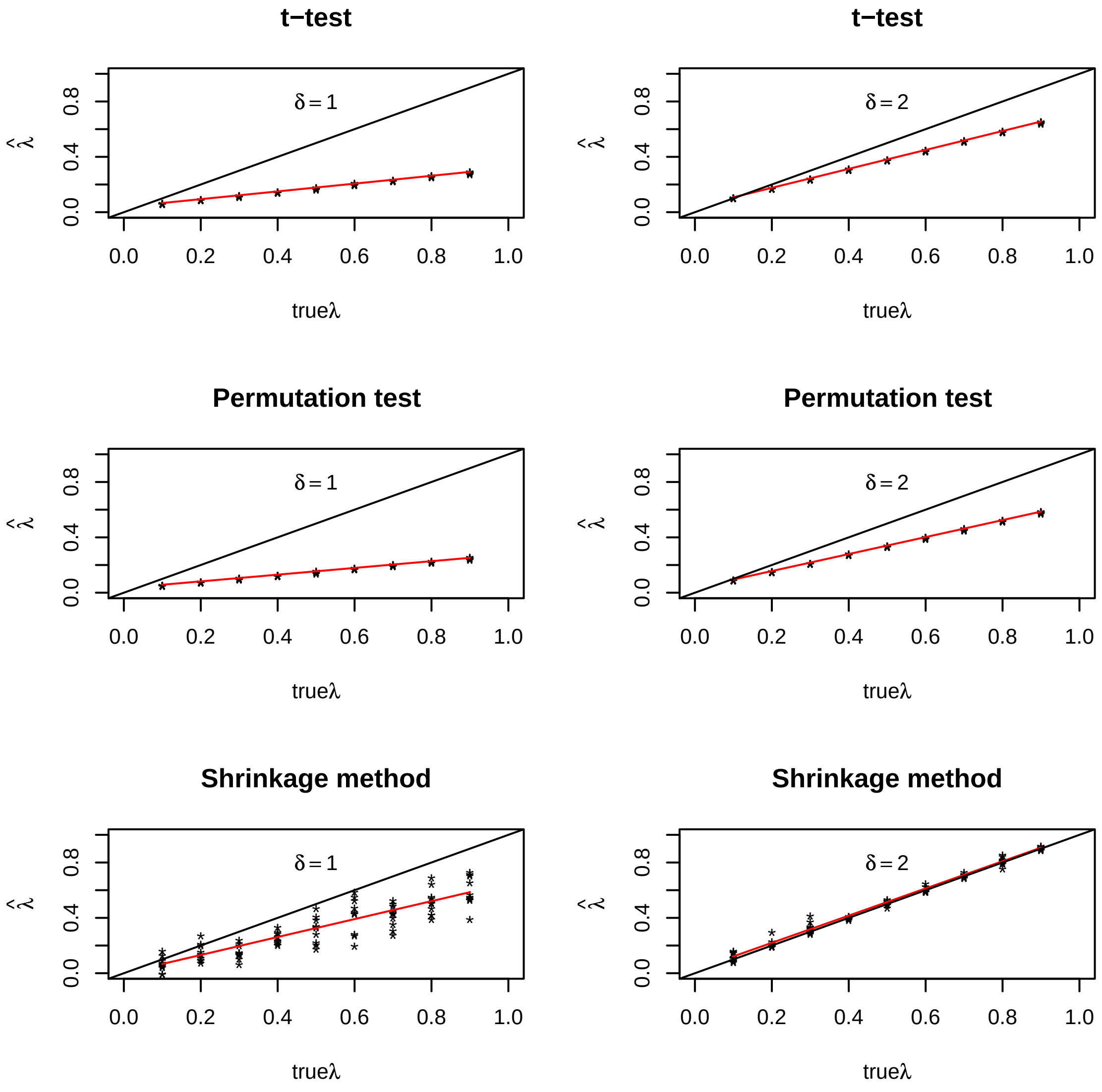


Figure 2.6: Gamma(1,a) with $a \sim \chi^2_{(3)}/3$

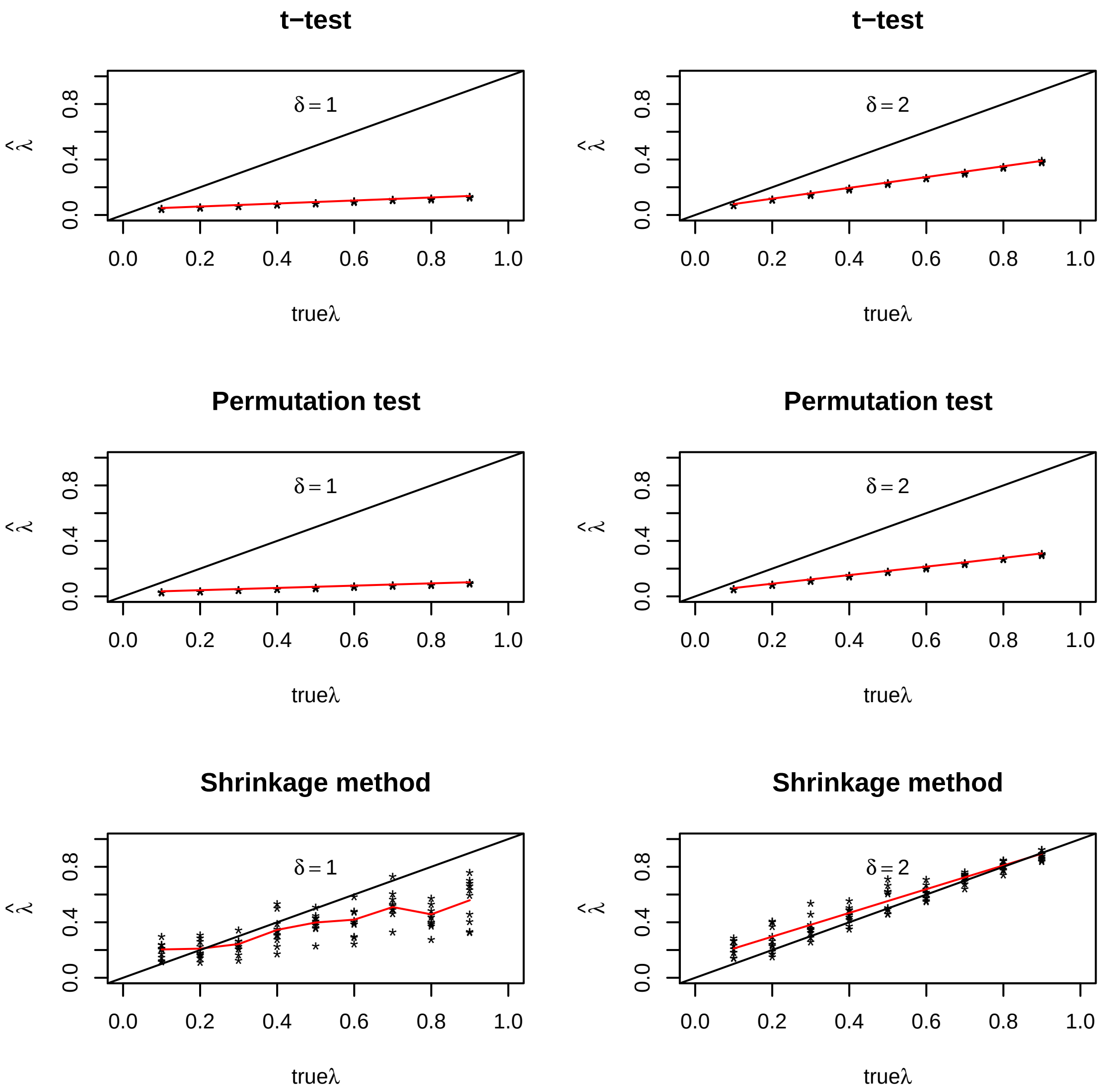

t−test
δ = 1
t−test
δ = 2
Permutation test
δ = 1
Permutation test
δ = 2
Shrinkage method
δ = 1
Shrinkage method
δ = 2
$\hat{\lambda}$
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
trueλ

Figure 2.7: t_3

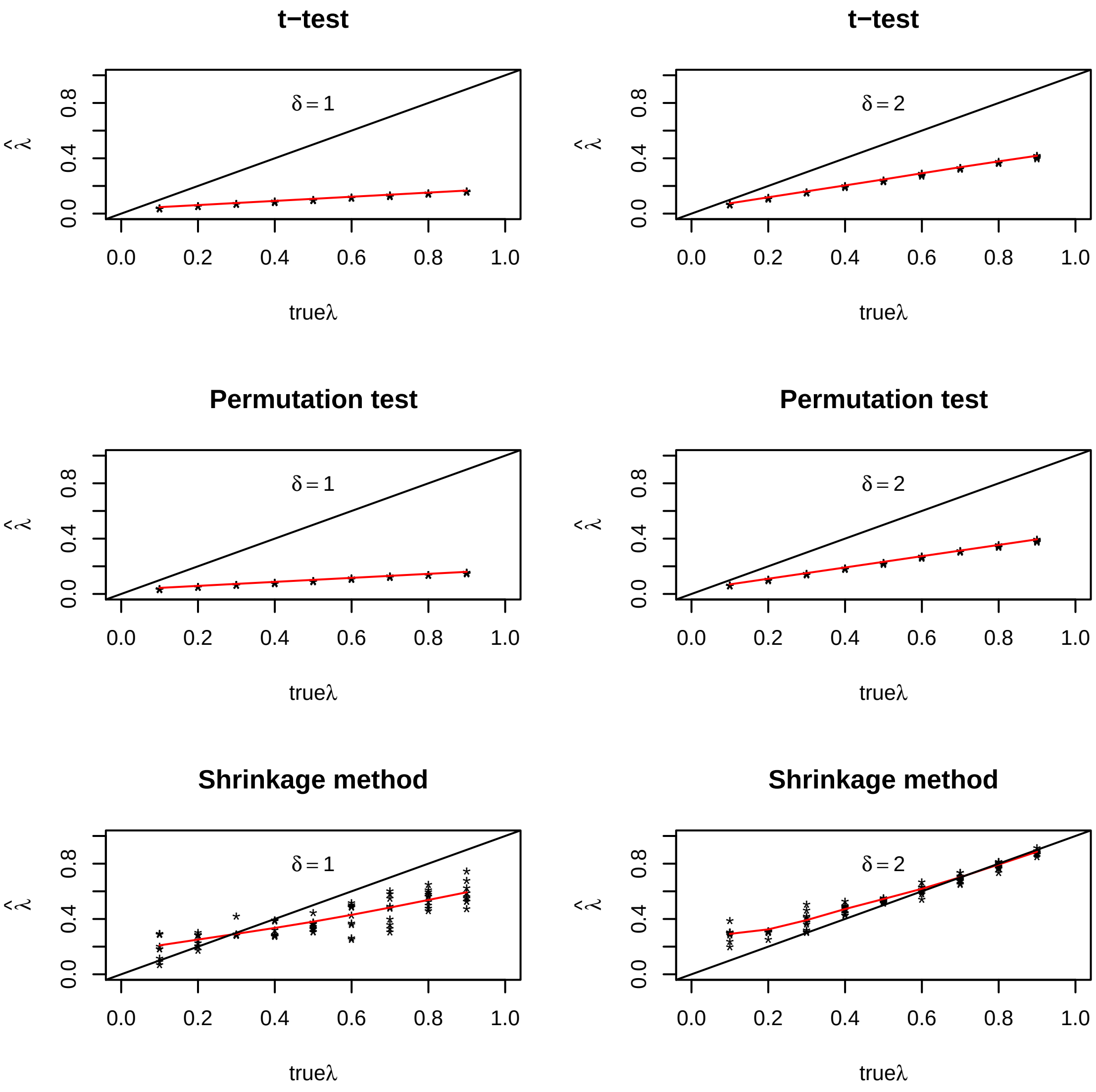


Figure 2.8: Lognormal(0,1)

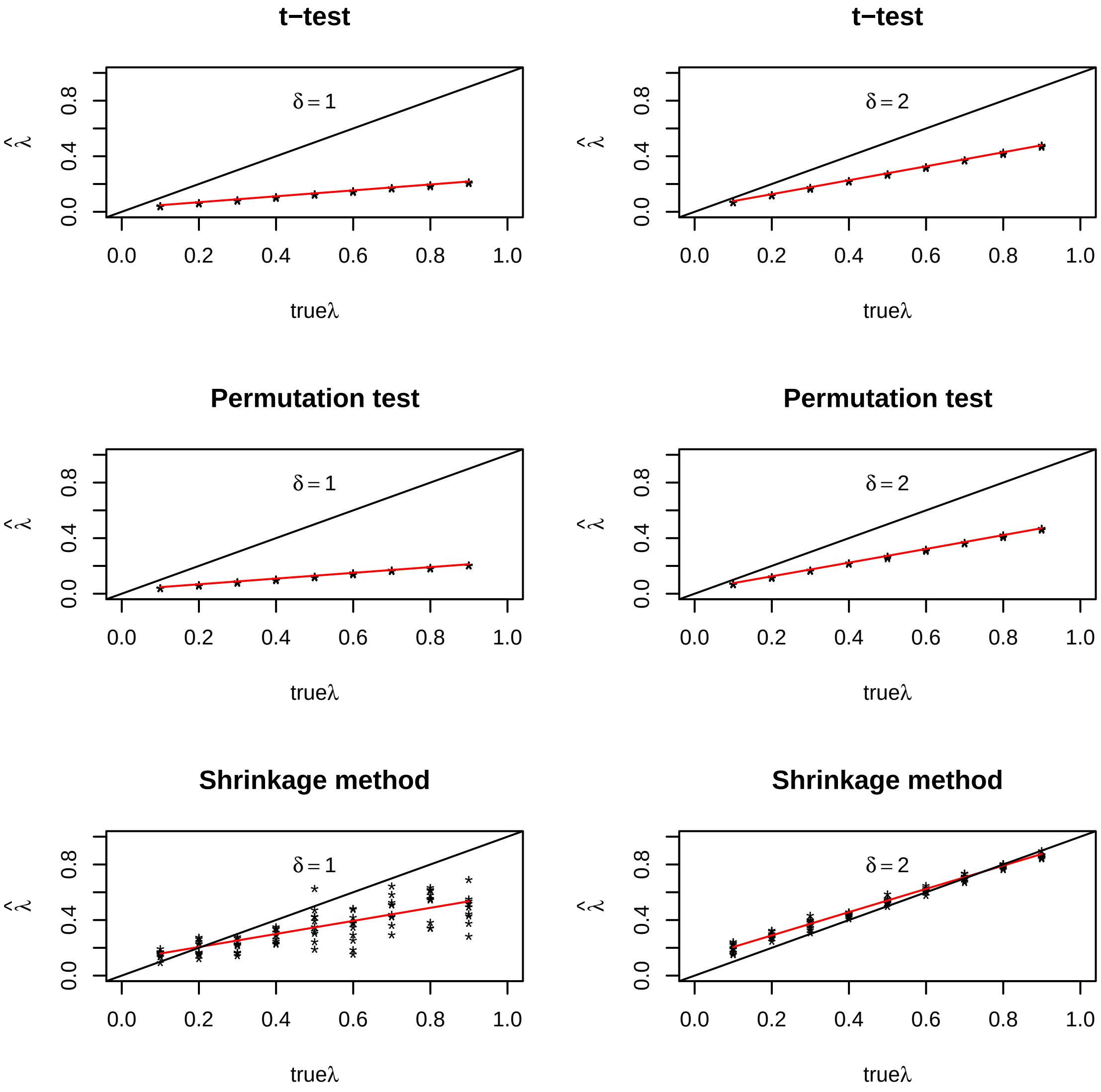


Figure 2.9: Lognormal (0,a), $a \sim \chi^2_{(3)}/3$

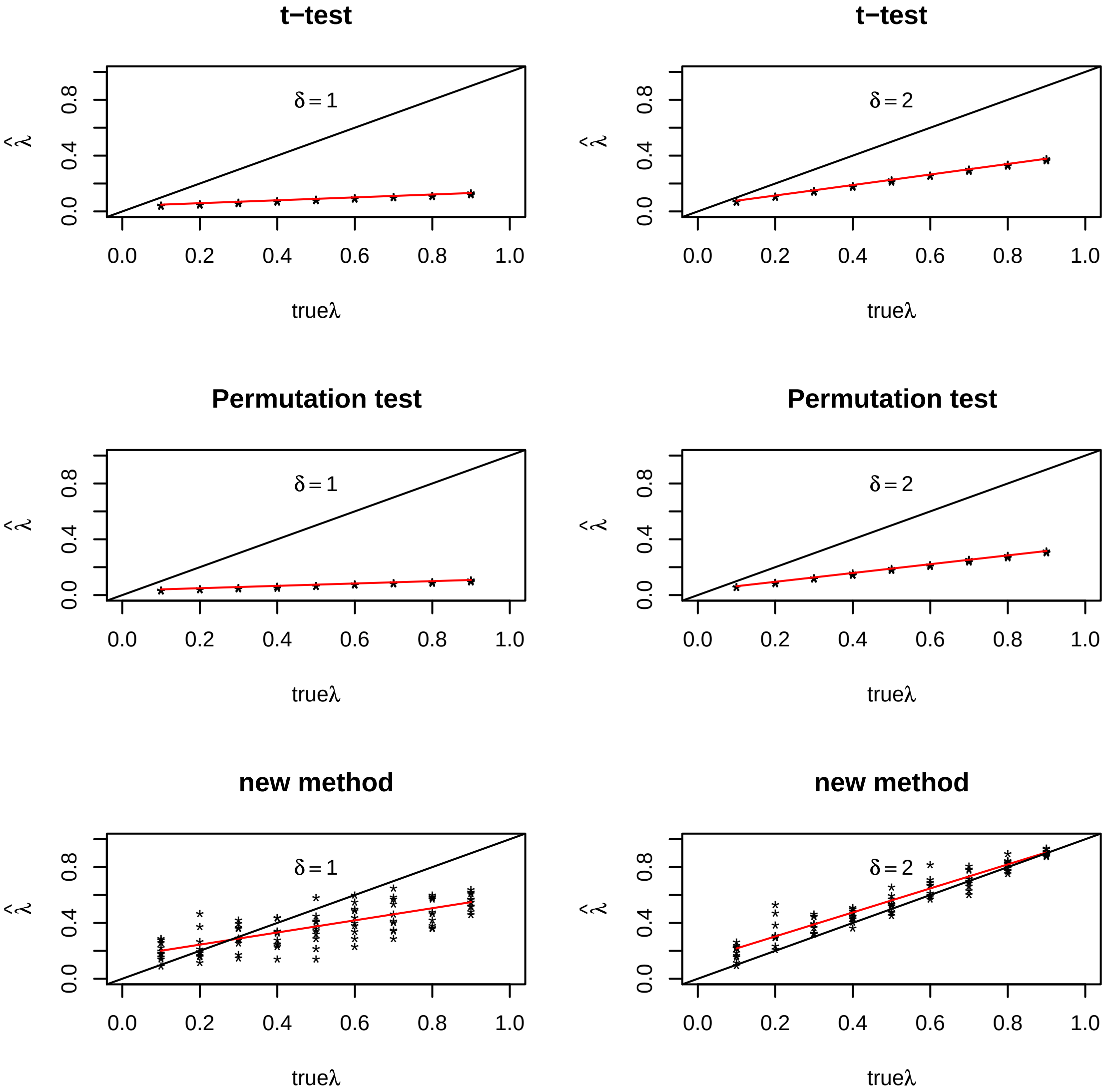


Figure 2.10: t_a with $a \sim \chi^2_{(3)}/3$

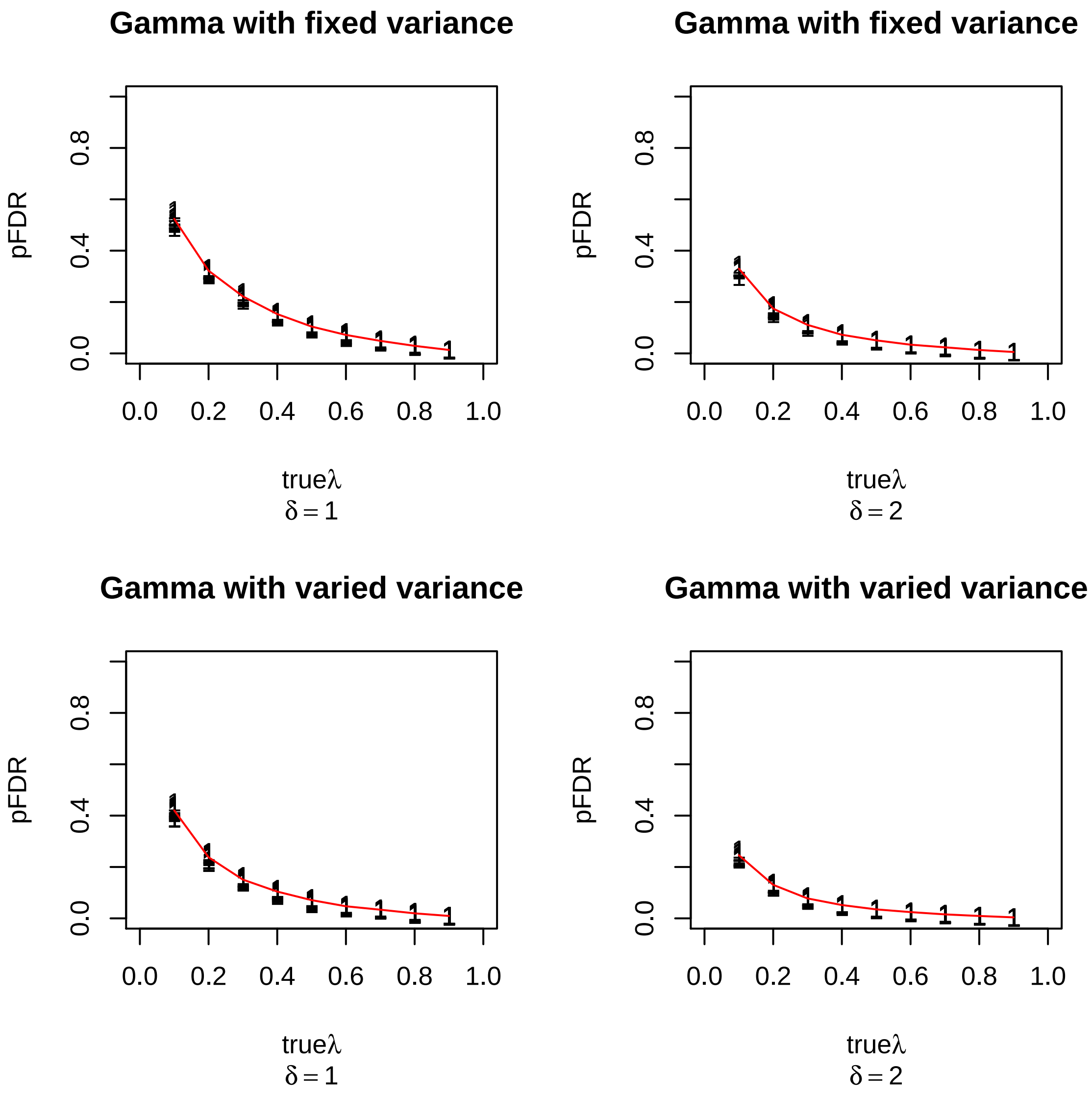


Figure 2.11: Gamma

Chapter 3

The estimations on error distributions

3.1 Introduction to estimation on error distributions and problems

It is well known among researchers that the data generated by microarray experiments is not normally distributed. The raw data is very skewed and for that reason that data is usually log-transformed, but in most cases, it is unlikely that either the log or any other transformation will produce normal data.

The second issue is that the number of samples is often very small such as $n = 3, 4, 5, \cdots$. The small sample sizes make some of the resampling statistical techniques like permutation tests or bootstrap limited. For example, if you generate two groups of 3 samples in each group, then for each gene there are only at most 10 absolute t-statistics ($|t|$) and hence the sampling distribution of $|t|$ does not have enough values to perform any credible inference.

The objective of this section is to propose methods for estimating the error distribution with enough accuracy that we are able to perform credible simulations and obtain the approximation of the sampling distribution that enables the performance of reliable inference procedures. We introduce two different methodologies. The first one is a non-parametric method to estimate the error distribution. The second one is an approximation of the tails of the error distribution using the t-family for symmetric error distribution and the stable-family for asymmetric error distribution. The second method is a simpler method and in practice may have similar performance with the first one.

The normalized intensities $\{X_{gij}\}$ of microarray data can be modeled as follows:

$$X_{gij} = \mu_g + \tau_{gi} + \sigma_g \epsilon_{gij}. \tag{3.1}$$

where μ_g and $\sigma_g^2, g = 1, \cdots, G$, are the effect and variance of the g^{th} gene respectively, τ_{gi} is the effect of the g^{th} gene in the i^{th} group ($i = 1, 2$), and $j = 1, \cdots, n_i$ indexes the samples. This is the same model in Amaratunga and Cabrera (2006) [6]. The treatment effect of the g^{th} gene is :

$$\tau_g = |\tau_{g2} - \tau_{g1}|$$

We assume that $\{\epsilon_{gij}\}$ are iid observations from an unknown distribution F and we assume that σ_g and τ_g are iid observations from unknown distributions F_σ and F_τ, respectively. F_σ is the distribution of the gene variances and F_τ is like to have mass zero with probability π representing the proportion of the genes that are not differentially expressed. In microarray data analysis, when the number of samples per group is very small $(3, 4, 5, \cdots)$ and residuals are subject to two constraints (sample mean $\bar{X} = 0$, sample standard deviation $s = 1$), then if we pool the residuals together, the estimated distribution that is obtained gives a very poor estimate of the error distribution which is a normal distribution. The idea to solve this problem have been discussed in chapter 2. The key part of the idea to solve this problem is selecting a subset of genes S_G that have small absolute t-values (say below 1 or some threshold that gives a large set of numbers), for each gene in S_G, both samples are pooled together and normalized by subtracting the sample mean(location parameter) and dividing by sample standard deviation(scale parameter). This gives a table of residuals $\{\hat{\epsilon}_{gij} = \frac{X_{gij} - \bar{X}_g}{s_g}, g \in S_G\}$. The error distribution F_ϵ is estimated by

$$\hat{F}_\epsilon = \text{ Empirical CDF}\{\hat{\epsilon}_{gij}, g \in S_G, i = 1, 2, j = 1, \cdots, n_i\}. \tag{3.2}$$

The main reason to combine the residuals is to double the sample size so that we can estimate the error distribution more accurate. In this chapter, we compare the performance of $\hat{F}_\epsilon$ among location and scale estimates as stated in section 3.2. By the definition of Q-Q normal plot, if the residual distribution is perfectly normally distributed, all points of Q-Q normal plot will be on the line $y = x$. In order to evaluate the performance of Q-Q plot among each residual distribution, we calculate the sum of squared projection distance(SSPD) from each point to $y = x$. The projection distance is the minimum of distance between points and $y = x$. The performance of two estimates then are compared base on SSPD.

3.2 Nonparametric methods

One of the popular methods to estimate the distribution is estimating location and scale parameters. In this section, we focus on this method to estimate the error distribution F_ϵ of DNA gene expression data. We consider 5 common used estimators for both location and scale. Location estimators are (1) mean, (2) median, (3) bisquare (Tukey's biweight robust estimator), (4) Huber, and (5)Hubers (Huber proposal 2 M-estimator) [97]. Scale estimators are (1) standard deviation, (2) MAD, (3) scale.tau (Huber tau estimate of scale), (4) scale.a (bisquare A-estimate of scale), and (5) Hubers(Huber M-estimator with proposal 2) [97]. Among 5 both scale and location estimators, we compare them on simulation studies. We found that when the location estimator is Bisquare and the scale is the Huber proposal 2 gives better estimates than other estimates; that is, it has smaller SSPD.

3.3 Parametric methods

3.3.1 t-distribution family to estimate symmetric error distribution

Lets first consider the case when F_ϵ is approximately symmetric. One idea that we considered is to approximate F_ϵ by a t-distribution family with unknown degrees of freedom that are to be estimated from the data. For the purpose of simulation, we do not concern with the shape of the center of the distribution, but only with the tails. Therefore, the t-family is a reasonable set of distributions to approximate F_ϵ.

3.3.2 Letter values

Letter values display is one of methods of exploratory data analysis. It starts a batch of data $X_1, X_2, \cdots, X_n$ and sorts it as order statistic $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$, where $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. The depth of each data value is defined as the value's position in an enumeration of values that starts at the nearer end of the data. The median(M) of the data $= \frac{X_{(k)}+X_{(k+1)}}{2}$ if $n = 2k$ or $X_{(k+1)}$ if $n = 2k+1$. So the depth of median $depth(M) = (n+1)/2$ if $n = 2k+1$, or $n/2$ if $n = 2k$. Then the depth of fourth $depth(F) = (depth(M){+}1)/2$. The lower or upper letter value of $depth(F)$ is the smaller or bigger value which has depth $depth(F)$, denote ad LLV(F) or ULV(F), respectively. The 5-number summaries(letter values) are Minimum, LLV(F) , Median, ULV(F), and Maximum. The depth of eighth $depth(E) = (depth(F) + 1)/2$ and so on. We tag the letters $M, F, E, D, C, B, A, Z, Y, X$ as median, fourth, eighth, so on. For example, the upper letter values of standard normal distribution are $ULV(M) = 0, ULV(F) = 0.67, ULV(E) = 1.15, ULV(D) = 1.53, ULV(C) = 1.86, ULV(B) = 2.15, ULV(A) = 2.42, ULV(Z) = 2.66, ULV(Y) = 2.89$, and $ULV(X) = 3.10$ that correspond to the $1 - 2^{-k}th$ quantile of standard normal distribution, where $k = 1, 2, 3, \cdots, 10$ and the upper letter values of t distribution with 2 degrees of freedom are $ULV(M) = 0, ULV(F) = 0.82, ULV(E) = 1.60, ULV(D) = 2.55, ULV(C) = 3.81, ULV(B) = 5.51, ULV(A) = 7.89, ULV(Z) = 11.20, ULV(Y) = 15.87, ULV(X) = 22.38$. For example, the lower fourth letter value for standard normal is LLV(F)$= -0.67$ and upper fourth letter value is ULV(F)$= 0.67$. The spread of depths is defined as $sp = ULV - LLV$. So the fourth spread of normal is $F_{sp} = ULV(F) - LLV(F) = 0.67 - (-0.67) = 1.34$.

In order to estimate the distribution of equivalent t-distribution, we use a method comparing the letter values of the data to a t-family. We calculate the spread of depths from standard normal distribution (Figure 3.1) and t-distribution with degrees of freedom $10, 5, 2$, and 1, (t_{10}, t_5, t_2, t_1). Comparing spread ratio of t_{10}, t_5, t_2, t_1 by normal(Table 3.1), we plot the log of spread ratio by the order $F, E, D, C, B, A, Z, Y, X$. The slopes of log spread ratio of t_{10}, t_5, t_2, t_1 are estimated by least square estimators. The estimated slopes are $0.047, 0.1046, 0.328, 0.782$, respectively (Figure 3.2). We get a very perfect linear relationship between degrees of freedom and inverse of slope (Figure 3.3). So, when error distribution is symmetric, we can calculate the log spread ratio of error distribution by normal and then get the slope from the least squared fits. Finally, the error distribution can be estimated by t-distribution with degrees of freedom $0.5 + 1/(2.25 * slope)$; that is,

$$\hat{DF} = \frac{1}{2} + \frac{1}{2.25 * slope}.$$

3.3.3 Stable distribution family to estimate asymmetric error distribution

NDA microarray data usually consists of log transformed and suitably normalized intensities. So, the distribution of microarray data is often asymmetric and longer tailed and heavily skewed than a normal and outliers are common across genes. The stable distribution [22] is the limit distribution of a suitably scaled sum of independent and identically distributed random variables. Statistically, they are used mostly when an example of a very long-tailed distribution is required. For small values of index, the distribution degenerates to point mass at 0. In this section, we use stable distribution to estimate the asymmetric error distribution which is very skewed. The density of stable function

$$\begin{aligned} f(t;\alpha,\beta) \quad &= exp[-|t|^{\alpha} exp(-\frac{1}{2}\pi i\beta k(\alpha) sign(t))], \text{ when } 0 < \alpha \leq 2, \alpha \neq 1. \\ &= exp[-|t|(1 + \frac{1}{2\pi} i\beta ln(|t|) sign(t))], \text{ when } \alpha = 1. \end{aligned}$$

where $k(\alpha) = 1 - |1 - \alpha|$, and $\beta \in [-1, 1]$. The parameters in stable distribution are index (α) and skewness (β). $\alpha = 2$ corresponds to the normal, $\alpha = 1$ to the Cauchy, and $\alpha = 1/2$ to Pearson distribution. Smaller values mean longer tails. β is the modified skewness [21]. This is not the ordinary skewness of a distribution. Positive values correspond to a long right tail, where the mean is greater than the median. Negative values correspond to a long left tail.

We figure out the relationship between degrees of freedom of t and stable distribution (Table 3.2). When estimate the asymmetric error distribution, we first estimate the tail distribution with t-family and use the relationship between t and stable distribution. We can estimate error distribution using stable distribution.

3.4 Simulation studies

3.4.1 Performance assessment on nonparametric methods

To assess the performance of location and scale estimators, we generate data matrices with 1000 rows and 8 columns from $N(0,1)$, t_4, and t_2 that are the error distributions. Rows represent genes and $1,2,3,4$ and $5,6,7,8$ columns represent replicates of each gene in different samples. For each matrix, we standardize the data set using two schemes. One is subtracting the grand mean and dividing the over all standard deviation. The other is subtracting the median and dividing the MAD. After standardization, for each row we estimate the error by subtracting location estimator and dividing scale estimator and then we get the estimated error distribution. The location estimators are mean, median, bisquare, Huber, and Hubers and the scale estimators are standard deviation, MAD, scale.tau, scale.a, and Hubers. In each pair of location and scale estimator, we repeat 10 simulations and calculate 10 SSPDs and then calculate mean and standard deviation of SSPDs. (Table 3.3-3.8) show the results of the simulations. In the normal case, Hubers and SD have smaller SSPD than scale estimators, and in t_4, Hubers scale usually dominates other scale estimators, but in t_2, the performance of MAD is better than others. In this simulation studies, we found that scale estimators is more sensitive than location estimators that means the values of SSPD are similar within scale estimators, but different between scale estimators. So, the scale estimators are much important than location estimators in this simulation study. We found that when the pair is Bisquare and Huber proposal 2, SSPD obtains the minimum in most cases. We can conclude that among these location and scale estimators, Bisquare and Huber proposal 2 are the best estimators for error estimation.

3.4.2 Performance assessment on parametric methods for symmetric error distribution

According to section 3.4.1, we know that Bisquare(location) and Huber proposal 2(scale) are the best nonparametric estimator by simulation studies. In this section, we compare nonparametric method mean and standard deviation, Bisquare and Huber proposal

2 with t-approach. All simulation settings are the same section 3.4.1. In each pair of location and scale estimator and t-approach estimate, we repeat 10 simulations and calculate 10 SSPDs and then calculate mean and standard deviation of SSPDs. (Table 3.9 - 3.11) show the results of the simulations. For symmetric error distributions Normal or t family, t-approach is a better estimator than nonparametric method (Bisquare-Hubers). When analyzing microarray data, if the error distribution is symmetric, we suggest using t-approach method to get a better result.

3.4.3 Performance assessment on parametric methods for asymmetric error distribution

According to section 3.4.1 and 3.4.2, we know that Bisquare(location) and Huber proposal 2(scale) are the best nonparametric estimators and the t-approach is better than nonparametric methods when the error distribution is symmetric. In this section, we compare nonparametric methods for the mean and standard deviation, Bisquare and Huber proposal 2 with parametric methods the t-approach and the Stable-approach. All simulation settings are the same section 3.4.1 and 3.4.2. In each pair of location and scale estimator and the t-approach and the stable-approach estimates, we repeat 10 simulations and calculate 10 SSPDs and then calculate mean and standard deviation of SSPDs. (Table 3.12 - 3.15) show the results of the simulations. For asymmetric error distributions (Gamma, Chi square, F or Stable), stable-approach and nonparametric are better than t-approach. So if the error distribution of microarray data is asymmetric, we suggest using stable-approach or nonparametric method to estimate the error distribution.

3.5 Discussion

In table 3.2 - 3.7, we find that scale estimators are more sensitive than location estimators that means the values of SSPD are similar within scale estimators, but different between them. So, the scale estimators are much important than location estimators. Hubers scale estimator is recommended in most cases, but when the error distribution is very skewed and heavy tailed(very few cases), then MAD is recommended. In this paper, we find a nice relationship of tails between normal and t family (Figure 1-3). When the empirical error distribution is symmetric, t-approach is a good technique, but when it is not symmetric, Bisqure - Hubers or stable distribution approach is better based on our simulation studies.

Table 3.1: Letter spreads for the t_{10}, t_5, t_2, and t_1 and ratios by standard normal spreads

	Normal	t_{10}	ratio	t_5	ratio	t_2	ratio	t_1	ratio
F_{sp}	1.35	1.40	1.04	1.45	1.08	1.63	1.21	2.00	1.48
E_{sp}	2.30	2.44	1.06	2.60	1.13	3.21	1.39	4.83	2.10
D_{sp}	3.07	3.35	1.09	3.68	1.20	5.11	1.67	10.05	3.28
C_{sp}	3.73	4.19	1.13	4.78	1.28	7.62	2.05	20.31	5.45
B_{sp}	4.31	5.01	1.16	5.93	1.38	11.05	2.56	40.71	9.45
A_{sp}	4.84	5.82	1.20	7.19	1.49	15.81	3.27	81.47	16.85
Z_{sp}	5.32	6.63	1.25	8.57	1.61	22.49	4.23	162.97	30.63
Y_{sp}	5.77	7.46	1.29	10.12	1.75	31.91	5.53	325.95	56.48
X_{sp}	6.19	8.32	1.34	11.85	1.91	45.19	7.29	651.90	105.24

Table 3.2: Relation between t and stable distributions.

DF of t	1	2	3	4	5	6	7	8
α	1.0	1.5	1.8	1.83	1.86	1.90	1.92	1.93
DF of t	9	10	11	12	13	14	>15	
α	1.94	1.95	1.96	1.97	1.98	1.99	2.0	

Table 3.3: N(0,1) standardized by mean and standard deviation

	Scale				
Location	SD	MAD	scale.tau	scale.a	Hubers
mean	0.34	0.96	1.16	6.6	0.21
(sd)	(0.05)	(0.14)	(0.30)	(1.68)	(0.03)
median	0.16	2.0	2.2	8.7	0.2
(sd)	(0.04)	(0.16)	(0.42)	(1.68)	(0.05)
bisquare	0.16	2.26	2.49	9.31	0.2
(sd)	(0.03)	(0.22)	(0.49)	(1.46)	(0.07)
Huber	0.13	1.70	1.78	7.79	0.20
(sd)	(0.04)	(0.15)	(0.39)	(1.83)	(0.06)
Hubers	0.21	1.21	1.42	6.81	0.26
(sd)	(0.04)	(0.15)	(0.34)	(1.56)	(0.06)

Table 3.4: t_4 standardized by mean and standard deviation

	Scale				
Location	SD	MAD	scale.tau	scale.a	Hubers
mean	2.10	0.30	0.50	3.79	0.53
(sd)	(0.15)	(0.16)	(0.39)	(1.58)	(0.05)
median	1.30	0.85	1.04	5.32	0.16
(sd)	(0.13)	(0.25)	(0.44)	(1.38)	(0.06)
bisquare	1.07	1.00	1.25	5.81	0.17
(sd)	(0.11)	(0.28)	(0.46)	(1.42)	(0.09)
Huber	1.36	0.75	0.91	4.74	0.26
(sd)	(0.12)	(0.26)	(0.36)	(1.41)	(0.07)
Hubers	1.51	0.57	0.78	4.30	0.37
(sd)	(0.13)	(0.24)	(0.42)	(1.37)	(0.08)

Table 3.5: t_2 standardized by mean and standard deviation

	Scale				
Location	SD	MAD	scale.tau	scale.a	Hubers
mean	8.05	0.33	0.46	1.45	1.12
(sd)	(1.22)	(0.18)	(0.30)	(0.80)	(0.41)
median	6.17	0.36	0.57	2.61	0.45
(sd)	(1.10)	(0.33)	(0.36)	(0.76)	(0.17)
bisquare	5.80	0.45	0.66	2.84	0.39
(sd)	(1.10)	(0.35)	(0.40)	(0.83)	(0.17)
Huber	6.22	0.38	0.57	2.44	0.54
(sd)	(1.11)	(0.35)	(0.39)	(0.76)	(0.17)
Hubers	6.39	0.31	0.46	1.87	0.68
(sd)	(1.12)	(0.38)	(0.32)	(0.73)	(0.17)

Table 3.6: N(0,1) standardized by median and MAD

	Scale				
Location	SD	MAD	scale.tau	scale.a	Hubers
mean	0.72	1.59	1.57	13.4	0.53
(sd)	(0.14)	(0.48)	(0.28)	(3.67)	(0.11)
median	0.27	5.78	7.17	25.8	0.36
(sd)	(0.08)	(1.08)	(0.69)	(5.42)	(0.13)
bisquare	0.18	4.80	5.0	23.7	0.23
(sd)	(0.02)	(0.92)	(0.55)	(5.17)	(0.07)
Huber	0.26	3.11	2.83	16.60	0.30
(sd)	(0.09)	(0.82)	(0.40)	(2.90)	(0.07)
Hubers	0.48	2.03	1.85	13.6	0.48
(sd)	(0.12)	(0.54)	(0.26)	(2.52)	(0.11)

Table 3.7: t_4 standardized by median and MAD

	Scale				
Location	SD	MAD	scale.tau	scale.a	Hubers
mean	5.08	0.53	0.89	12.0	2.60
(sd)	(0.66)	(0.25)	(0.58)	(6.25)	(0.46)
median	1.86	4.29	7.43	32.3	0.29
(sd)	(0.34)	(1.14)	(2.61)	(12.8)	(0.04)
bisquare	2.00	3.72	4.53	28.6	0.60
(sd)	(0.39)	(1.07)	(2.22)	(12.1)	(0.13)
Huber	2.91	2.06	2.36	19.9	1.15
(sd)	(0.49)	(0.80)	(1.45)	(8.78)	(0.27)
Hubers	3.71	1.13	1.30	14.9	1.85
(sd)	(0.59)	(0.52)	(0.80)	(6.98)	(0.44)

Table 3.8: t_2 standardized by median and MAD

	Scale				
Location	SD	MAD	scale.tau	scale.a	Hubers
mean	36.3	3.03	3.35	11.5	15.4
(sd)	(4.07)	(1.81)	(1.90)	(11.0)	(3.25)
median	20.9	3.03	7.73	38.4	4.00
(sd)	(2.85)	(2.23)	(4.46)	(13.5)	(1.65)
bisquare	22.0	2.70	4.00	34.7	5.21
(sd)	(2.88)	(2.02)	(2.56)	(13.7)	(1.68)
Huber	25.7	1.68	2.05	23.6	7.88
(sd)	(3.20)	(1.19)	(1.31)	(14.2)	(2.22)
Hubers	28.9	1.71	1.84	13.6	10.9
(sd)	(3.48)	(0.81)	(0.68)	(9.30)	(2.74)

Table 3.9: N(0,1)

Method	Sd	Bisq-Hubers	t-approach
SSPD	0.336	0.21	0.235

Table 3.10: t_2

Method	Sd	Bisq-Hubers	t-approach
SSPD	7.8	0.3	0.057

Table 3.11: t_4

Method	Sd	Bisq-Hubers	t-approach
SSPD	2.05	0.16	0.07

Table 3.12: Gamma(5,1)

Method	Sd	Bisq-Hubers	t-approach	Stable-approach
SSPD	247	220	249	220

Table 3.13: $F_{(3,1)}$

Method	Sd	Bisq-Hubers	t-approach	Stable-approach
SSPD	69	0.02	36	7

Table 3.14: $\chi^2_{(3)}$

Method	Sd	Bisq-Hubers	t-approach	Stable-approach
SSPD	80	54	84	59

Table 3.15: stable(1.5,1)

Method	Sd	Bisq-Hubers	t-approach	Stable-approach
SSPD	19.2	1.38	16.1	0.031

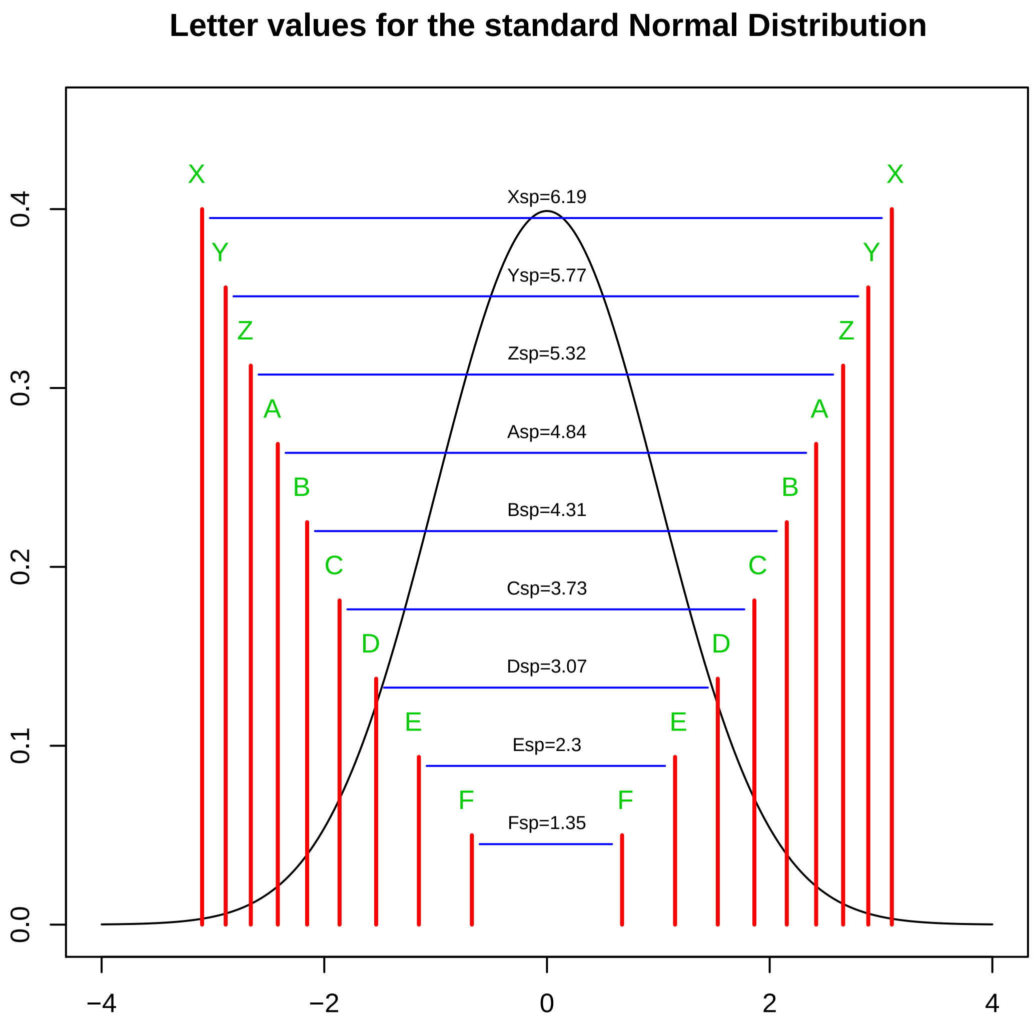


Figure 3.1: Letter value for N(0,1)

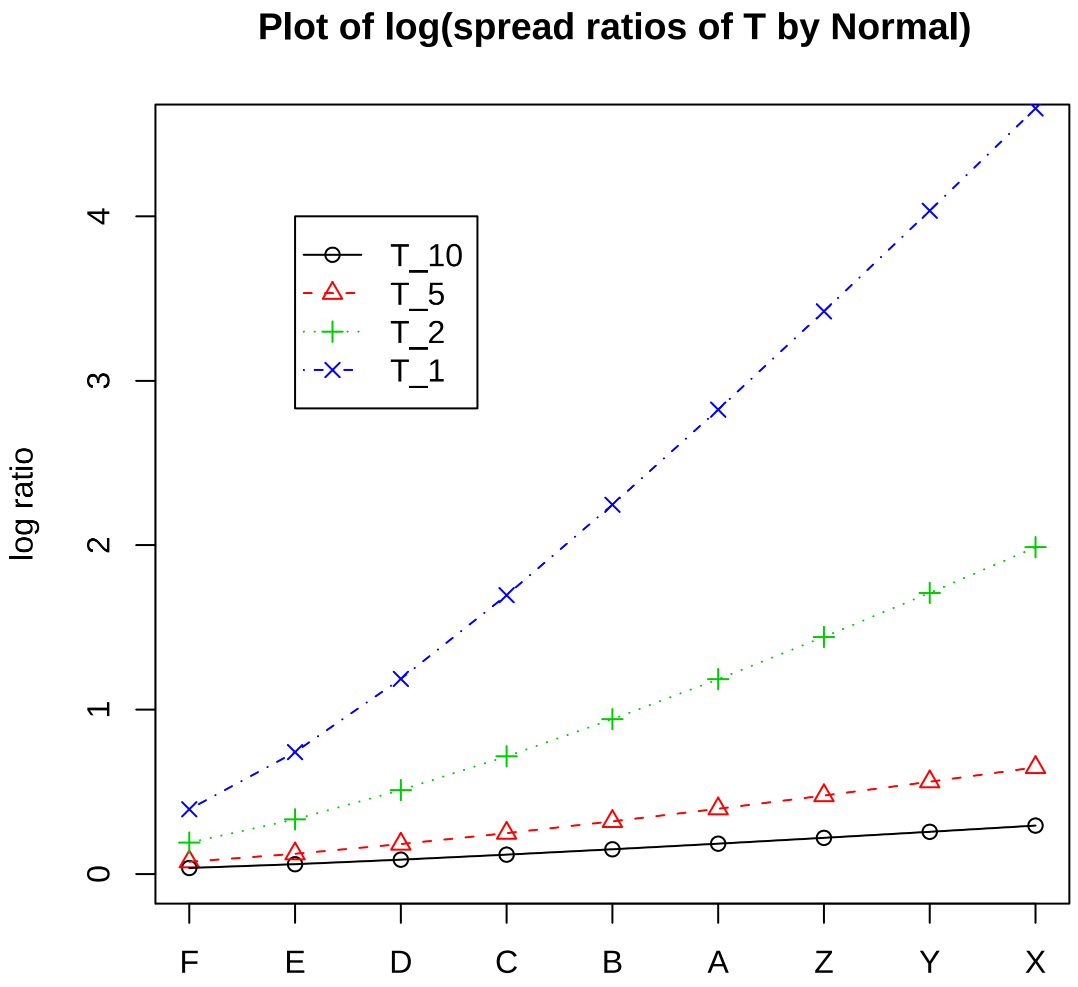


Figure 3.2: Log of spread ratios of T by N(0,1)

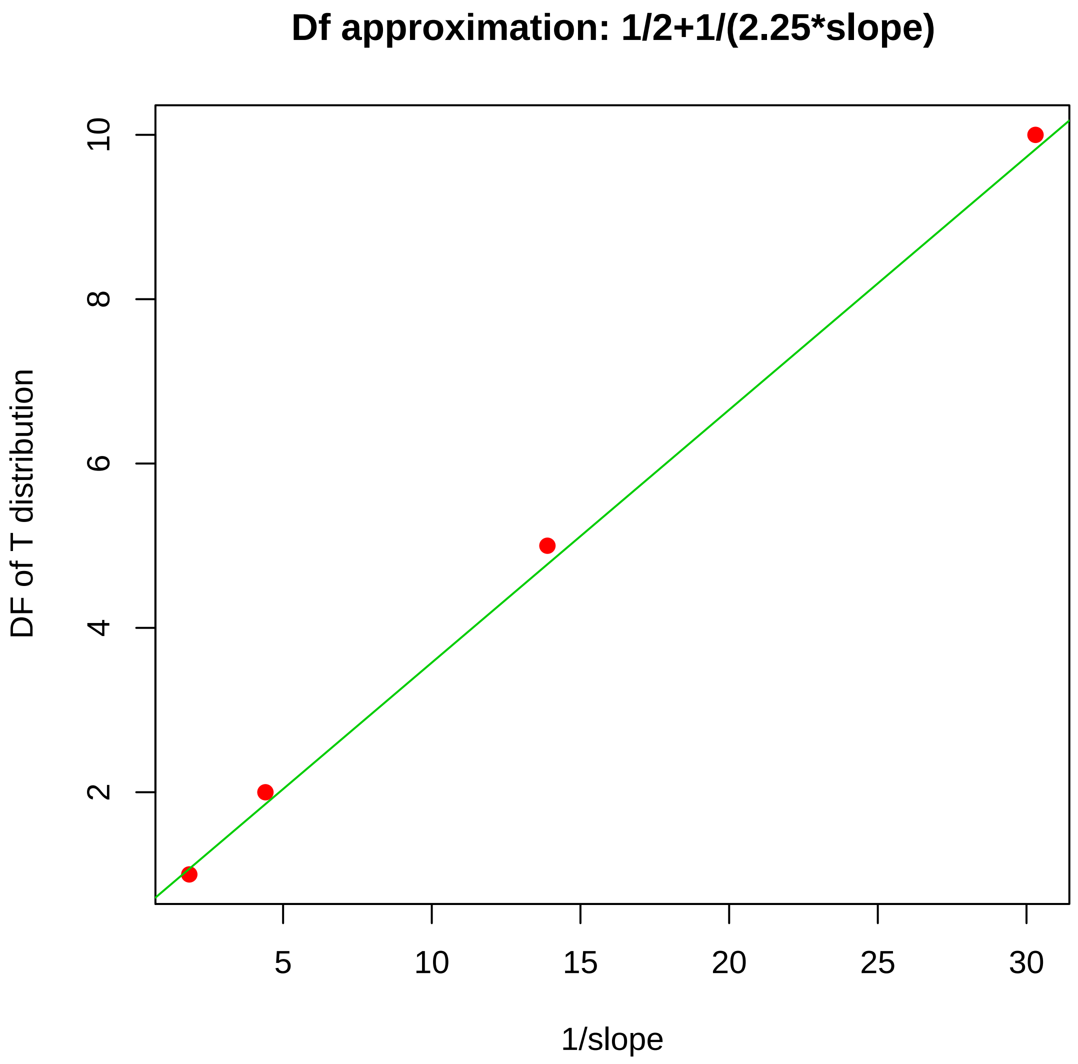


Figure 3.3: Plot of linear relationship between df of t and the inverse of slopes

Chapter 4

DNA barcode of Life and Data analysis

4.1 Introduction to DNA barcode

In the past two years, a series of studies [46][47] have been published in which ”DNA barcoding” was proposed as a tool for differentiating biological species. Barcoding is based on two assumptions that (1) mitochondrial DNA, double stranded DNA, does not recombine and it is past without any variation from mother to children, (2) the rate of mutation is slow enough to preserve short sequences among spcies with very small divergences. With million of species and their life-stage transformations, the animal kingdom provides a challenging target for taxonomy. Recent work has suggested that a DNA-based identification system, found that the mitochondrial gene, cytochrome *c* oxidase subunit 1 (COI) with 648 base pairs long, can aid the resolution of this diversity. COI has emerged as a suitable barcode region for most taxonomic groups of animals. Some articles [46] [47] showed that the sequence divergences at COI sequence regularly enable the discrimination of closely allied species in most animal phyla. This success in species diagnosis reflects both the high rates of sequence change at COI in most animal groups and constrains on intraspecific mitochondrial DNA divergence arising, at least in part, through selective sweeps mediated via interactions with the nuclear genome. There is no compelling a priori reason to focus analysis on a specific gene, but COI sequence does have two important advantages. First, it is robust in the sense that it has low variation within species. Second, COI appears to possess a greater range of phylogenetic signal than any other mitochondrial gene. So the species-level diagnoses can routinely be obtained through COI analysis.

4.2 Challenges of DNA Barcode data

What are some of the scientific and technical challenges associated with barcode data?

1. Specimen identification versus ”species discovery”.

   Barcode data are being used in two ways: to assign unidentified specimens to known species, and to improve our knowledge of species differences (including the occasional discovery of potential new species.) What analytical methods are appropriate for these different tasks, and what new approaches to ”novelty detection” could be applied to barcode data?

2. Using character-based barcodes.

   The nucleotide found at each site (A, G, C or T) can be used as a data point, which opens an alternative approach to comparing specimens in terms of overall percent sequence similarity (or difference). How can we analyze barcode data that are treated as discrete characters?

3. Measuring confidence.

   How should our confidence in decisions based on barcode data be calculated when we assign a specimen to a known species, or when we say that two clusters of specimens are distinct and may be separate species? How should the quality of the sequence data, sample size, and our knowledge of the biology of populations and species be incorporated into confidence measures?

4. Optimizing sample size.

   How many specimens per species are needed to create a reliable ”reference barcode” for a species? These reference barcodes must have sufficient information about intraspecific variability to enable accurate assignment of unidentified specimens to their correct species. How should these minimum sample sizes reflect the biology and evolutionary history of each species?

5. Shrinking the barcode.

   How long a gene sequence is needed to assign specimens to known species, and to uncover potentially new species? Do we need multiple gene regions, or a single region, or just certain nucleotide sites within one region?

It has been converted to more concrete scientific problems as following.

1. Assignment to known species.

   Construct a classification rule using the training set that assigns specimens in the testing set to the species contained in the training set. Your classification rule should both maximize correct assignments and minimize incorrect assignments.

2. Data visualization.

   Most barcode studies display results as phonetic cluster diagrams, which can easily be confused with phylogenetic trees. An important challenge is to develop new visualization methods to display and analyze barcode data. These new methods should improve our ability to see and explore the degree and structure of variability within species and divergence among species. (Illustrate the new methods on the datasets provided.)

3. Character-based approaches to barcode data.

   All barcoding studies to date have measured variation within species and divergence among species with phonetic measures - measures of distance based on overall similarly (or dissimilarity) among barcode sequences. Barcode sequences can also be compared using the nucleotides at equivalent sites in the sequence. This approach parallels the use of homologous characters in phylogenetic analysis, and there may be significant synergies with those analytical methods. Character-based approaches may also reduce drastically the barcode sequence lengths needed as diagnostics. The challenge is to develop protocols and software that analyze character-based barcode data to assign specimens to known species and to identify barcode clusters that might be new species.

4. Detection of possible new species.

   A good portion of the specimens in the testing set do not belong to the species included in the training set. Your classification rule should also be able to do two things in addition to assigning specimens to their correct species: (1) determine which specimens should not be assigned to known species, and (2) how these unclassified specimens in the testing set should be partitioned among potentially new species. Your classification rule should therefore identify clusters among unclassified specimens in the testing set that might be new species.

5. Confidence measures.

   For some specimens the classification into known species might be unclear or borderline. One way to think more formally about these situations is to assign a measure of confidence to the assignments of specimens to species. Another is assigning measures of confidence to new clusters representing new species. The challenge is to propose ways to measure confidence of assignments to species and separation among clusters.

6. Clusters within species.

   All species include some level of variation among individuals and in some cases this variation takes the form of splits among local populations and even subspecies. The challenge is to develop a classification rule that will identify clusters within individual species that rise above background variation and therefore might represent subspecies or other significant biological units. This is similar to Challenge 2 but instead of clustering a big dataset you will cluster many smaller datasets since you need to find clusters separately within each individual species.

7. Sample size.

   The first four challenges are doable as long as there is a sufficient number of specimens per species. ”Sufficient” will be a relative term, varying with a number of biological variables (population size, intraspecific variability, and gene flow are three important ones). The challenge is to provide guidelines for sample size - guidelines that will allow your clustering method and/or classification rule to produce decisions with a determined level of confidence. You should also explore how your method is robust relative to small absolute and relative sample sizes.

8. Metric for barcode data.

One important question is related to the clustering and classification methods needed for the first five challenges: What kind of distance metrics should be used to measure the difference (similarity) between barcodes? Barcode data is high-dimensional and categorical, and very little is known about how to analyze this kind of data. Many "off-the-shelf" methods can be applied to this data. One approach to improving the results obtained with "off-the-shelf" methods is by tapping into the specific structure of the data. One might start by analyzing the variability structure within species and between species and understanding the "correlation structure" of the data in high dimensions. How can one exploit the complexity of this correlation structure to obtain better results than the standard methods? How can one model this structure? Finally, you should address how new methods you propose compare in performance to "off-the-shelf" methods.

9. New clustering methods.

   Clustering methods are challenged by small datasets and are often not robust under dynamically changing datasets. Will your approaches lead to some new clustering methods? More generally, will barcoding lead to some new clustering methods? How about new Bayesian clustering methods? The challenge is to describe such new methods and illustrate them on the datasets given.

4.3 Genetic distance

Genetic distance is one of the most popular methods in bioinformatics to construct the evolutionary tree that represents the historical relationships between the species being analyzed. The measure of genetic distance between the sequences being classified requires a multiple sequence alignment(MSA) as an input. Distance is often defined as the fraction of mismatches at aligned positions, with gaps either ignored or counted as mismatches. There are many types of genetic distance as follows:

Let X_i, Y_i be the frequencies of the i^{th} nucleotide or amino acid from the population A and B, respectively. Then

1. Euclidean distance D_{EU}

$$D_{EU} = \sqrt{\sum_i (X_i - Y_i)^2}.$$

2. Sanghvi distance (1953) X^2

$$X^2 = 2\sum_i \frac{(X_i - Y_i)^2}{X_i + Y_i}.$$

3. Cavalli-Sforze and Edwards chord distance (1967) D_{CH}

$$D_{CH} = \frac{2}{\pi}\sqrt{2(1 - \sum_i \sqrt{X_i Y_i})}.$$

4. Rogers distance (1972) D_R

$$D_R = \sqrt{\frac{\sum_i (X_i - Y_i)^2}{2}}.$$

5. Prevosti distance (1975) O_p

$$O_p = \sum_i \frac{|X_i - Y_i|}{2}.$$

6. Nei distance (1983) D_A

$$D_A = 1 - \sum \sqrt{X_i Y_i}.$$

7. Bhattacharyya and Nei distance (1987) θ^2

$$\theta^2 = (\arccos \sum_i \sqrt{X_i Y_i})^2.$$

Some distance measures performed well in some circumstances, but worse in others. No one can dominate others. So it is very important to construct a robust distance measure or it is always best if we compare several distance measures under conditions in which we know what the answer should be. But it is very hard to find a best and robust distance measure and it is very tedious to compare several distance measures under some circumstances. In stead of using distance measure, we can use clustering tools in statistics that are very powerful and very simple.

4.4 Optimal Scoring Rule

In statistics, clustering techniques have been applied to a wide variety of research problems and provide an excellent summary of the many published studies reporting the results of cluster analysis. For example, in the field of medicine, clustering diseases, cures for diseases, or symptoms of diseases can lead to very useful taxonomies. In the field of psychiatry, the correct diagnosis of clusters of symptoms such as paranoia, schizophrenia, etc. is essential for successful therapy. In archeology, researchers have attempted to establish taxonomies of stone tools, funeral objects, etc. by applying cluster analytic techniques. In the new field DNA barcoding, it can cluster the species from COI sequences. In general, whenever one needs to classify a 'mountain' of information into manageable meaningful piles, cluster analysis is of great utility.

In order to use the clustering method, the COI sequences should convert to real numbers for all specimens. One possibility is transferring $X_i \in \{A,\ C,\ T,\ G\}$ to $\{1,\ 2,\ 3,\ 4\}$ for the i^{th} loci in COI sequence, but this method did not take any statistical sense into account. We put a constrain on X_is that we standardized the variance for each position of the sequence and then estimate X_is by optimizing the F statistics across the species for $i = 1, \cdots, 648$, where
F = (Mean Square error Between Species)/(Mean Square error Within Species).
For convenient, we always code the smallest number as 1. This procedure is called 'Optimal Scoring Rule'(OSR). After procedure OSR, the variance of the numbers that A, C, T, G transform to is the same as 5/3, Var{1,2,3,4}, for all positions. In order to display the scheme of OSR, a simple example is illustrated as following. We use small part of the dataset which consists of COI sequences of birds of North America [114]. In that dataset, there are 180 species among 2369 different birds. We take 3 species which have 10 sequences with length 5 (Table 4.1). First, we code the characters A, C, T, G to 1, 2, 3, and 4, respectively (Table 4.2). OSR procedure is applied on a column at each time when optimizing the F statistics. We get the relationship between original codes and new codes (Table 4.3). Then we can transform the original codes in Table 4.2 to new codes (Table 4.4). Then we use the clustering method on new coding of the data set.

Table 4.1: COI sequence

Species	COI sequences
Accipiter	G T C C G
Accipiter	G T G C G
Accipiter	G T C C G
Accipiter	G T C C G
Accipiter	C T C C G
Actitis	C T G C C
Actitis	C T G C C
Aegolius	A T C C C
Aegolius	A T C C C
Aegolius	A T C C G

Table 4.2: COI sequences transform to numbers

Species	COI sequences
Accipiter	4 3 2 2 4
Accipiter	4 3 4 2 4
Accipiter	4 3 2 2 4
Accipiter	4 3 2 2 4
Accipiter	2 3 2 2 4
Actitis	2 3 4 2 2
Actitis	2 3 4 2 2
Aegolius	1 3 2 2 2
Aegolius	1 3 2 2 2
Aegolius	1 3 2 2 4

Table 4.3: Relationship between original and new coding

Column	original coding	new coding
1	(1, 2, 3, 4)	(1, 2.750, 3.509, 3.920)
2	(1, 2, 3, 4)	(1, 2, 3, 4)
3	(1, 2, 3, 4)	(1, 2.861, 3.471, 3.943)
4	(1, 2, 3, 4)	(1, 2, 3, 4)
5	(1, 2, 3, 4)	(1, 2.643, 3.190, 4.067)

Table 4.4: Sequence after OSR

Species	COI sequences
Accipiter	3.920 3.000 2.861 2.000 4.067
Accipiter	3.920 3.000 3.943 2.000 4.067
Accipiter	3.920 3.000 2.861 2.000 4.067
Accipiter	3.920 3.000 2.861 2.000 4.067
Accipiter	2.750 3.000 2.861 2.000 4.067
Actitis	2.750 3.000 3.943 2.000 2.643
Actitis	2.750 3.000 3.943 2.000 2.643
Aegolius	1.000 3.000 2.861 2.000 2.643
Aegolius	1.000 3.000 2.861 2.000 2.643
Aegolius	1.000 3.000 2.861 2.000 4.067

4.5 Principal Components Analysis (PCA)

Principal components analysis (PCA) is a method of classical multivariate analysis that is the most commonly used technique for dimension reduction. PCA usually useful in situation where we are dealing with many variables and we want to reduce them to a few new uncorrelated variables, linear combinations of the original variables without losing much information. None of the variables is designated as dependent. In seeking a linear combination with maximal variance, we are essentially searching for a dimension along which the observations are maximally separated or spread out. In general, the principal components define different dimensions from discriminant functions. In some applications, principal components are often obtained for use as input to another analysis.

Principal components analysis deals with a single sample of observations with no structure in the observations. We have a sample of n observations $Y_1, Y_2, \cdots, Y_n$, where $(Y_i)_{p\times 1} = (y_1, y_2, \cdots, y_p)', i = 1, 2, \cdots, n$, is a p-dimensional vector. Principal components can be applied to any distribution F_Y. If the variables $y_1, y_2, \cdots, y_p$ in each Y_i are correlated, the ellipsoidal swarm of points is not orientated parallel to any of the axes represented by $y_1, y_2, \cdots, y_p$. We wish to find the natural axes of the swarm of points with origin to sample mean $\bar{Y}$. This can be done by translating the origin to $\bar{Y}$ and rotating the axes. After rotation so that axes become the natural axes of the ellipsoid, the new variables (principal components) will be uncorrelated.

Without loss of generality, we can assume that $Y_i, i = 1, 2, \cdots, n$ has been centered; that is, the average of Y_i is zero for all i. The singular value decomposition (SVD) of $(Y)_{p\times n} = [Y_1\ Y_2\ \cdots\ Y_n]$ is defined as

$$Y = UDV',$$

where U is a $p \times p$ matrix which projects the p-dimensional samples into p-dimensional samples, V is a $n \times p$ matrix with $VV' = I_n$, and D is a $p \times p$ diagonal matrix, whose diagonal elements, s_i, are called singular values. We assume that $s_1 \leq s_2 \leq \cdots \leq s_p$.

From the SVD, the sample variance-covariance matrix of Y is as

$$S = YY' = UD^2U'.$$

Hence the column vectors of U are the principal components of S and the square of the diagonal elements of D are their respective variances:

$$D^2 = \begin{pmatrix} s_1^2 & 0 & \cdots & 0 \\ 0 & s_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & s_p^2 \end{pmatrix}, \tag{4.1}$$

and $U = (u_1, \cdots, u_p)$.
We denote the eigenvalue of S: $\lambda_1 = s_1^2, \lambda_2 = s_2^2, \cdots, \lambda_p = s_p^2$. The objective is to select a subset of k principal components containing most of the information in the original data. Because the eigenvalues are variances of the principal components, it means that the proportion of variance are explained by the first k components.

$$\text{Proportion of variance } = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\sum_{i=1}^{p} s_{z_i}^2}, \tag{4.2}$$

since $\sum_{i=1}^{p} \lambda_i = tr(AS_YA') = tr(S_YA'A) = tr(S_Y)$. However we should decide how many components should be retained. Usually, we should retain sufficient components to account for a high percentage of total variance.

4.6 Linear Discriminant Analysis(LDA)

4.6.1 Introduction

A simplest method of supervised classification, Linear Discriminant Analysis (LDA) (Fisher, 1936 [40]), endures to this day as one of most popular classification techniques. The term 'group' presents either a population or a sample from the population. There are two objectives in separation of groups.

1. Description of group separation, in which linear functions (discriminant function) of the variables are used to describe or elucidate the differences between two or more groups. The goals of discriminant analysis include identifying the relative contribution of the p variables to separation of the groups and finding the optimal plane on which the points can be projected to best illustrate the configuration of the groups.

2. Prediction or allocation, in which linear or quadratic function (classification functions) of the variables are employed to assign an individual sampling unit to one of the groups. The measured values (in the observation vector) for an individual or object are evaluated by the classification functions to see to which group the individual most likely belongs.

In this section, we focus on linear discriminant function analysis.

4.6.2 LDA on two groups

In the case of two groups, we have a sampling unit to be classified, but we do not know to which of the two groups the subject or object belongs. We assume that the two populations to be compared have the same covariance matrix Σ but distinct mean vectors μ_1 and μ_2, where Σ is the p by p covariance matrix and μ_1, μ_2 are the p by 1 vectors. And samples $Y_{11}, \cdots, Y_{1n_1}$ and $Y_{21}, \cdots, Y_{2n_2}$ are from two groups 1 and 2, respectively, where $Y'_{ij} = (Y_{ij1}, Y_{ij2}, \cdots, Y_{ijp}), i = 1, 2,\ j = 1, \cdots, p$ consists of measurements on p variables. The discriminant function is the linear combination of these p variables that maximizes the distance between the two group mean vectors.

Our goal is to find a p by 1 vector a, where $a' = (a_1, a_2, \cdots, a_p)$, such that a linear combination $z_{ij} = a'Y_{ij}, i = 1, 2,\ j = 1, \cdots, p$ transforms each observation vector to a scalar:

$$z_{1j} = a'Y_{1j} = a_1Y_{1j1} + a_2Y_{1j2} + \cdots + a_pY_{1jp}, \quad j = 1, 2, \cdots, n_1,$$
$$z_{2j} = a'Y_{2j} = a_1Y_{2j1} + a_2Y_{2j2} + \cdots + a_pY_{2jp}, \quad j = 1, 2, \cdots, n_2.$$

Hence the $n_1 + n_2$ observation vectors

$$Y_{11}, Y_{12}, \cdots, Y_{1n_1}$$

$$Y_{21}, Y_{22}, \cdots, Y_{2n_2}$$

are transferred to scalars

$$z_{11}, z_{12}, \cdots, z_{1n_1}$$

$$z_{21}, z_{22}, \cdots, z_{2n_2}.$$

We find the means

$$\bar{z}_1 = \sum_{j=1}^{n_1} a'Y_{1j}/n_1 = a'\bar{Y}_1 \tag{4.3}$$

and

$$\bar{z}_2 = a'\bar{Y}_2, \tag{4.4}$$

and wish to find the vector a that maximized the standardized difference $(\bar{z}_1 - \bar{z}_2)/s_z$. s_z^2 is the pooled covariance of $z_{1i}, i = 1, \cdots, n_1$ and $z_{2i}, i = 1, \cdots, n_2$. s_z^2 also can be expressed as

$$\begin{aligned}
s_z^2 &= \frac{1}{n_1+n_2-2}((n_1-1)\sum_{i=1}^{n1}(z_{1i}-\bar{z}_1)^2+(n_2-1)\sum_{i=1}^{n_2}(z_{2i}-\bar{z}_2)^2) \qquad (4.5)\\
&= \frac{1}{n_1+n_2-2}((n_1-1)\sum_{i=1}^{n1}(a'Y_{1i}-a'\bar{Y}_1)(a'Y_{1i}-a'\bar{Y}_1)'\\
&+ (n_2-1)\sum_{i=1}^{n2}(a'Y_{2i}-a'\bar{Y}_2)(a'Y_{2i}-a'\bar{Y}_2)')\\
&= a'\{\frac{1}{n_1+n_2-2}((n_1-1)\sum_{i=1}^{n1}(Y_{1i}-\bar{Y}_1)(Y_{1i}-\bar{Y}_1)'\\
&+ (n_2-1)\sum_{i=1}^{n1}(Y_{1i}-\bar{Y}_1)(Y_{1i}-\bar{Y}_1)')\}a\\
&= a'\{\frac{1}{n_1+n_2-2}((n_1-1)S_1+(n_2-1)S_2)\}a\\
&= a'S_{pl}a,
\end{aligned}$$

$$(4.6)$$

where S_1 is the p by p covariance matrix of group 1, S_2 is the p by p covariance matrix of group 2, and S_{pl} is the p by p pooled covariance matrix of group 1 and 2. Since $(\bar{z}_1-\bar{z}_2)/s_z$ can be negative, we use the squared distance $(\bar{z}_1-\bar{z}_2)^2/s_z^2$, by (4.5), (4.6), and (4.9), that can be written as

$$\frac{(\bar{z}_1-\bar{z}_2)^2}{s_z^2}=\frac{[a'(\bar{Y}_1-\bar{Y}_2)]^2}{a'S_{pl}a}. \tag{4.7}$$

The maximum of (4.10) occurs when

$$a=S_{pl}^{-1}(\bar{Y}_1-\bar{Y}_2)(\text{ if } S_{pl}^{-1} \text{ exists}), \tag{4.8}$$

or when a is any multiple of $S_{pl}^{-1}(\bar{Y}_1-\bar{Y}_2)$.

The maximizing vector a is not unique, but its direction is unique; that is, the relative values or ratios of $a_1,\cdots,a_p$ are unique. Note that in order for S_{pl}^{-1} to exist, we must have $n_1+n_2-2>p$.

Suppose Y is the vector of measurements on a new sampling unit that we wish to classify into one of the two groups and $n_1+n_2-2>p$. For each observation $Y_{1i}, i=1,\cdots,n_1$ from group 1, we calculate $z_{1i}=a'Y_{1i},\cdots,z_{1n_1}=a'Y_{1n_1}$. From (4.6) we have $\bar{z}_1=a'\bar{Y}_1=(\bar{Y}_1-\bar{Y}_2)'S_{pl}^{-1}\bar{Y}_1$ (Since $(S_{pl}^{-1})'=S_{pl}^{-1}$). Similarly, from (4.7) we have $\bar{z}_2 = (\bar{Y}_1 - \bar{Y}_2)'S_{pl}^{-1}\bar{Y}_2$. Fisher's (1936) *linear classification procedure* assigns a p by 1 vector Y to group 1 if $z = a'Y$ is closer to $\bar{z}_1$ than to $\bar{z}_2$ and assigns Y to group 2 if $z = a'Y$ is closer to $\bar{z}_2$ than to $\bar{z}_1$. Then

$$\bar{z}_1 - \bar{z}_2 = a'(\bar{Y}_1 - \bar{Y}_2) = (\bar{Y}_1 - \bar{Y}_2)'S_{pl}^{-1}(\bar{Y}_1 - \bar{Y}_2) > 0, \tag{4.9}$$

where S_{pl}^{-1} is positive definite. So, if $a' = (\bar{Y}_1 - \bar{Y}_2)'S_{pl}^{-1}$ then it is always true that $\bar{z}_1$ is always greater than $\bar{z}_2$. That is $\bar{z}_1 > \bar{z}_2$. If $a' = (\bar{Y}_2 - \bar{Y}_1)'S_{pl}^{-1}$, then $\bar{z}_2 > \bar{z}_1$. Since $\frac{1}{2}(\bar{z}_1 + \bar{z}_2)$ is the midpoint of $\bar{z}_1$ and $\bar{z}_2$, $z > \frac{1}{2}(\bar{z}_1 + \bar{z}_2)$ implies z is closer to $\bar{z}_1$.
To express the classification rule in terms of Y, we write $\frac{1}{2}(\bar{z}_1 + \bar{z}_2)$ as

$$\frac{1}{2}(\bar{z}_1 + \bar{z}_2) = \frac{1}{2}(\bar{Y}_1 - \bar{Y}_2)'S_{pl}^{-1}(\bar{Y}_1 + \bar{Y}_2). \tag{4.10}$$

Then the classification rule becomes : Assign Y to group 1 if

$$a'Y = (\bar{Y}_1 - \bar{Y}_2)'S_{pl}^{-1}Y > \frac{1}{2}(\bar{Y}_1 - \bar{Y}_2)'S_{pl}^{-1}(\bar{Y}_1 + \bar{Y}_2), \tag{4.11}$$

and assign Y to group 2 if

$$a'Y = (\bar{Y}_1 - \bar{Y}_2)'S_{pl}^{-1}Y < \frac{1}{2}(\bar{Y}_1 - \bar{Y}_2)'S_{pl}^{-1}(\bar{Y}_1 + \bar{Y}_2). \tag{4.12}$$

Or equivalently, we write the linear discriminant function $L(Y)$ as

$$\begin{aligned} L(Y) &= a'Y - \frac{1}{2}(\bar{Y}_1 - \bar{Y}_2)'S_{pl}^{-1}(\bar{Y}_1 + \bar{Y}_2) \qquad (4.13) \\ &= (\bar{Y}_1 - \bar{Y}_2)'S_{pl}^{-1}Y - \frac{1}{2}(\bar{Y}_1 - \bar{Y}_2)'S_{pl}^{-1}(\bar{Y}_1 + \bar{Y}_2) \\ &= a'Y + a_0, \end{aligned}$$

where $a' = (\bar{Y}_1 - \bar{Y}_2)'S_{pl}^{-1}$ is a 1 by p vector and $a_0 = -\frac{1}{2}(\bar{Y}_1 - \bar{Y}_2)'S_{pl}^{-1}(\bar{Y}_1 + \bar{Y}_2)$ is a real number. The linear discriminant function $L(Y)$ is linear in Y, that is, $L(Y) = a_{11}y_1 + a_{12}y_2 + \cdots + a_{1p}y_p + a_0$ is a linear combination of components of the vector Y. Then we can classify the new measurement Y as

$$Y \in \begin{cases} \text{group 1} & \text{if } L(Y) > 0 \\ \text{group 2} & \text{if } L(Y) < 0 \end{cases} \tag{4.14}$$

Fisher's [40] approach using (4.14) and (4.15) is essentially nonparametric because no distributional assumptions were made. However, if the two groups are normal with equal covariance matrices, then this method is optimal.

It is very interesting that the mutual connection between multiple regression and two-group discriminant analysis (Fisher[40] and Flury and Riedwyl [41]) is that the roles of independent and dependent variables are reversed in the two models. The dependent variables of discriminant analysis become the independent variable in regression.

Let $w \in \{0, 1\}$ be the group variable(identifying groups(populations) 1 and 2) such that $\bar{w} = 0$, and define $b = (b_1, b_2, \cdots, b_p)'$ as the vector of regression coefficients. Then we know that b is proportional to the discriminant function coefficient vector $a = S_{pl}^{-1}(\bar{Y}_1 - \bar{Y}_2)$:

$$b = \frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 2) + n_1 n_2 D^2} a,$$

where $D^2 = (\bar{Y}_1 - \bar{Y}_2)' S_{pl}^{-1} (\bar{Y}_1 - \bar{Y}_2)$. The F test statistics for the hypothesis that q of the $p(> q)$ variables are redundant for separating the groups can be obtained as

$$F = \frac{n_1 + n_2 - p - 1}{q} \frac{R_p^2 - R_{p-q}^2}{1 - R_p^2},$$

where R_p^2 and R_{p-q}^2 are the R^2 from regressions with p and $p - q$ variables, respectively. More detail about the link between two-group discriminant analysis and multiple regression is in Fisher [40] and Flury and Riedwyl [41].

Example 4.6.1. *Two species 'Accipiter' and 'Aegolius' consist of 9 birds that are from North America [114]. In each sample (bird), there are total 155 base pairs(variables). We use Optimal Scoring Rule to preprocess this data set(Table 4.5). We denote the species 'Accipiter' is the group 1 and the species 'Aegolius' is the group 2. In this example S_{pl}^{-1} does not exist since $n_1 + n_2 - 2 = 7 < 155 = p$. So we can not use linear discriminant function to classify the new observation of the measurement unit.*

Species	a_1			$\cdots$			a_{155}
Accipiter	*3.566633*	*2.960976*	*1*	$\cdots$	*1*	*3.0369078*	*2.857944*
Accipiter	*3.566633*	*2.960976*	*1*	$\cdots$	*1*	*3.0369078*	*2.857944*
Accipiter	*3.566633*	*2.960976*	*1*	$\cdots$	*1*	*3.0369078*	*2.857944*
Accipiter	*3.566633*	*2.960976*	*1*	$\cdots$	*1*	*3.0369078*	*2.857944*
Accipiter	*3.425382*	*2.960976*	*1*	$\cdots$	*1*	*1.0000000*	*2.857944*
Aegolius	*1.000000*	*0.944621*	*1*	$\cdots$	*1*	*3.0369078*	*1.000000*
Aegolius	*1.000000*	*0.944621*	*1*	$\cdots$	*1*	*3.0369078*	*1.000000*
Aegolius	*1.000000*	*1.000000*	*1*	$\cdots$	*1*	*0.6391904*	*1.000000*
Aegolius	*1.000000*	*1.000000*	*1*	$\cdots$	*1*	*0.6391904*	*1.000000*

Table 4.5: Birds of North America

The data for the group 1

$$(X_1)_{155\times 5} = \begin{bmatrix} 3.566633 & 3.566633 & 3.566633 & 3.566633 & 3.425382 \\ 2.960976 & 2.960976 & 2.960976 & 2.960976 & 2.960976 \\ 1 & 1 & 1 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 & 1 \\ 3.0369078 & 3.0369078 & 3.0369078 & 3.0369078 & 1 \\ 2.857944 & 2.857944 & 2.857944 & 2.857944 & 2.857944 \end{bmatrix} \tag{4.15}$$

$$= \begin{bmatrix} X_{11} & X_{12} & X_{13} & X_{14} & X_{15} \end{bmatrix}.$$

The data for the group 2

$$(X_2)_{155\times 4} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0.944621 & 0.944621 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 3.0369078 & 3.0369078 & 0.6391904 & 0.6391904 \\ 1 & 1 & 1 & 1 \end{bmatrix} \tag{4.16}$$

$$= \begin{bmatrix} X_{21} & X_{22} & X_{23} & X_{24} \end{bmatrix}.$$

From table (4.6), we can calculate the sample mean of group 1

$$(\bar{X}_1)_{155\times 1} = \frac{1}{5}\sum_{i=1}^{5} X_{1i} = (3.5383830, \cdots, 2.8579431)',$$

and the sample mean of group 2

$$(\bar{X}_2)_{155\times 1} = \frac{1}{4}\sum_{i=1}^{4} X_{2i} = (1.0000000, 0.9723103, \cdots, 1.0000000)'.$$

Then we can calculate sample covariance matrix from group 1

$S_1 = \frac{1}{5-1}\sum_{i=1}^{5}(X_{1i} - \bar{X}_1)(X_{1i} - \bar{X}_1)' =$

$$\begin{bmatrix} 3.990402 * 10^{-03} & \cdots & -9.405345 * 10^{-09} \\ -2.399915 * 10^{-08} & \cdots & -5.432866 * 10^{-13} \\ \vdots & \ddots & \vdots \\ -9.405345 * 10^{-09} & \cdots & 8.721949 * 10^{-13} \end{bmatrix}_{155\times 155}, \tag{4.17}$$

and sample covariance matrix from group 2

$S_2 = \frac{1}{4-1}\sum_{i=1}^{4}(X_{2i} - \bar{X}_2)(X_{2i} - \bar{X}_2)' =$

$$\begin{bmatrix} 8.529508 * 10^{-13} & \cdots & 4.219947 * 10^{-13} \\ 7.376264 * 10^{-09} & \cdots & 1.643308 * 10^{-08} \\ \vdots & \ddots & \vdots \\ 4.219947 * 10^{-13} & \cdots & 1.627237 * 10^{-12} \end{bmatrix}_{155\times 155}. \tag{4.18}$$

Then from S_1, S_2, the pooled covariance matrix is

$S_{pl} = \frac{1}{(5+4-2)}[(5-1)S_1 + (4-1)S_2] =$

$$\begin{bmatrix} 2.280230 * 10^{-03} & \cdots & -5.374302 * 10^{-09} \\ -1.055254 * 10^{-08} & \cdots & 7.042439 * 10^{-09} \\ \vdots & \ddots & \vdots \\ -5.374302 * 10^{-09} & \cdots & 1.195784 * 10^{-12} \end{bmatrix}_{155\times 155}. \tag{4.19}$$

Since $p > n_1 + n_2 - 2$, S_{pl} is singular, and so S_{pl}^{-1} does not exist.

4.6.3 LDA on several groups

In discriminant analysis for several groups, we are concerned with finding linear combinations of variables that best separate groups of multivariate observations. Discriminant analysis for several groups may have many purposes:

- Examine group separation in a two-dimensional plot. When there are more than two groups, it requires more than one discriminant function to describe group separation. If the points in the p-dimensional space are projected onto a 2-dimensional space represented by the first two discriminant functions, we can obtain the best view of how the groups are separated.
- Find a subset of the original variables that separates the groups almost as well as the original set.
- Rank the variables in terms of their relative contribution to group separation.
- Interpret the new dimensions represented by the discriminant functions.
- Follow up to fixed-effects MANOVA.

Suppose we have a data set, which has N observations with K groups. There are p variables in each observation. The i^{th} group has distribution $F_i \sim F(\mu_i, \Sigma)$, where $\mu_i, i = 1, 2, \cdots, K$ is the p by 1 true mean vector and Σ is the p by p covariance matrix, and S_i is the p by p sample covariance matrix for i^{th} group for $i = 1, \cdots, K$. Then the p by p pooled standard covariance matrix can be estimated as following:

$$S_{pl} = \frac{1}{N-K} \sum_{i=1}^{K} (n_i - 1) S_i,$$

where n_i is the number of observations in the i^{th} group, $\sum_{i=1}^{K} n_i = N$, and $(S_i)_{p\times p}$ is the sample covariance matrix of the i^{th} group. We assign a new observation $(Y)_{p\times 1} = (y_1, y_2, \cdots, y_p)'$ to one of groups by minimizing the distance from Y to the center of each group. i.e.

$$\begin{aligned} Y \in \text{group } j \Leftrightarrow j &= Arg\ min_{i=1}^{K} D_i^2(Y) \qquad (4.20) \\ &= Arg\ min_{i=1}^{K} (Y - \bar{Y}_i)' S_{pl}^{-1} (Y - \bar{Y}_i), \end{aligned}$$

where $\bar{Y}_i$ is the sample mean of i^{th} groups. From (4.23), we can obtain a linear classification rule:

$$\begin{aligned} D_i^2(Y) &= Y'S_{pl}^{-1}Y - Y'S_{pl}^{-1}\bar{Y}_i - \bar{Y}_i'S_{pl}^{-1}Y + \bar{Y}_i'S_{pl}^{-1}\bar{Y}_i \\ &= Y'S_{pl}^{-1}Y - 2\bar{Y}_i'S_{pl}^{-1}Y + \bar{Y}_i'S_{pl}^{-1}\bar{Y}_i(\text{ Since } (Y'S_{pl}^{-1}\bar{Y}_i)' = \bar{Y}_i'S_{pl}^{-1}Y) \end{aligned}$$

The first term on the right can be neglected since it is not a function of i and does not change from group to group. The second term is a linear function of Y, and the third term does not involve Y. We thus delete $Y'S_{pl}^{-1}Y$ and obtain a linear classification function and denote it as $L_i(Y)$. If we multiply by $-1/2$, the classification rule becomes :

$$L_i(Y) = \bar{Y}_i'S_{pl}^{-1}Y - \frac{1}{2}\bar{Y}_i'S_{pl}^{-1}\bar{Y}_i \tag{4.21}$$

and assigning Y to the groups $i, i = 1, 2, \cdots, K$ with the largest $L_i(Y)$. This will be the same group for which $D_i^2(Y)$ in (4.23) is smallest, that is, the group whose mean vector $\bar{Y}_i$ is closest to Y. To highlight the linearity of (4.24) as a function of Y, we can express it as

$$L_i(Y) = a_i'Y + a_{i0} = a_{i1}y_1 + a_{i2}y_2 + \cdots + a_{ip}y_p + a_{i0},$$

where $a_i' = \bar{Y}_i'S_{pl}^{-1}$ and $a_{i0} = -\frac{1}{2}\bar{Y}_i'S_{pl}^{-1}\bar{Y}_i$.

4.7 LDA on HLdata

However, when the number of variables is too big(p is large), but only have few observations(n_i is small), the inverse of the pooled p by p covariance matrix S_{pl}^{-1} does not exist. We called this kind of data set as '**HLdata**', which means high dimensions and low sample size. For example, when $K = 2$, S_{pl}^{-1} exists only when $p < n_1 + n_2 - 2$. But how about $p \geq n_1 + n_2 - 2$? (Example 4.5.1) To avoid this problem, we use principal component analysis (section 4.5) to reduce the number of variables. We can choose the the first k principal components, such that $k < n_1 + n_2 - 2$. Then the pooled covariance matrix $(S_{pl}^{-1})_{k\times k}$ exists. How large k should be? We can consider the guidelines discussed in section 4.5. We apply PCA before using LDA, say LDA on HLdata, on the two groups and several groups cases:

4.7.1 LDA on HLdata : two groups

We have data with high dimensions but low sample size X_1 and X_2. Both of them have p variables and X_1 has n_1 observations and X_2 has n_2 observations, where $p > n_1, p > n_2$ and $p > n_1 + n_2 - 2$. So X_1 is a matrix with p rows and n_1 columns denoted as $(X_1)_{p\times n_1} = [X_{11} \; X_{12} \; \cdots \; X_{1n_1}]$ and X_2 is a matrix with p rows and n_2 columns denoted as $(X_2)_{p\times n_2} = [X_{21} \; X_{22} \; \cdots \; X_{2n_2}]$. When applying PCA, the first k components are chosen such that $k < n_1 + n_2 - 2$. That is, only k eigenvectors with the first k largest eigenvalues are picked and formed a matrix $(P)_{p\times k}$(section 4.5), where $P'P = (I)_{k\times k}$. The transformed data $(Y_1)_{k\times n_1} = [Y_{11} \; Y_{12} \; \cdots \; Y_{1n_1}] = P'X_1, (Y_2)_{k\times n_2} = [Y_{21} \; X_{22} \; \cdots \; Y_{2n_2}] = P'X_2$ that are with lower dimensions k. The covariance matrix for Y_1 is

$$\begin{aligned}
(S_1^Y)_{k\times k} &= \frac{1}{n_1-1}\sum_{j=1}^{n_1}(Y_{1j}-\bar{Y_1})(Y_{1j}-\bar{Y_1})' \\
&= \frac{1}{n_1-1}\sum_{j=1}^{n_1}(P'X_{1j}-P'\bar{X}_1)(P'X_{1j}-P'\bar{X}_1)' \\
&= \frac{1}{n_1-1}\sum_{j=1}^{n_1} P'(X_{1j}-\bar{X_1})(X_{1j}-\bar{X_1})'P \\
&= P'(\frac{1}{n_1-1}\sum_{j=1}^{n_1}(X_{1j}-\bar{X_1})(X_{1j}-\bar{X_1})')P \\
&= P'S_1^XP,
\end{aligned}$$

and similarly,

$$(S_2^Y)_{k\times k} = P'S_2^XP,$$

where S_1^X, S_2^X are the covariance matrix of X_1, X_2, respectively, and $\bar{X}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} X_{1j} = (\bar{x}_{11}, \cdots, \bar{x}_{1p}), \bar{Y}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j} = (\bar{y}_{11}, \cdots, \bar{y}_{1k})$. Then the pooled covariance matrix for Y_1 and Y_2 is

$$\begin{aligned} (S_{pl}^{Y})_{k\times k} &= \frac{1}{n_1+n_2-2}((n_1-1)S_1^{Y_1}+(n_2-1)S_2^{Y_2}) \\ &= \frac{1}{n_1+n_2-2}((n_1-1)P'S_1^{X_1}P+(n_2-1)P'S_1^{X_2}P) \\ &= P'S_{pl}^{X}P, \end{aligned}$$

where $(S_{pl}^{X})_{p\times p}$ is the pooled covariance matrix of X_1 and X_2.

So, we have relations between sample covariance matrices

$$\begin{aligned} (S_1^{Y})_{k\times k} &= P'S_1^{X}P && (4.22) \\ (S_2^{Y})_{k\times k} &= P'S_2^{X}P && (4.23) \\ (S_{pl}^{Y})_{k\times k} &= P'S_{pl}^{X}P && (4.24) \end{aligned}$$

Then from (4.16), the linear discriminant function for a vector $(Y)_{k\times 1}$ is

$$L(Y) = (\bar{Y}_1-\bar{Y}_2)'(S_{pl}^{Y})^{-1}Y - \frac{1}{2}(\bar{Y}_1-\bar{Y}_2)'(S_{pl}^{Y})^{-1}(\bar{Y}_1+\bar{Y}_2). \tag{4.25}$$

Since $Y = P'X$, then the linear discriminant function $L(X)$ for a vector $(X)_{p\times 1}$ is:

$$\begin{aligned} L(X) &= (P\bar{'}X_1 - P\bar{'}X_2)'(S_{pl}^{P'X})^{-1}P'X && (4.26) \\ &- \frac{1}{2}(P\bar{'}X_1 - P\bar{'}X_2)'(S_{pl}^{P'X})^{-1}(P\bar{'}X_1 + P\bar{'}X_2) \\ &= (\bar{X}_1-\bar{X}_2)'P(P'S_{pl}^{X}P)^{-1}P'X \\ &- \frac{1}{2}(\bar{X}_1-\bar{X}_2)'P(P'S_{pl}^{X}P)^{-1}P'(\bar{X}_1+\bar{X}_2) \\ &= a'X + a_0, \end{aligned}$$

where

$$a' = (\bar{X}_1-\bar{X}_2)'P(P'S_{pl}^{X}P)^{-1}P' \tag{4.27}$$

is a 1 by p vector and

$$a_0 = -\frac{1}{2}(\bar{X}_1-\bar{X}_2)'P(P'S_{pl}^{X}P)^{-1}P'(\bar{X}_1+\bar{X}_2) \tag{4.28}$$

is a real number. The thus scalar function $L(X)$ is linear in X.

After transformation $Y = P'X$, $((S_{pl}^{Y})^{-1})_{k\times k} = ((PS_{pl}^{X}P')^{-1})_{k\times k}$ exists since $k < n_1 + n_2 - 2$. From (4.17), we can use LDA on HLdata to classify a new p by 1 vector of measurements X as:

$$X \in \begin{cases} \text{group 1} & \text{if } L(X) > 0 \\ \text{group 2} & \text{if } L(X) < 0 \end{cases} \tag{4.29}$$

Example 4.7.1. *In Example 4.6.1, the Birds data of Table 4.5 showed that S_{pl}^{-1} does not exist, since $p = 155 > 7 = n_1 + n_2 - 2$. In this example, we use PCA on the data and obtain the first nine eigenvalues (others are too small that we do not list)*

$$\lambda_1 = 50, \lambda_2 = 14, \lambda_3 = 11, \lambda_4 = 9,$$

$$\lambda_5 = 2.1 * 10^{-9}, \lambda_6 = 2 * 10^{-9}, \lambda_7 = 1.9 * 10^{-9}, \lambda_8 = 1.5 * 10^{-9}, \lambda_9 = 1.3 * 10^{-30},$$

and associate eigenvectors form a 155 by 9 matrix:

$$\begin{bmatrix} -1.87 * 10^{-1} & 1.28 * 10^{-2} & \cdots & -2.3 * 10^{-1} \\ -1.46 * 10^{-1} & 1.62 * 10^{-2} & \cdots & 3.1 * 10^{-2} \\ \vdots & \vdots & \ddots & \vdots \\ -1.37 * 10^{-1} & 1.04 * 10^{-2} & \cdots & -3.3 * 10^{-2} \end{bmatrix}_{(155\times 9)} . \tag{4.30}$$

The first 4 eigenvalues $(50, 14, 11, 9)$ are significant larger than others. We choose $k = 4$ and $4 < 7 = n_1 + n_2 - 2$. So we have a 155 by 4 orthogonal matrix

$$P = \begin{bmatrix} -1.87 * 10^{-1} & 1.28 * 10^{-2} & -3.1 * 10^{-2} & -1.28 * 10^{-2} \\ -1.46 * 10^{-1} & 1.62 * 10^{-2} & -2.54 * 10^{-2} & 9.3 * 10^{-4} \\ \vdots & \vdots & \vdots & \vdots \\ -1.37 * 10^{-1} & 1.04 * 10^{-2} & -2.25 * 10^{-2} & 1.52 * 10^{-3} \end{bmatrix}_{(155\times 4)} . \tag{4.31}$$

The next step is that we transform original data set X_1, X_2 *which is a* 155 *by* 5 *matrix and* 155 *by* 4 *matrix, respectively, to new data*

$$(Y_1)_{4\times 5} = P'X_1 = \begin{bmatrix} -7.555960 & -7.555976 & \cdots & -8.0559306 \\ -4.378160 & -4.378172 & \cdots & -1.9267418 \\ -4.755289 & -4.755263 & \cdots & 0.0395366 \\ -3.172523 & -3.172525 & \cdots & 6.0114615 \end{bmatrix}, \tag{4.32}$$

$$(Y_2)_{4\times 4} = P'X_2 = \begin{bmatrix} 4.710504 & 4.710492 & 5.8084313 & 5.8084256 \\ -8.373081 & -8.373071 & 1.8426948 & 1.8427011 \\ 1.395684 & 1.395671 & -0.6301066 & -0.6300934 \\ -1.050878 & -1.050867 & -1.9466131 & -1.9466052 \end{bmatrix}. \tag{4.33}$$

From (4.20)-(4.22) and (4.25)-(4.27), we can calculate the pooled covariance matrix S^Y_{pl} *for transformed data* Y *as*

$S^Y_{pl} = P'S^X_{pl}P =$

$$\begin{bmatrix} 0.4355421 & 1.194099276 & -2.091258515 & 0.111862258 \\ 1.1940993 & 15.835917941 & 0.043450754 & -0.002324198 \\ -2.0912585 & 0.043450754 & 12.744057403 & 0.004070432 \\ 0.1118623 & -0.002324198 & 0.004070432 & 10.133204176 \end{bmatrix}. \tag{4.34}$$

$(S^Y_{pl})^{-1} =$

$$\begin{bmatrix} 102690913 & -690503641 & 1496108680 & -101245590 \\ -690503641 & 52379597 & -113490450 & 7680196 \\ 1496108680 & -113490450 & 245898845 & -16640618 \\ -101245590 & 7680196 & -16640618 & 1126114 \end{bmatrix}, \tag{4.35}$$

which exists.

From (4.30) and (4.31), we can calculate the coefficients a, a_0 *as:*

$$a_1 = (23843684851, 18751164819, \cdots, 17468593002)' \tag{4.36}$$

$$a_0 = -147060767959 \tag{4.37}$$

In order to show that LDA on HLdata works, we randomly pick a measurement

$$X_1 = (3.57, 2.96, \cdots, 2.8579437)'$$

in group 1.

The linear discriminant function for X_1 is

$$L(X_1) = a'X_1 + a_0 = 836486707001 > 0 \tag{4.38}$$

By (4.32), we can classify X_1 as in group 1.

In the same way, we randomly choose a measurement

$$X_2 = (1, 0.94, \cdots, 1)'$$

in group 2.

The linear discriminant function for X_2 is

$$L(X_2) = a'X_2 + a_0 = -836486875169 < 0 \tag{4.39}$$

By (4.32), we can classify X_2 as in group 2.

In Example 4.7.1, original LDA can not apply since $p = 155 >= 7n_1 + n_2 - 2$, but LDA on HLdata works and can classify the new measurement correctly.

4.7.2 LDA on HLdata : several groups

Suppose we have a data set, which has N observations $X_1, X_2, \cdots, X_N$ with K groups. There are p variables in each observation, that is each $X_i, i = 1, 2, \cdots, N$ is a p by 1 vector. The i^{th} group has distribution $F_i \sim F(\mu_i, \Sigma)$, where $\mu_i, i = 1, 2, \cdots, K$ is the p by 1 true mean vector and Σ is the p by p covariance matrix, and S_i^X is the p by p sample covariance matrix for i^{th} group for $i = 1, \cdots, K$. Then the pooled standard covariance p by p matrix can be estimated as following:

$$S_{pl}^X = \frac{1}{N-K} \sum_{i=1}^{K} (n_i - 1) S_i^X,$$

where n_i is the number of observations in the i^{th} group, $\sum_{i=1}^{K} n_i = N$, and $(S_i^X)_{p\times p}$ is the sample covariance matrix of the i^{th} group, but if $p > N-K$ then $(S_{pl}^X)^{-1}$ does not exist. So we should use PCA on the variables to find the largest k principal components, that is, the k eigenvectors forms a p by k orthogonal matrix P projecting X from dimension p to $Y = P'X$ with dimension k (section 4.5). We can choose k such that $k < N-K$. Then $(S_{pl}^Y)^{-1}$ exists. We assign a new observation $(Y)_{p\times 1} = (y_1, y_2, \cdots, y_p)'$ to one of groups by minimizing the distance from Y to the center of each group.

From (4.18) or (4.19) we have

$$\begin{aligned} X \in \text{group } j \Leftrightarrow j &= Arg\ min_{i=1}^{K} D_i^2(P'X) \qquad (4.40)\\ &= Arg\ min_{i=1}^{K} (P'X - P^{\bar{\prime}}X_i)'(P'S_{pl}^X P)^{-1}(P'X - P^{\bar{\prime}}X_i)\\ &= Arg\ min_{i=1}^{K} (X - \bar{X}_i)'P(P'S_{pl}^X P)^{-1}P'(X - \bar{X}_i), \end{aligned}$$

where S_{pl}^X is the pooled covariance matrix for all groups. This is equivalent to using the linear discriminant function

$$\begin{aligned} L_i(X) &= (P^{\bar{\prime}}X_i)'(P'S_{pl}^X P)^{-1}P'X - \frac{1}{2}(P^{\bar{\prime}}X_i)'(P'S_{pl}^X P)^{-1}P^{\bar{\prime}}X_i \qquad (4.41)\\ &= \bar{X}_i'P(P'S_{pl}^X P)^{-1}P'X - \frac{1}{2}\bar{X}_i'P(P'S_{pl}^X P)^{-1}P'\bar{X}_i\\ &= a_i'X + a_{i0}, \end{aligned}$$

where $a_i' = \bar{X}_i'P(P'S_{pl}^X P)^{-1}P'$ is a 1 by p matrix and $a_{i0} = -\frac{1}{2}\bar{X}_i'P(P'S_{pl}^X P)^{-1}P'\bar{X}_i$ is a real number. $L_i(X)$ is a linear function of a new p by 1 vector measurement X. We assign X to the groups i with the largest $L_i(X)$. From LDA on HLdata, we can classify the data with many variables but few observations, but we should address here is that choosing the number of components k is very important. It should satisfy the guidelines in section 4.5 and be less than $N-K$, where N is the total number of samples analyzed and K is the number of groups.

4.8 Application of the LDA on HLdata

In the DNA barcode data, it is usually large number of samples (specimens), but the COI sequences are very similar between species and noisy. It causes large misclassification rate. For example, we use training-test data technique to evaluate the performance of LDA in (4.1) on data set, Birds of North America. We randomly select testing sets from whole data with size from 25 to 150. Training sets are used to estimate the parameters in LDA and then we calculate the misclassification rate from test sets. We repeat this procedure for 100 times and calculate the mean of the misclassification in each test set(Figure 4.1). The mean misclassification rate increases when the size of test set increases. The result is not good because the mean misclassification rate is greater 10% for size of test set > 70.

To improve this situation, we introduce a method called pairwise comparison. The procedure of pairwise comparison is that if we classify a new sample X to one of K species, we use LDA on HLdata to classify X to one of two species, say species i versus species $j(\neq i)$, for $\binom{K}{2}$ times. We calculate the times that X classified to species $i, i = 1, 2, \cdots, K$, and assign X to a species j if species got a largest number of times. That is, at each comparison (species i versus species $j(\neq i)$), if the specimen belongs to species i using LDA on HLdata, then we add 1 to the i^{th} position of an array $A = (a_1, \cdots, a_K)$. There are total $\binom{K}{2}$ comparisons. After all pairwise comparisons, we assign the new specimen X to the species i if $a_i = max_{j=1}^{K} a_j$. However when N is large, $NK(K-1)/2$ pairwise comparisons are not an easy task. In order to avoid this problem, the two-step procedure will reduce the workload.

Example 4.8.1. *To show the performance of the pairwise comparison, we choose a subset (468 samples $N = 468$) of Birds of North America with 6($K = 6$) species, whose name(sample size) are Branta (156), Dendroica(75), Larus(65), Anas(64), Vireo(62), and Aythya(46). We randomly pick the n samples as the testing set from 468 samples, where $n = 5, 10, 15, \cdots, 150$, and calculate the misclassification rate. So for each sample in testing set, there are $\binom{6}{2} = 15$ comparisons. We repeat this procedure 100 times and report the mean of the misclassification rate(Fig. 4.2). We get a very good result that the misclassification rate is under 10 percent. The improve is significant from global LDA (Fig 4.2).*

4.9 Discussion

Robin Lloyd, who is a LiveScience senior editor, said on Nov. 13, 2007 that "a stunningly egotistical Swedish naturalist, Carl Linnaeus, wrote a book called "Systema Naturae," first published in 1735 at 13 pages long, and proposed a hierarchical system for classifying plants, animals and minerals and launched an effort to identify and inventory all the world's living things. Now 250 years after publication of the book's latter editions, scientists still have discovered as few as 10 percent of the species now living on Earth, said Harvard biologist Edward O. Wilson, who spoke last week at an event at the New York Botanical Garden to celebrate a visit of Linnaeus' personal copy of the book's first edition. After 250 Years of Classifying Life, **90 Percent of species remain unknown**. $\cdots$. For instance, the number of species of nematodes or roundworms, the most abundant animals on Earth, stands at about 16,000 species known, but the numbers of actual species could run into the millions, experts estimate".

Now, the Barcode of Life Initiative (BOLI) began in 2003 with a proposal that scientists could tell species apart by using a very short gene sequence from a standardized position in the genome. Since that time, DNA barcoding has begun to emerge as a global standard for assigning biological specimens to the correct species. Research projects on insects, birds, fish, algae, and many other taxonomic groups are underway, and many more are being planned. Some are global research campaigns involving dozens to hundreds of contributors, and others are the work of a small team focusing on a small taxonomic group. All these barcoding projects share the goal of building an open-access database of reference barcodes that will improve our understanding of biodiversity and will allow non-taxonomists to identify species. We also joint projects and attended the international DNA barcode conferences. We hope we can contribute our statistical techniques on BOLI. However, the technique of DNA barcode analysis is a new area in statistics and very few statisticians involved this projects. Our mission is that we give some new statistical methods to analyze this high dimension and low sample size data. Since DNA barcode is a categorical data set, we use OSR procedure, which optimizes the F statistic when converting the characters A, T, C, G to numerical numbers. We also introduced a method LDA on HLdata and two comparison procedure to classify the species and get a comparable results. Other mission is that we can hopefully involved more and more statisticians to joint this project. At Rutgers University, Center for Discrete Mathematics and Theoretical Computer Science Founded as a National Foundation Science and Technology Center, DIMCAS, also organized this project called ”The DNA Barcode Data Analysis Initiative (DBDAI)” http://dimacs.rutgers.edu/Workshops/DNABarcode/.

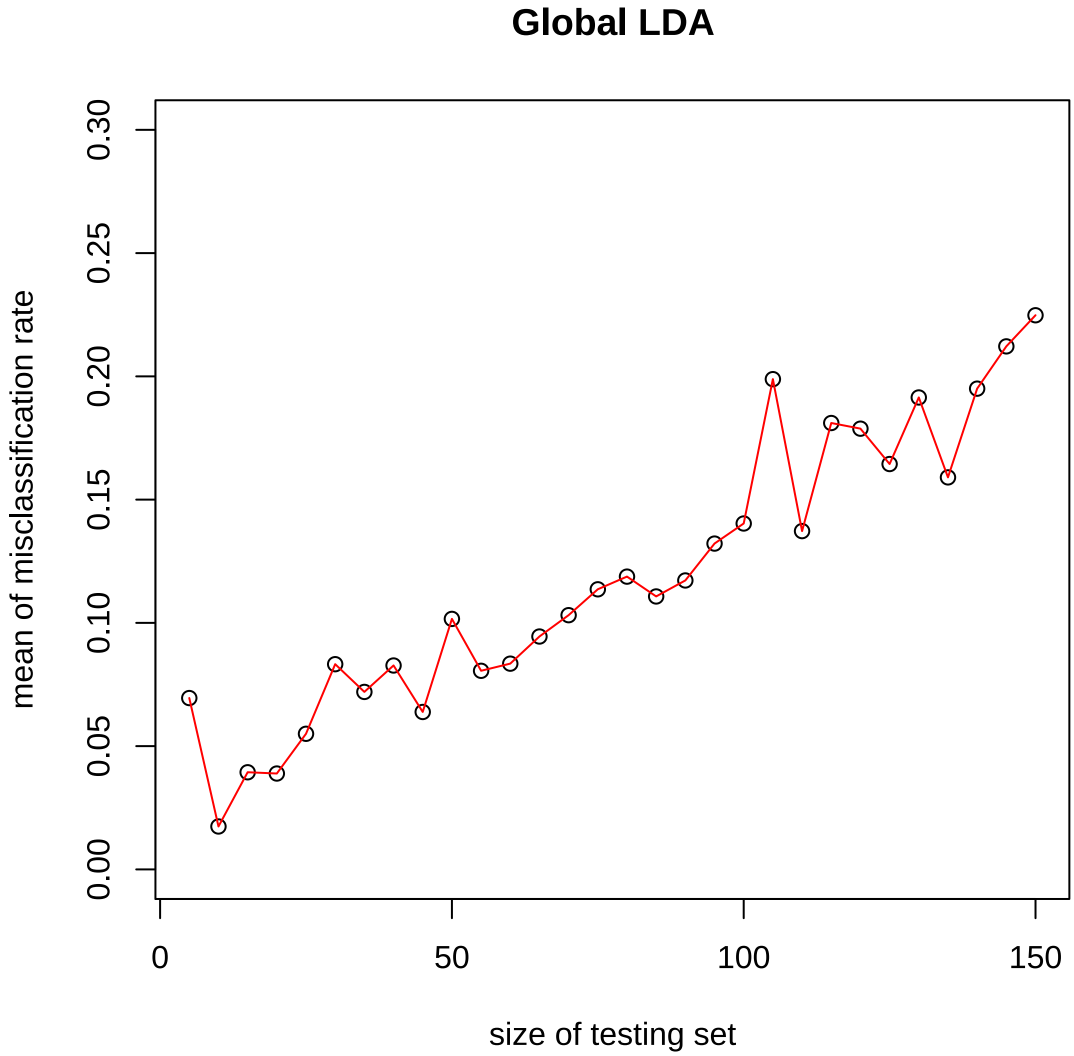


Figure 4.1: Birds of North America

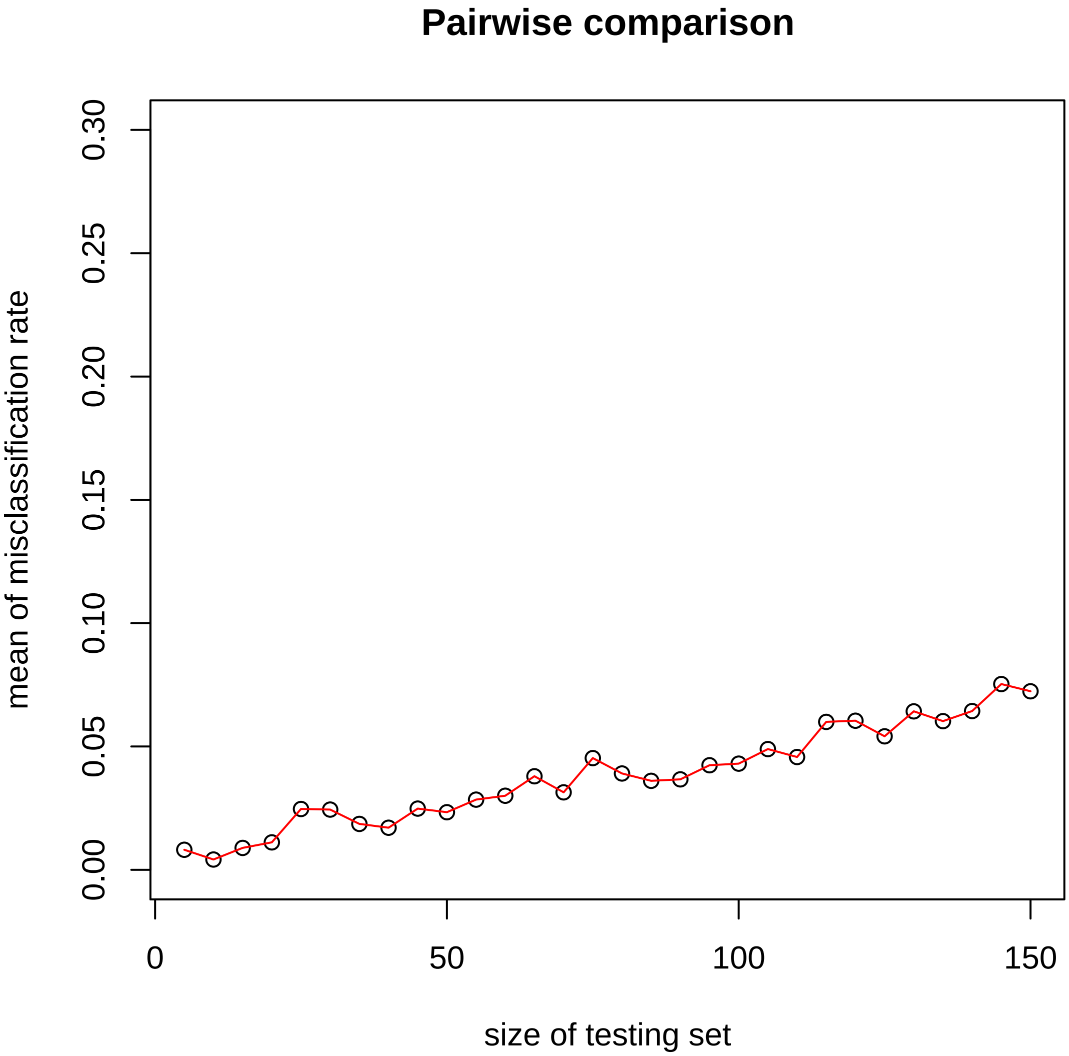


Figure 4.2: Pairwise comparison

References

[1] Adams, R. and Bischof, L.(1994). Seeded region growing. *IEEE transactions on Pattern Analysis and Machine Intelligence* **16**, 641–647.

[2] Alizadeh, A., Eisen, M., Davis, R., Ma,C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., Powell, J., Yang, L., Marti, G., Moore, T., Hudson, J., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., Weisenburger, D., Armitage,J., Warnke, R., Levy, R., Wilson, W., Gever,M., Byrd, J., Botstein, D., Brown, P., Staudt, L.(2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403** (6769), 503–511.

[3] Alon, U., Barkai, N., Notterman, D., Gish, K., Ybrra, S., Mack, D., Levine, A.,(1999). Broad patterns of expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS*, **96**, 6745–6750.

[4] Amaratunga, D. and Cabrera, J (2001). Statistical analysis of viral microchip data. *Journal of the American Statistical Association,* **96** 1161–1170.

[5] Amaratunga, D. and Cabrera, J. (2004). *Exploration and Analysis of DNA Microarray and Protein Array Data.* Wiley, New York.

[6] Amaratunga, D. and Cabrera, J. (2006). Differential expression in DNA microarray and protein array experiment. Technical Report 06-001, Department of Statistics, Rutgers University.

[7] 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*, 7, 509–519.

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

[9] Beucher, S., Meyer, F.(1993). The morphological approach to segmentation: the watershed transformation. Mathematical morphology in image processing. *Optical Engineering* **34**, 433–481.

[10] Bildi, Pierre and Long, A.(2001). A Bayesian framework for the analysis of microarray expression data: regularity t-test and statistical inferences of gene changes. *Bioinformatics,* **Vol. 17**. No. 6, 509–519.

[11] Breiman, L.(1996). Bagging predictors. *Machine Learning* **24**, 123–140.

[12] Breiman, L.(1998). Arcing classifiers. *Annals of Statistics* **26**, 801–824.

[13] Breiman, L., Friedman, J., Olsen, R. and Stone, C.(1984). *Classification and Regression Trees.* Wadsworth, Monterey, CA.

[14] Broberg, P. (2003). Ranking genes with respect to differential expression. *Genome Biology,* **4** R41.

[15] Brown, C., Goodwin, P., and Sorger, P.(2000). Image metrics in the statistical analysis of DNA microarray data. *Proceedings of the National Academy of Sciences* **98**, 8944–8949.

[16] Brown, M., Grundy, W., Lin, D., Cristianini, N. Sugnet, C., Furey, T., Ares Jr, M., and Haussler, D.(2000). Knowledge-based analysis of microarray gene expression data by using support vector machine. *Proceedings of the National Academy of Sciences* **97**, 262–267.

[17] Buckley, M.(2000). *Spot User's Guide.* CSIRO Mathematical and Information Sciences, Sydney, Australia. http://www.cmis.csiro.au/iap/Spot/spotmanual.htm.

[18] Buhler, J., Ideker, T., and Haynor, D.(2000). Dapple: improved techniques for finding spots on DNA microarrays. University of Washton CSE Technical Report UWTR 2000-08-05.

[19] Burgoyne, P. S. (1982). Genetic homology and crossing over in the X and Y chromosomes of Mammals. *Hum. Genet.* **61**(2), 85-90e.

[20] Cabrera, J. and Fernholz, L.(1999). Target estimation for bias and mean square reduction. *Annals of Statistics.* **27**, 1080–1104.

[21] Chambers, J., Mallows, C., and Stuck, B. (1976). A Method for Simulating Stable Random Variables. *Journal of the American Statistical Association* **71**, 340–344.

[22] Chen, Y., Dougherty, E., and Bittner, M.(1997). Ratio based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics* **2**, 364–374.

[23] Chen, Y., Kamat, V., Dougherty, E. Bittner, M., Meltzer, P. and Trent, J. (2002). Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics,* **Vol. 18**. No. 9. 1207–1215.

[24] Cleveland, W. (1979). Robust locally weighted regression and smoothing scatter plots. *J. AM. Stat. Assoc.,* **74**, 829–836.

[25] Crick, F. (1970). Central Dogma of Molecular Biology. *Nature* **227**, 561-563.

[26] Daniel, W. (1990). *Applied Nonparametric Statistics*, Second Edition, Boston: PWS-KENT.

[27] Debashis, G. and Chinnaiyan, A. (2002). Mixture modeling of gene expression data from microarray experiments. *Bioinformatics.* **18** (2): 275–286.

[28] Dempster, AP, Lair, N.,Rubin, D.(1976). Maximum likelihood estimation from incomplete data using the EM algorithm. *Journal of Royal Statistical Society, Series B* **39** 1–38.

[29] Dudley, Amimee M., Aach, John, Steffen, Martin A., and Church, George M. (2002). Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *PNAS* **99**, No. 11, 7554–7559.

[30] Dudoit, S., Yang, Y., Speed, T., and Callow, M.(2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12** ,111-140.

[31] Dudoit, S., Fridlyand, J., and Speed, T.(2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97**, 77–87.

[32] Dudoit, S., Yang, Y., and Bolstad, B.(2002). Using R for the analysis of DNA microarray data. *R News* **2**(1), 24–32.

[33] Dudoit, S., Yang, Y.(2002). Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data. In G. Parmigiani, E. Garret, R. Irizarry and S. Zeger, editors, *The Analysis of Gene Expression Data : Methods and Software.* Springer, New York.

[34] Durbin, B., Hardin, J., Hawkins, D., and Rocke, D.M. (2002). A variance-stabilizing transformation for gene-expression microarray data. Manuscript.

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

[36] Eisen, M.(1999). *ScanAlyze User Manual.* Standford University, Palo Alto. http://rana.lbl.gov.

[37] Eisen, M., Spellamn, P., Brown, P., and Bostein, D.(1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, **95**, 14863–14868.

[38] Finkelstein, D., Gollub, J., Ewing, R., Sterky, F. Somerville, S., and Cherry, J. (2001). Iterative linear regression by sector. In *Methods of Microarray Data Analysis. Papers from CAMDA 2000.* eds. S.M. Lin and K.F. Johnson, Kluwer Academic, 57–68.

[39] Fisher, R.A. (1934). *Statistical Methods for Researcher Workers.* Oxford University Press.

[40] Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *An. Eugen.*, **7**, 179–188.

[41] Flury, B. and Riedwyl, H. (1985). T^2 Tests, the Linear Two-Group Discriminant Function, and Their Computation by Linear Regression. *American Statistician,* **39**, 20–25.

[42] Gene Ontology Consortium (2000). Gene ontology: Tool for the unification of biology. *Nature Genet.* **25** 25–29.

[43] Glonek, G. and Solomon, P.(2002). Factorial designs for microarray experiments. Technical report, Department of Applied Mathematics, University of Adelaide, Australia.

[44] Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfielf, C., and Lander, E.(1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

[45] Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Stadut, L., Chan, W., Bostein, D., Brown, P.(2000). G̈ene shavingäs a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology 1*(2): research0003.1–0003.21.

[46] Hebert, P., Ratnasingham, S., and deWaard, J.(2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. Lond.* **B**(Suppl.), DOI 10.1098.

[47] Hebert, P., Cywinska, A., Ball, S., and deWaard, J.(2002) Biological identifications through DNA barcodes. *Proc. R. Soc. Lond.* **B**, 02PB0653.1.

[48] Hollander, M. and D. Wolfe (1999). *Nonparametric Statistical Methods*, 2nd ed. New York: Wiley.

[49] Huber, P. J. (1981) *Robust Statistics.* Wiley, New York.

[50] Ideker, T., Thorsson, V., Siegel, S., and Hood, L. (2000). Testing for differentially expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology* **7** (6) 805–817.

[51] Kayser et. al. (2000) Characteristics and Frequency of Germline Mutations at Microsatellite Loci from the Human Y Chromosome, as Revealed by Direct Observation in Father/Son Pairs, *American J. Human Genetics* **66**: 1580–1588.

[52] Khan, J., Wei, J., Ringner, M., Saal, Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C., Peterson, C., and Meltzer, P.(2001). Classification and diagnostics prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7, 673–679.

[53] Kepler, T. Crosby, L. and Morgan, K. (2000). Normalization and analysis of DNA microarray data by self-consistency and local regression. Santa Fe Institute Working Paper, Santa Fe, New Mexico.

[54] Kerr, M., Martin, M., and Churchill, G. (2000). Analysis of variance for gene expression data via mixed models. *Journal of Computational Biology* **8**, 625–637.

[55] Kerr, M. and Churchill, G.(2001). Experimental design for gene expression microarrays. *Biostatistics,* **2** 183–201.

[56] Koenker, R. and Park, B.(1994). An interior point algorithm for nonlinear quantile regression, *Journal of Econometrics*, **71** : 265–283.

[57] Kooperberg, C., Fazzio, T., Delrow, J. and Tsukiyama, T. (2002). Improved background correction for spotted cDNA microarrays. *Journal of Computational Biology* **9**, 55–66.

[58] Lazzeroni, L. and Owen, A.(2002). Plaid models for gene expression data. *Statistica Sinica*, **12**, 61–86.

[59] Lee, M. L.T., Kuo, F.C., Whitmore, G.A. and Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academic of Sciences,* **97** 9834–9839.

[60] Li, Cheng and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays : Expression index computation and outliers detection. *PNAS.* **Vol. 98**. No. 1 31–36.

[61] Lin, D., Yang, Y., Scolnick, J., Brunet, L., Peng, V., Speed, T., and Ngai, J.(2002). A spatial map of gene expression in the olfactory bulb. Technical report, Department of Molecular and Cell biology, University of California, Berkeley.

[62] Lipshutz, R. Fodor, S., Gingeras, T., and Lockhart, D.(1999). High density synthetic oligonucleotide array. *Natural genetics supplement,* **Vol. 21** 20–24.

[63] Lockhart, D., Dong, H., Byrne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14**, 1675–1680.

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

[65] Lonnstedt, I., Grant, S., Begley, G. and Speed, T.(2001). Microarray analysis of two interacting treatments: a linear model and trends in expression over time. Technical Report, Department of Mathematics, Uppsala University, Sweden.

[66] Marazzi, A.(1993). *Algorithms, Routines and S Functions for Robust Statistics.* Wadsworth & Brooks/Cole.

[67] Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis.* Academic Press, Cambridge.

[68] McLachlan, G.(1992) *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York.

[69] Nadon, R., Shi, P., Skandalis, A., Woody, E., Hubschle, H., Susko, E., Rghei, N., and Ramm, P.(2001).Statistical methods for gene expression arrays. In *Microarrays: Optical Technologies and Informatics,* M.L. Bittner, Y. Chen, A. Dorsel, and E. Dougherty(eds), Proceedings of SPIE, **Vol. 4266**, 46–55.

[70] Nadon R., and Shoemaker, J.(2002). Statistical issues with microarrays: processing and analysis. *Trend in Genetics,* **Vol. 18**. No. 5, 265–271.

[71] Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., and Tsui, K.W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *J. Comp. Biol.* **8** 37–52.

[72] Ohno, S.(1967). *Sex Chromosomes and Sex-Linked Genes.* Springer, Berlin.

[73] Pan, Wei, Lin, J., and Le, C.(2003). A mixture model approach to detecting differentially expressed genes with microarray data. *Functional and Integrative Genomics,* **Vol. 3**. No. 3, 117–124.

[74] Pan, Wei, Lin, J., and Le, C.(2002). How many replicates of arrays are required to detect gene expression changes in microarray experiment? A mixture model approach. *Genome Biology* **Vol. 3 (5)** 0022.1–0022.10.

[75] Parmigiani, G., Garrett, E., Anbazhagan, R. Gabreilson, E.(2002). A statistical framework for expression-based molecular classification in cancer. *JRSS B* **64**, part 4, 717–736.

[76] Pavlidis, P. et al. (2004). Using the Gene Ontology for Microarray Data Mining: A comparison of Methods and Application to Age Effect in Human Prefrontal Cortex. *Neurochemical Research* **29** 1213–1222.

[77] Quackenbush, J.(2001). Computational analysis of microarray data. *Nature Review Genetics* 2, 418–427.

[78] Raghavan, R., Amaratunga, D., Cabrera, J. Nie, A. Qin, J., and Mcmillian, M. (2006). On Methods for Gene Function Scoring as a Mean of Facilitating the Interpretation of Microarray Results. *J. Comp. Biol* **13**. No 3: 798–809.

[79] Riply, B.(1996). *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge.

[80] Rocke, D. and Durbin, B.(2001). A model for measurement error for gene expression arrays. *Journal of computational biology.* **Vol. 8**. No. 6: 557–569.

[81] Rose, A., Barret, T., Powell, J., Morton, M., Femandez, J., Zhang, Y., Cabrera, C., Amaratunga, D. and Shankley, N. (2003). Changes in rat gastric mucosal gene expression in response to gastrin-mediated cholecystokinin 2 receptor activation. Submitted.

[82] Schadt, E.,Li, Cheng, Su, Cheng and Wong, W.(2001). Analyzing high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry* **80**: 192–202.

[83] Schadt, E.,Li, C., Ellis, B. and Wong, W.(2002). Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry* **84**, S37, 120–125.

[84] Schena, M.(1999). *DNA Microarrays: A Practical Approach*, Oxford University Press.

[85] Schena, M., Shalon, D., Davis, D.R., and Brown, P. O., (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science,* **270(5235)** 467–470.

[86] Shaffer, J.(1995). Multiple hypothesis testing . *Annual Review of Psychology* **46**, 561–576.

[87] Skaletsky, H. et al.(2003). *Nature* **423**, 825-837 .

[88] Smyth, G. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**. No. 1, Article 3.

[89] Smyth, G. and Yang, Y. and Speed T. (2002). Statistical issue in cDNA microarray data analysis. *Methods in Molecular Biology* **224**, 111–136.

[90] Soille, P.(1999). *Morphological Image Analysis: Principles and Applications.* Springer, New York.

[91] Speed, T. and Yang, Y.(2002). Direct versus indirect design for cDNA microarray experiments. Technical report 616, Department of Statistics, University of California, Berkeley.

[92] Storey, J. and Tibshirani, R. (2001). Estimating false discovery rates under dependence with applications to DNA microarrays. Technical Report 2001-18, Department of Statistics, Stanford University.

[93] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E., Golub, T., (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *PNAS* **96**, 2907–2912.

[94] Ting, M-L, Kuo, F., Whitmore, G. and Sklar, J.(2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *PNAS,* **Vol. 97**. No. 18, 9834–9839.

[95] Tseng, G., Oh, M., Rohlin, L., Liao, J., Wong, W. (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Reseach* **29**, 2549–2557.

[96] Tusher,V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS,* **98** 5116–5124.

[97] Venable and Ripley (2003). *Modern Applied Statistics with S* Springer, New York.

[98] Walsh, Bruce (2001). Estimating the Time to the Most Recent Common Ancestor for the Y chromosome or Mitochondrial DNA for a Pair of Individuals. *Genetics society of America* **158**, 897–912.

[99] Wang, X., Ghosh, S. and Guo, S.(2001). Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Research* **29**(15), e75.

[100] Watson, J.D. and F.H. Crick (1953). A structure for deoxyribose nucleic acids. *Nature* **171** :737-738.

[101] Westfall, P. and Young, S. (1993). *Re-Sampling Based Multiple Testing.* Wiley, New York.

[102] Wodicka, L., Dong, H., Mittmann, M., Ho, M., and Lockhart, D.(1997). Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nat. Biotechnol.* **15**, 1359–1367.

[103] Yang, M., Ruan, Q-G, Yang, J., Eckenrode, S., Wu, S., McIndoe, R., and She, J-X.(2001). A statistical procedure for flagging weak spots greatly improves normalization and ratio estimates in microarray experiments. *Physiological Genomics* **7**, 45–53.

[104] Yang, T., Buckley, M., Dudoit,S.,and Speed, T.(2002). Comparison of methods for image analysis on cDNA microarrays. *Journal of Computational and Graphical Statistics* **11**, 108–136.

[105] Yang, Y., Dudoit,S., Luu,P., and Speed, T.(2001). Normalization for cNDA microarray data. In M.L. Bittner, Y. Chen, A.N. Dorsel, and E.R. Dougherty(eds.), *Microarrays: Optical Technologies and Informatics* Volume 4266 of Proceedings of SPIE.

[106] Yang, Y., Dudoit,S., Luu,P., Lin, D., Peng, V., Peng, V., Ngai, J. and Speed, T.(2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30(4)**, e15.

[107] Zhang, W., Shmulevich, I., Astola, J.(2004). *Microarray Quality Control*, John Wiley and Son.

[108] Zhao, L., Prentice, R., Breeden, L. (2001). Statistical modeling of large microarray data sets to identify stimulus-response profiles. *PNAS* **98**. No. 10,5631–5636.

[109] Array Vision, Imaging Research Inc. http://imaging.brocku.ca

[110] GenePix Pro microarray and array analysis software, Axon Instruments Inc. http://www.axon.com

[111] QuantArray Analysis Software. http://lifescience.perkinelmer.com

[112] R project CRAN. http://http://www.r-project.org

[113] Scanalytics MicroArray Suite. http://www.scanalytics.com

[114] Birds of North America.
http://dimacs.rutgers.edu/Workshops/BarcodeResearchChallenges2007/

Vita

Ching-Ray Yu

1995	BS in Mathematics, National Central University, Taiwan
1997	MS in Mathematics, National Central University, Taiwan
2008	Ph.D. in Statistics, Rutgers University, the State University of New Jersey