Statistical analysis of microarrays

Daphne Mouzaki

Statistical Analysis of Microarrays

Daphne Mouzaki

December 13, 2005

Statistical Analysis of Microarrays

Daphne Mouzaki
Thesis for MSc by Research
Department of Mathematical Sciences
University of Durham
Submitted November 2005

Abstract

Microarray statistical analysis involves thousands of hypothesis tests to consider at the same time. Empirical Bayes methods which are well-suited for large scale inference problems seem to be the most appropriate approach for microarray data. In this thesis we describe and compare Efron's ([3],[1],[4]) nonparametric empirical statistical analysis and Newton's and Kendziorski's ([12]) parametric empirical statistical analysis on microarray data. Both methods estimate Efron's ([3],[1],[4]) local false discovery rate, which identifies interesting genes and provides information about the power of the experiment.

Acknowledgements

Firstly, I would like to thank my supervisor Dr. Allan Seheult who has been encouraging, enthusiastic and patient throughout the period of the research.

Of course much needed financial and moral support has come from my parents. They have been positive throughout the period of the studies and maintained their interest in my work the whole time.

Finally, I would like to thank Dimos for his support during the course. Dimos was always optimistic and made me see things brighter. He has been supportive and always believed in me.

This thesis is an original work not previously submitted in whole or part for examination or publication. It is submitted in 2005 to the University of Durham, for the award of MSc by research.

Contents

Chapter 1

Introduction

1.1 Technical foundations for DNA microarray technology

Medical researchers have always sought new technologies and tools in order to be as precise as possible in diagnosis and to personalize medical care. This appeared likely to happen after the discovery of a vast amount of information about the DNA sequence of the human genome and the emergence of DNA microarray technology.

These two advances both in knowledge and technology encouraged biomedical research concerning the study of gene expression and also helped to discover the connection between particular genes and specific diseases. Through the compilation of knowledge about the DNA sequence novel genes could be identified and studied.

However, the great importance of the revelation of the DNA sequence, could not have been realized without the corresponding technological development. The great amount of information about the DNA sequence should be organized in such a way that the genes would be as quickly as possible classified and identified. DNA microarray technology measures the amount of transcription of a large number of genes and consequently helps to this process of classification and identification of hundreds or thousands genes.

1.2 What is DNA

All the instructions for making the structures and materials the body needs to function are included in DNA. Most living cells encode in their nucleus DNA.

Scientifically DNA is a molecule which consists of two long strands of nucleotides. There are four kinds of nucleotides which make up DNA: adenine (A), thymine (T), guanine (G) and cytosine (C). Each strand of DNA is made up from these nucleotides, linked together end to end; for example, a fragment of a single strand of DNA could be: A T C C T G. When a single strand finds its complement they hybridize; for example, the complement of the above DNA fragment is: T A G G A C.

An important process that reveals unique properties of a cell is the transcription of DNA information into messenger RNA (mRNA). mRNA is a chemical found in the cytoplasm of a cell and more rarely in its nucleus, similar to DNA, which is responsible for the transmission of DNA information in the cell and for protein synthesis. The scientific name for the process of the transmission of information contained in DNA into mRNA is " gene expression". In other words, the abundance and the kind of mRNA in a cell reveals which genes are expressed.

1.3 What is a DNA microarray?

DNA microarray technology was introduced in 1996 [9]. It was revolutionary because it allowed scientists to analyse expression of many genes in a single experiment and could be efficient when the sample of living cells to be studied was small. A microarray experiment is used to measure gene expression within a single sample or to compare gene expression among samples.

A DNA microarray could be either a plastic or a glass slide, often one by three inches long, on which single stranded molecules of DNA are arranged at different locations and every spot includes thousands of copies of a DNA strand. A microarray measures gene expression by exploiting the ability of messenger RNA (mRNA) to hybridize, ie to find its complementary single stranded DNA on the chip.

Suppose that a drug company wants to determine whether the painkiller, which is about to be introduced to the market, is harmful for the stomach. This issue could be examined by comparing gene expression activity in stomach cells on which the drug has been applied and stomach cells on which it hasn't, using DNA microarrays.

Necessary for the experiment is to construct or buy a microarray (or chip) and to obtain two samples of stomach cells. The drug should be applied to one of the two samples of stomach cells and not applied to the other sample. In what follows we will call the former treated cells and the latter untreated.

From both samples, treated and untreated, messenger RNA (mRNA) is extracted, which is the substance that makes up gene expression.

mRNA from the two samples is transcribed into cDNA. mRNA is used as a template to generate more stable complementary cDNA. Fluorescent labels are added to cDNA from treated and cDNA from untreated cells. cDNA from treated cells is labelled, say with red, and cDNA from untreated cells is labelled, say with green.

The mixture of the red and green labelled cDNA is applied to the DNA microarray. When cDNA from the sample finds its complementary sequence of bases on the chip, there is hybridization. Of course not all genes are always expressed. The DNA microarray is scanned and a special computer programme is used to calculate the red to green fluorescent ratio at each spot and to analyse results. This ratio estimates any possible changes in gene activity caused by the drug.

1.4 Application of microarray analysis

Microarray analysis, although revolutionary for microbiology and medical diagnosis, produces enormous amounts of data, that make statistical analysis complicated using traditional techniques. Empirical Bayes methods seem to be very effective in high dimensional inference problems and for this reason they are likely to be an effective approach for statistical analysis of microarrays.

Empirical Bayes methods, in contrast to other statistical techniques, do not apply statistical inferences for every gene separately but take advantage of a kind of information sharing among genes. Inference for each component is influenced by data from other components. This is what Robbins meant by the term "borrowing information"[10].

1.5 Empirical Bayes methods

In microarray analysis, researchers intend to measure gene expression of a vast amount of genes simultaneously. This can be done by using empirical Bayes methodology. Empirical Bayes methods take into account the common parameters and information shared by genes to make inferences about each gene in terms of this shared information.

For example, let's presume that $y = (y_1, \ldots, y_J)$ are measured expression levels for every gene j, where $j = 1, \ldots, J$ and $\mu = (\mu_1, \ldots, \mu_J)$ is a vector of the corresponding mean expression levels for every gene j, ie a hypothetical profile. Suppose that mean expression level μ_j depends on an unknown parameter λ. In a classical Bayesian formulation μ will be treated as ran-

dom with prior distribution $\pi(\mu|\lambda)$, where the parameter λ is assumed to be specified or elicited without reference to the data y. We write $f(y|\mu)$ for distribution of y given μ. Thus, in the Bayesian method, λ would be assumed known and posterior inference about μ would be based on:

$$p(\mu|y,\lambda) = \frac{f(y|\mu)\pi(\mu|\lambda)}{\int f(y|\mu)\pi(\mu|\lambda)d\mu} \tag{1.1}$$

In contrast with the above analysis, Empirical Bayes methods regard λ as unknown and estimate it using the whole dataset y. To be more specific λ is estimated by maximizing the marginal likelihood

$$m(y|\lambda) = \int f(y|\mu)\pi(\mu|\lambda)d\mu \tag{1.2}$$

In this case inference about μ is based on the estimated posterior distribution

$$p(\mu|y,\hat{\lambda}) = \frac{f(y|\mu)\pi(\mu|\hat{\lambda})}{m(y|\hat{\lambda})} \tag{1.3}$$

where $\hat{\lambda}$ is the maximum marginal likelihood estimate of λ.

Empirical Bayes methods can be either parametric [8] or nonparametric [10]. In the parametric case, it is supposed that the prior $\pi(\mu|\lambda)$ has a known parametric form, which means that the posterior distribution of μ can be easily calculated having estimated $\hat{\lambda}$. On the other hand, in the nonparametric case the prior is not specified. Then, for example, the posterior mean is expressed in terms of the unknown prior and the data are used to estimate the posterior mean directly. We will see in chapter three reference to Efron's work in which essentially he estimates the numerator and denominator of (1.1) nonparametrically when each μ_j, corresponds to expressed and non-expressed genes.

Chapter 2

Parametric empirical Bayes methods

In statistical analysis of microarrays, it is intended to describe gene expression levels using a probability distribution for measurements. We will see that this distribution describes each gene's behaviour, depending on information given from the other genes. This could involve the comparison between genes from two possible conditions or a more complicated situation for more than two cellular conditions. In what follows, we borrow the notation and method used by Newton and Kendziorski [12].

2.1 Two conditions

Let $x_j = (x_{j1}, x_{j2}, \ldots, x_{jI})$ denote I expression measurements taken on gene j, on either the original or on logarithmic scale. They are considered to be independent random deviations from a gene mean value μ_j, arising from an observation distribution

$$f_{obs}(\cdot|\mu_j)$$

In the case where the gene expression measurements describe two different conditions, the vector of observed expression values is partitioned into two subsets s_1, s_2 where s_k contains the indices from group k. If the distribution of mean expression measurements is not affected by this grouping, then there is equivalent expression EE_j for gene j. In the opposite case there is differential expression DE_j for gene j.

Regardless of whether we are comparing two or more conditions, the null hypothesis always refers to equivalent expression within genes.

Concentrating on the simple case of two conditions, the measurement

distribution for equivalently expressed genes is given by

$$f_0(x_j) = \int (\prod_{i=1}^{I} f_{obs}(x_{ji}|\mu))\pi(\mu)d\mu$$

where $\pi(\mu)$ is the probability distribution of the true expression level μ_j of any gene j. The corresponding distribution for differentially expressed genes can be written

$$f_1(x_j) = f_0(x_{j,1})f_0(x_{j,2})$$

where $x_{j,k}$ are the measurements for group $k = 1, 2$. The proportion of differentially expressed genes is given by the unknown fraction p and the remainder, $1 - p$, describes the proportion of equivalently expressed genes.

It is clear that the marginal distribution of the data is a mixture of equivalently and differentially expressed genes given by

$$pf_1(x_j) + (1 - p)f_0(x_j)$$

Inference about the proportion of differentially expressed genes is based on the posterior probability that gene j is differentially expressed:

$$\frac{pf_1(x_j)}{pf_1(x_j) + (1 - p)f_0(x_j)}$$

Notice that this posterior distribution will depend on the other genes because their measurements will be used to estimate unknown parameters in $\pi(\mu)$ and p.

2.2 Multiple conditions

Usually research concerns comparison among multiple patterns of mean expression. Although it is more difficult to deal with, the process of depicting data through a marginal distribution is quite similar with the simple case of two cellular conditions. The gene patterns will not be just two, EE and DE, but could be m distinct patterns. The null hypothesis in every situation refers to the equivalent expression of genes. When the analysis concerns m distinct patterns, marginal distribution is expressed by

$$\sum_{k=0}^{m} p_k f_k(x_j)$$

where p_k is the proportion of pattern k in the data and $f_k(x_j)$ is the probability distribution of measurements for each pattern of expression. Inference about expression pattern k is based on the posterior of k pattern given by

$$p(k|x_j) = \frac{p_k f_k(x_j)}{\sum_{k=0}^{m} p_k f_k(x_j)}$$

2.3 Gamma-Gamma model

One parametric application of the general mixture model is the Gamma Gamma model. The observation component is a Gamma distribution having shape parameter $\alpha > 0$ and a mean value μ_j, with scale parameter $\lambda = \frac{\alpha}{\mu_j}$

$$f_{obs}(x|\mu_j) = \frac{\lambda^\alpha x^{\alpha-1} \exp(-\lambda x)}{\Gamma(\alpha)}$$

for measurements $x > 0$. The coefficient of variation in this case is $\frac{1}{\sqrt{\alpha}}$ taken to be constant across genes. The mean effect of gene j is distributed according to $\pi(\mu_j)$ which in this case is an inverse Gamma. This means that with α taken as fixed, λ has Gamma distribution with shape parameter α_0 and scale parameter v. Therefore the unknown parameters will be $\theta = (\alpha, \alpha_0, v)$. The distribution for an equivalently expressed gene j in the Gamma-Gamma model, of the measurements $x_j = (x_{j1}, x_{j2}, \ldots, x_{jI})$ is:

$$f_0(x_1, x_2, \ldots, x_I) = k \frac{(\prod_{i=1}^{I} x_i)^{\alpha-1}}{(v + \sum_{i=1}^{I} x_i)^{I\alpha+\alpha_0})}$$

$$k = \frac{v^{\alpha_0}\Gamma(I\alpha + \alpha_0)}{(\Gamma(\alpha))^I \Gamma(\alpha_0)}$$

which is a specification of the general form

$$f_0(x_j) = \int (\prod_{i=1}^{I} f(x_{ji}|\mu))\pi(\mu)d\mu$$

(The proof for the above expression is given in the Appendix.)

2.4 Log-normal-Normal model

In this model the observation component is distributed according to the log-normal model and the mean expression of gene j, μ_j, is normally distributed

with mean μ_0 and variance τ_0^2. The coefficient of variation will be constant and equal to $\sqrt{\exp(\sigma^2) - 1}$ on the raw scale.

The density for $f_0(x_1, x_2, \ldots, x_I)$ will become Gaussian with mean vector $(\mu_0, \mu_0, \ldots, \mu_0)^t$ and covariance matrix equal to

$$\Sigma_I = \sigma^2 I_I + \tau_0^2 J_I$$

where I_I is the $I \times I$ identity matrix and J_I is the $I \times J$ matrix of ones. (The proof for the above expression, restricted to one-dimension is given in the Appendix).

2.5 EM algorithm

In order to obtain estimates of the proportion p and the parameters $\theta = (\alpha, \alpha_0, v)$ for the Gamma-Gamma model and $\theta = (\mu_0, \sigma^2, \tau_0^2)$ in case of Log-normal-Normal model, we use an EM algorithm.

The EM algorithm is used to calculate maximum likelihood estimates, 'ignoring' missing or unknown data. This algorithm involves two steps, the E-step and the M-step.

In the E-step the conditional expectation of missing data, in our case $\Phi = (\Phi_1, \Phi_2, \ldots, \Phi_J)$, is calculated given the observed components $x = (x_1, x_2, \ldots, x_J)$ and current estimates of θ and p, where $\Phi_j = 1$ when gene j is differentially expressed or 0 when gene j is equivalently expressed.

In the M-step a new value of θ is estimated using maximum likelihood based on the estimate of Φ in the E-step. This process is iterated until convergence.

More specifically, in microarray analysis the null hypothesis refers to equivalent expression. The proportion p of equivalently expressed genes is unknown. In order to estimate this proportion and the parameter θ, an EM algorithm is used.

Initially, a pattern indicator Φ_j is defined. $\Phi_j = 1$ when gene j is differentially expressed and $P(\Phi = 1|p) = p$ and $\Phi_j = 0$, when gene j is equivalently expressed with $P(\Phi_j = 0|p) = 1 - p$.

The joint distribution of (x_j, Φ_j) for gene j will be

$$P(x_j, \Phi_j|\theta, p) = P(x_j|\Phi_j, \theta, p)P(\Phi_j|\theta, p)$$

Writing $f_1(x_j|\theta)$ for the probability distribution of measurements for a differentially expressed gene given θ, and correspondingly $f_0(x_j|\theta)$ the probability distribution for an equivalently expressed gene given θ, the joint distribution

is given by

$$P(x_j, \Phi_j|\theta, p) = \begin{cases} pf_1(x_j|\theta), & \text{for } \Phi_j = 1 \\ (1-p)f_0(x_j|\theta), & \text{for } \Phi_j = 0 \end{cases}$$

which can be written $P(x_j, \Phi_j|\theta, p) = [pf_1(x_j|\theta)]^{\Phi_j}[(1-p)f_0(x_j|\theta)]^{1-\Phi_j}$. The log-likelihood of θ, p given x and Φ is

$$\log(P(x, \Phi|\theta, p)) = \sum_{j=1}^{I}[\Phi_j(\log f_1(x_j|\theta)+\log p)+(1-\Phi_j)(\log f_0(x_j|\theta)+\log(1-p))]$$

The E-step is performed to estimate θ, p as follows. The conditional expectation of $\log(P(x, \Phi|\theta, p))$ given x is:

$E(\log P(x, \Phi|\theta, p)|x) =$

$$\sum_{j=1}^{I} E(\Phi_j|x_j, \theta, p)(\log f_1(x_j|\theta) + \log p) + (1 - E(\Phi_j|x_j, \theta, p))(\log f_0(x_j|\theta) + \log(1-p)) \tag{2.1}$$

where

$$E(\Phi_j|x_j, \theta, p) = P(\Phi_j = 1|x_j, \theta) = \frac{pf_1(x_j|\theta)}{pf_1(x_j|\theta) + (1-p)f_0(x_j|\theta)}$$

Given current estimates $\theta^{(t)}$, $p^{(t)}$ for θ and p we estimate $E(\Phi_j|x_j, \theta, p)$ by

$$\Phi_j^{(t)} \equiv E(\Phi_j|x_j, \theta^{(t)}, p^{(t)}) = \frac{p^{(t)}f_1(x_j|\theta^{(t)})}{p^{(t)}f_1(x_j|\theta^{(t)}) + (1-p^{(t)})f_0(x_j|\theta^{(t)})}$$

In the M-step we replace $E(\Phi_j|x_j, \theta, p)$ with $\Phi_j^{(t)}$ in (2.1) and find the next estimate of θ, $\theta^{(t+1)}$ which maximizes

$$E(\log P(x, \Phi = \Phi^{(t)}|\theta, p = p^{(t)})|x)$$

Maximizing (2.1) with respect to p leads to the next estimate of p

$$p^{(t+1)} = \frac{1}{J}\sum_{j=1}^{I} \Phi_j^{(t)}$$

This process is repeated until convergence.

Chapter 3

Nonparametric empirical Bayes methods

Nonparametric Empirical Bayes methods are very effective for statistical analysis of microarray data. In this chapter it is intended to describe the application of nonparametric methods using an example of microarray experiment.

3.1 Example of nonparametric analysis

In order to demonstrate nonparametric statistical analysis, an example of a microarray experiment will be presented. The following example is from [2].

The study concerns the comparison of gene expression levels between 24 patients with more aggressive stomach cancer (Type 2) and 24 patients with less aggressive stomach cancer (Type 1). The aim of the study is to identify those genes that were more active or less active across the two conditions.48 microarrays were used, one for every patient and 2640 genes were measured for every microarray. A small part of those data is given in Table 5.1.

Assuming that the underlying data are approximately normally distributed, the most appropriate way of testing gene expression levels for every gene i between the two conditions, Type 1, Type 2 would be by using the two sample t-statistic. Taking that into account, it can be presumed that for every gene i there will be a corresponding two sample t-statistic, y_i. As in the parametric Empirical Bayes analysis, in the nonparametric case the statistical analysis is based on a simple hypothesis. Genes are divided in two categories: genes that are interesting (differentially expressed) and genes that are uninteresting (equivalently expressed). The former sometimes will be called non-null and the latter called null. The proportion of "non-null" genes is p_1 and

the proportion of "null" is $p_0 = 1 - p_1$ with prior densities $f_1(y)$ and $f_0(y)$ respectively. The mixture density of y is :

$$f(y) = p_1 f_1(y) + p_0 f_0(y)$$

It is important to determine the posterior probability of "non-null" genes given the data, which is given by:

$$\text{P("non-null"}|y) = \frac{p_1 f_1(y)}{f(y)} = 1 - \frac{p_0 f_0(y)}{f(y)}$$

In contrast with the parametric model where the densities $f_0(y), f(y)$ and the proportion of "non-null" and "null" genes p_1, p_0 can be easily estimated using maximum likelihood, in the nonparametric model $p_1, p_0, f_0(y), f(y)$ are unknown and have to be estimated using more complicated methods.

In the example, to estimate the distribution $f(y)$, a smooth curve is fitted to the histogram of the 2640 y_i values using a Poisson GLM fit (Figure 6.1). This process is explained in sections 3.2 and 3.3.

The prior density of "null" genes, $f_0(y)$ is assumed to be a Student's t-distribution with 46 degrees of freedom. The appropriateness of the t-distribution for $f_0(y)$ can be checked by using permutation methods (see [2]).

Finally, important for the estimation of the proportion of non-null genes given the y values, is to estimate the proportion of "non-null" and "null" genes p_1, p_0. p_1, p_0 are difficult to estimate without parametric assumptions about the densities $f_0(y), f(y)$. However, taking into account that the posterior probability of a non-null gene should be above or equal to zero, p_1 can be restricted to:

$$p_1 \geq 1 - \frac{f(y)}{f_0(y)}$$

for all y, or equivalently

$$p_0 \leq \frac{f(y)}{f_0(y)}$$

In conclusion, having estimated the densities $f(y), f_0(y)$ and the proportion of "non-null" and "null" genes p_1, p_0, the posterior probability of a "non-null"gene can be estimated from:

$$\text{P("non-null"}|y) = 1 - \frac{p_0 f_0(y)}{f(y)}$$

3.2 Poisson regression for density estimation

Estimates of probability densities such as $f(y)$ can be constructed using maximum likelihood when the parametric family is specified or by nonparametric methods such as the kernel density estimation. These two methods can be combined by putting an exponential family through a kernel estimator. To be exact, in order to obtain the estimator of a probability density $f(y)$, exponential families are used and the probability density estimate will be of the form:

$$f_\beta(y) = f_0(y) \exp(\beta_0 + \beta_1 t(y))$$

This formula includes two parts; an exponential term $\exp(\beta_0 + \beta_1 t(y))$ and $f_0(y)$ a carrier density. The estimated probability density will be equal to:

$$f_{\hat{\beta}}(y) = \hat{f}_0(y) \exp\left(\hat{\beta}_0 + \hat{\beta}_1 t(y)\right)$$

where $\hat{f}_0(y)$ is a normal kernel density estimator, $t(y)$ is the sufficient statistic and β_0, β_1 are estimated by maximizing the likelihood $\prod_{i=1}^{n} f_\beta(y_i)$. Calculations concerning these specially constructed exponential families are usually based to Poisson regression models introduced by Lindsey, (1974a, b). In other words, Poisson regression is an effective way of fitting specially designed exponential families by using generalized linear model software, such as the glm function in R.

3.3 Example of Poisson regression estimation

The following example is from Efron and Tibshirani, 1996 [5]. Suppose that $y = (y_1, \ldots, y_{67})$ is a vector of 67 pain scores obtained from a questionnaire administered to 67 women after an operation. y_i with $i = 1, 2, \ldots, 67$ runs from 0 to 4 ie $y_i \in [0, 4]$ where 0 = no pain and 4 = worst pain.

To estimate the probability density $\hat{f}(y)$, Poisson regression methods are applied. Particular fitted densities $\hat{f}_0(y)$ and $\hat{f}(y)$ are demonstrated in Figure 6.2.

The sample space $[0, 4]$ is partitioned into 40 cells and every cell's length is 0.1. The respective counts in each cell represented in Table 4.2 are $s_1, s_2, \ldots, s_{40}$. For example s_1 is the number of counts in $[0, 0.1]$, s_2 is the number of counts in $[0.1, 0.2]$, etc.

It can be observed, that the sum of all counts is the total number of pain scores ie $\sum_{j=1}^{40} s_j = 67$. If $\mathcal{Y} = [0, 4]$, $\mathcal{Y} = \bigcup_{j=1}^{40} \mathcal{Y}_j$ and $\{f_\theta(y), \theta e \Theta\}$ is a family of probability densities on $\mathcal{Y}$ then the probability of observing y_i on j cell is $\pi_j(\theta) = \int_{\mathcal{Y}_j} f_\theta(y) dy$.

In other words, the number of counts at every cell s_j can be thought as $s_j \sim P_0(\mu_j(\gamma,\theta))$, where $\mu_j(\gamma,\theta)$ is the expected number of counts at cell j ie $\mu_j(\gamma,\theta) = \gamma\pi_j(\theta)$, where γ is a positive parameter.

Taking into consideration that $s_j \sim P_0(\mu_j(\gamma,\theta))$ and using the specially designed exponential formula $f_\beta(y) = f_0(y)\exp(\beta_0 + \beta_1 t(y))$, $\mu_j(\gamma,\theta)$ can be written as

$$\mu_j(\beta) = \mu_j^0 \exp(\beta_0 + \beta_1 t_j)$$

where μ_j is proportional to $\pi_j^0 = \int_{y_j} f_0(y)dy$ and $t_j = t(y_{(j)})$, is the value of the sufficient statistic at the centre point $y_{(j)}$ of interval j and $\gamma = \exp(\beta_0)$ is a free parameter.

If $\mu_j(\beta)$ is expressed in the log-scale, $\log\mu_j(\beta) = \beta_0 + \beta_1 t_j + \log\mu_j^0$ which is a formula for the general linear model. It can be observed that $\log\mu_\beta = X\beta$ (ignoring the term $\log\mu^0$), where X is the 40×2 matrix.

$$X = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_{40} \end{bmatrix}$$

The parameter vector is

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

In order to estimate $f_0(y)$ and $f(y)$ it is necessary to estimate $\mu_j(\beta) = \mu_j^0\exp(\beta_o + \beta_1 t_j)$ and $\beta = (\beta_0, \beta_1)$, which are calculated using maximum likelihood equations for every parameter. In this case these equations will be summarised in:

$$X'[s - \mu(\hat{\beta})] = 0$$

where

$$s = \begin{bmatrix} s_1 \\ \vdots \\ s_{40} \end{bmatrix}$$

and $\mu(\hat{\beta})$ is the vector with j-th component equal to $\mu_j^0\exp(x_j\hat{\beta})$ (The proof of the above statement can be seen in the Appendix)

To complete the estimation we calculate $\hat{\mu}_j^0$ which is equal to:

$$\hat{\mu}_j^0 = M(\lambda)s$$

where s is the 40×1 vector

$$\begin{bmatrix} s_1 \\ \vdots \\ s_{40} \end{bmatrix}$$

and $M(\lambda)$ is a 40×40 smoothing matrix. The ki element of $M(\lambda)$ is

$$M_{ki}(\lambda) = \frac{c_k}{\lambda}\phi\left(\frac{y_{(k)} - y_{(i)}}{\lambda}\right)$$

where $y_{(k)} = \frac{k-0.5}{10}$ is the midpoint of cell $\mathcal{Y}_j$ and ϕ is the standard gaussian density. Note that the larger λ is the smoother will be the kernel estimate of f_0. In this particular case, $\lambda = 1$ and c_k is a positive constant, corresponding to midpoint $y_{(k)}$, chosen to make M_{s_+}, the normal kernel smoother M for the total $s_+ = 67$ counts, equal to 1. Having estimated μ_j^0 and β, it is very simple to estimate $f_0(y)$ and $f(y)$. $\hat{f}_0(y_)$ is nothing else but $\hat{\mu}_j^0$ plotted as a function of $y_{(j)}$ and $\hat{f}(y)$ is $\hat{\mu}_j$ plotted as a function of $y_{(j)}$ calculated from

$$\hat{\mu}_j = \hat{\mu}_j^0 \exp\{\hat{\beta}_0 + \hat{\beta}_1 t_j\}$$

when $t_j = (y_{(j)}, y_{(j)}^2)$, X is a 40×3 matrix for $\hat{f}(y)$, because in this case β has three components.

3.4 The local false discovery rate

An interesting approach to large scale inference problems is Efron's approach through local false discovery rate ([3], [1], [6]). Initially, it is presumed that the N genes in a microarray experiment are divided in two categories: "non-null" genes and "null" genes. As we discussed before, p_0 is the prior for the "null" genes and p_1 is the prior for the "non-null" genes. $f_0(y)$ and $f_1(y)$ are the densities for the null and non-null genes, respectively. The mixture density is:

$$f(y) = p_0 f_0(y) + p_1 f_1(y)$$

and the local false discovery rate is defined to be the posterior probability of being "null" given data y

$$\text{fdr}(y) = P\{\text{null}|y\} = \frac{p_0 f_0(y)}{f(y)}.$$

The local false discovery rate is a useful tool for identifying non-null genes or more generally interesting cases. Normally, the majority of genes in a microarray experiment are null genes and the non-null genes are a small proportion of the total cases. This means that p_0 is near 1 (usually $p_0 \geq 0.90$), so that

$$\text{fdr}(y) \approx \frac{f_0(y)}{f(y)}$$

Then it can be concluded, that cases with small fdr, usually are reported as interesting. Efron ([1]) chooses the cutoff point to be 0.2; ie $\text{fdr}(y) \leq 0.2$. When $\text{fdr}(y) \leq 0.2$, we can bound the posterior odds ratio as follows:

$$\frac{p_1 f_1(y)}{p_0 f_0(y)} = \frac{1 - \frac{p_0 f_0(y)}{f(y)}}{\frac{p_0 f_0(y)}{f(y)}} = \frac{1 - \text{fdr}(y)}{\text{fdr}(y)} \geq \frac{1 - 0.2}{0.2} \Leftrightarrow \frac{1 - \text{fdr}(y)}{\text{fdr}(y)} \geq 4$$

Taking into account that $p_0 \geq 0.90$ the prior odds ratio will be

$$\frac{p_1}{p_0} \leq \frac{1}{9}$$

From the above, the density ratio

$$\frac{f_1(y)}{f_0(y)} \geq 36$$

in favor of the "non-null" cases. On the contrary, if we take

$$\text{fdr}(y) \leq 0.6$$

$$\frac{p_1 f_1(y)}{p_0 f_0(y)} \geq \frac{2}{3}$$

and the prior ratio still

$$\frac{p_1}{p_0} \leq \frac{1}{9}$$

the Bayes factor will be only

$$\frac{f_1(y)}{f_0(y)} \geq 6$$

in favour of "non-null" cases. In applications, we will need to estimate $f_0(y), f(y)$ and p_0 to estimate $\text{fdr}(y)$. Poisson regression will be used to estimate $f(y)$, but (as we will see in chapter 4) estimation of $f_0(y)$ and p_0 is more subtle and depends initially on the prior assumption that p_0 is close to 1 and that the number of genes N is large.

Chapter 4

Nonparametric versus parametric statistical analysis

Although parametric and nonparametric empirical Bayes methods are different, their application to microarray analysis depends on the same simple assumption: genes are either differentially expressed or they are equivalently expressed.

In this chapter, it is intended to clarify the differences between the two methods and also to make them more comprehendable through their application to data from a particular microarray experiment.

4.1 Application of nonparametric methods to a microarray experiment

The following example is taken from Efron [1]. The methodology used is called a "nonparametric Empirical Bayes method".

In this microarray experiment, using the simple mixture model, genes are divided in two categories, genes that appear to be interesting (differentially expressed) and genes that appear to be uninteresting (equivalently expressed). The former sometimes will be called non-null and the latter called null. The proportion of null genes will be p_0 and the proportion of non-null genes will be $p_1 = 1 - p_0$. The null density and non-null density will be respectively $f_0(z)$ and $f_1(z)$. The mixture density from (3.1) is

$$f(z) = p_0 f_0(z) + p_1 f_1(z)$$

with subdensities $p_0 f_0(z)$ and $p_1 f_1(z)$, written as $f_0^+(z)$ and $f_1^+(z)$. As it has been stressed previously, this research focuses on estimating the probability

of genes being null given the data. Efron ([3], [1], [6]) refers to this posterior probability as the local false discovery rate, fdr,

$$\text{fdr}(z) = \frac{p_0 f_0(z)}{p_1 f_1(z) + p_0 f_0(z)} = \frac{f_0^+(z)}{f(z)} \tag{4.1}$$

In the example studied and discussed by Efron, 8 microarrays are used, 4 from HIV infected cells and 4 from uninfected cells. The total number of genes being measured is 7680. For the comparison of HIV infected and uninfected cells, every gene yields a two sample t-statistic t_i, $i = 1, \ldots, 7680$, each with 6 degrees of freedom. Let $F_6(t_i)$ denote the cumulative distribution function of t_i. The two sample t-statistic t_i can be transformed to a z score

$$z_i = \Phi^{-1}(F_6(t_i)) \tag{4.2}$$

where Φ is the cumulative distribution function of a standard normal variable. The histogram of the 7680 z-values of the microarray experiment is depicted in Figure 6.3. Two curves have been fitted to the histogram. The beaded curve estimates the null density $f_0(z)$ and the solid curve the mixture density $f(z)$. The null density $f_0(z)$ should be N(0,1) if t_i has a t-distribution with 6 degrees of freedom under the null hypothesis. Unfortunately this is not always the case. In some microarray experiments, as in the HIV experiment, the theoretical null does not satisfactorily represent the density for the null genes. In such cases an empirical null distribution should be fitted instead. In the HIV experiment the empirical distribution for the null genes is taken to be $N(\mu, \sigma^2)$, where μ and σ is estimated from Efron [1] to be 0.10 and 0.74 respectively.

The estimation of an empirical null distribution is based on the "zero" assumption, that null genes are concentrated mainly at the central peak of the histogram where the z values are approximately zero. Normally, null genes are the vast majority of the total of genes in a microarray experiment; that is, we expect p_0 to be large, usually above 0.9. Thus, null genes usually concentrate at the central peak of the histogram of z-values, whereas the non-null genes are at the extremes. Under-expressed genes appear on the left of zero, while over-expressed appear to the right.

As it has been noted before, it is important to estimate the posterior probability of null genes, the local false discovery rate, fdr. The process of estimating fdr involves the estimation of the mixture density $f(z)$ and the estimation of the subdensity $f_0^+(z)$. As discussed in sections 3.2, 3.3, the mixture density $f(z)$ may be estimated using Poisson regression methods. The subdensity $f_0^+(z)$ is more challenging to estimate.

To estimate the subdensity $f_0^+(z)$ we will assume that $f_0(z)$ is $N(\mu, \sigma^2)$ and use the log of the estimated mixture density $\hat{f}(z)$ of the histogram of

Figure 6.3. Taking into account that null cases are concentrated at the central peak of the histogram, a quadratic curve can be fitted to $\log \hat{f}(z)$ around $z = 0$ to give estimates of p_0, μ and σ^2. This assumes that $\log \hat{f}(z)$ around $z = 0$ is approximately equal to $\log \hat{f}_0^+(z)$ (Figure 6.4). In other words, p_0, μ, σ^2 are chosen to quadratically approximate the histogram counts near $z = 0$ (Figure 6.3), because $f_0(z)$ is taken to be the density of some normal distribution $N(\mu, \sigma^2)$. The estimates of μ, σ^2 indicate whether or not the theoretical null $N(0, 1)$ is suitable for a particular microarray experiment. Of course in the case where it is assumed that $f_0(z)$ is $N(0, 1)$ it is only necessary to select p_0 to quadratically fit the histogram heights near zero.

In the HIV microarray experiment the theoretical null $N(0, 1)$ was not suitable to describe the behaviour of null genes. This can be concluded from Figure 6.4 and from the impossible value of $p_0 = 1.15$ which was estimated from the quadratic approximation of Efron [1] to the histogram counts near $z = 0$. In Figure 6.4 it can be observed that $\log \hat{f}_0^+(z)$ when f_0 is N(0,1) is even more dispersed than $\log \hat{f}(z)$, which should not happen as $f_0^+ \leq f$.

Taking into consideration that the theoretical null is not suitable for the HIV microarray experiment and using the methodology discussed above, p_0 is estimated to be 0.917 and the empirical null is estimated to have probability density function

$$f_0(z) = \phi_{-0.10,0.74}(z) \tag{4.3}$$

where

$$\phi_{\mu,\sigma}(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2\right) \tag{4.4}$$

denotes the $N(\mu, \sigma^2)$ density.

According to Efron [1], the assumption that null genes should be concentrated near $z = 0$ is more valid when p_0 exceeds 0.90 because then the percentage of 0.10 or less of non-null genes only has a small effect on the estimate $\hat{f}_0^+(z)$ of the numerator of the fdr.

The solid curve in Figure 6.5 is the estimated fdr curve for the HIV data and the dashed curve, which estimates the effect size density for the non-null cases [1], will be discussed later in this section. As it has been noted in section 3.4, non-null genes are reported when the estimated fdr ≤ 0.2. The estimated proportion of non-null genes is $\hat{p}_1 = 1 - 0.917 = 0.083$ which means that we would expect $0.083 \times 7680 = 637$ non-null genes to be reported from the experiment. However, only 71 genes on the left of zero, those with $z_i \leq -2.34$, have fdr ≤ 0.2 and only 115 genes on the right of zero, those with $z_i \geq 2.17$, have fdr ≤ 0.2. This means that although we expect to identify 637 non-null genes, only 186 are identified as such. We consider this observation in more detail in section 4.3.

4.2 Estimation of the effect size density

Questions concerning the power of a test procedure, i.e the ability of the experiment to identify non-null cases and whether or not it is necessary to increase the experiment's size can be sufficiently answered through the estimation of the effect size density $g_1(\mu)$. Suppose that the expected z-values or true scores μ are randomly generated in two varieties: the null true scores and the non-null true scores. The effect size density of null scores is $g_0(\mu)$ and the effect size density of non-null scores is $g_1(\mu)$. To describe more clearly the estimation of $g_1(\mu)$ it will be assumed that null true scores and non-null true scores are generated from a one group model:

$$\mu \sim g(\mu) \qquad \text{and} \qquad z|\mu \sim N(\mu, \sigma_0^2) \tag{4.5}$$

with σ_0^2 fixed. Using the above model, the mixture density of z-scores can be expressed as $f(z) = g * N(0, \sigma_0^2)$ where * stands for the convolution of the two distributions. This will be the case if we can write $z = \mu + \epsilon$ when μ and ϵ are independent with $\mu \sim g_1(\mu)$ and $\epsilon \sim N(0, \sigma_0^2)$. The mixture density of z-scores is $f(z) = f_1^+(z) + f_0^+(z)$, so $f_1^+(z) = f(z) - f_0^+(z)$ and $g_1(\mu)$ can be obtained by inverting the subdensity $f_1^+(z)$ or by inverting $f_1(z)$ considering that $f_1(z) = \int_{-\infty}^{+\infty} f(z|\mu) g_1(\mu) d\mu$. Efron [1] applied to $f_1(z)$, Brown's [11] and Stein's [7] formula for the posterior mean function $E(\mu|z)$ under the one group model described above

$$\mu|z \sim (z + \sigma_0^2 l'(z), \sigma_0^2(1 + \sigma_0^2 l''(z))) \tag{4.6}$$

where $l(z) = \log f(z)$ and l', l'' indicate first and second derivatives. (The proof of 4.6 is described in the Appendix). When the above formula is applied to $f_1(z)$, $l_1(z) = \log f_1(z)$, the posterior mean function and the posterior standard deviation function of μ are :

$$\alpha(z) = z + \sigma_0^2 l_1'(z) \qquad b(z) = \sigma_0(1 + \sigma_0^2 l_1''(z))^{\frac{1}{2}}$$

The posterior density of μ given z for a non-null gene is

$$g_1(\mu|z) = \frac{g_1(\mu)\phi_{\mu,\sigma_0}(z)}{f_1(z)}$$

Assuming that $f(\mu|z)$ is $N(\alpha(z), b^2(z))$, a first guess for $g_1(\mu)$ is

$$g_1^{(0)}(\mu) = \int_{-\infty}^{+\infty} f_1(z)\phi_{\alpha(z),b(z)}(\mu)dz \tag{4.7}$$

To accept $g_1^{(0)}(\mu)$ it is important that $f_1^{(0)}(z) = g_1^{(0)} * N(0, \sigma_0^2)$ is close to f_1. To test this Efron uses the following measure of discrepancy from f_1:

$$D(f_1, f_1^{(0)}) = \int_{-\infty}^{+\infty} \frac{(f_1(z) - f_1^{(0)}(z))^2}{f_1(z)} dz \tag{4.8}$$

When $f_1^{(0)} = g_1^{(0)} * N(0, \sigma_0^2)$ is not equal to f_1, $g_1^{(0)}$ is updated to $g_1^{(1)}$ and then it is checked whether $f_1^{(1)} = g_1^{(1)} * N(0, \sigma_0^2)$ is close to f_1. Using $g_1^{(0)}(\mu|z) = \frac{g_1^{(0)}(\mu)\phi_{\mu,\sigma_0}(z)}{f_1^{(0)}(z)}$ we update $g_1^{(0)}$ to $g_1^{(1)}$ by noting that

$$g_1^{(1)}(\mu) = \int_{-\infty}^{+\infty} g_1^{(0)}(\mu|z) f_1(z) dz \tag{4.9}$$

and $f_1^{(1)} = g_1^{(1)} * N(0, \sigma_0^2)$

This process is iterated until $f_1^{(n)}(z) = g_1^{(n)} * N(0, \sigma_0^2)$, where n is the number of iterations, is approximately f_1 or in other words the distance, $D(f_1, f_1^{(n)})$ between $f_1^{(n)}$ and f_1 is small. Beginning with estimates of $f_1, g_1^{(0)}$ and $f_1^{(0)}$ and iterating the procedure of updating $g_1^{(n)}$ and estimating $f_1^{(n)} = g_1^{(n)} * N(0, \sigma_0^2)$ several times is how the effect size density $g_1(\mu)$ in Figure 6.5 was computed by Efron.

Although the above procedure was efficient in estimating the effect size density $g_1(\mu)$ Efron's 2005 study [4] describes easier techniques to estimate the non-null density $f_1(z)$, which we now discuss. Important for both methods that he describes as applied to the data from the HIV experiment is to divide the z-value sample space into 79 bins, each of width $\Delta = 0.1$. The bin counts $s_1, s_2, \ldots, s_{79}$ sum to 7680, the total number of z-values in the HIV experiment. If the probability density of z-values is $f(z)$, the probability density at midpoint $z_{(k)}$ of bin k will be $f(z_{(k)}) = f_k$ and the respective local false discovery rate $\text{fdr}(z_{(k)}) = \text{fdr}_k$. As it was noted in section 3.3, the s_k may be regarded as independent Poisson counts

$$s_k \sim P_0(\mu_k) \qquad k = 1, 2, \ldots, 79$$

where $\mu_k = 7680 \times 0.1 \times f_k$ is the expected number of cases in bin k. The non-null density $f_1(z)$ can be estimated using the fdr_k, f_k estimations for every bin or by fitting a regression curve to what Efron calls thinned counts, and will be explained later in this section. The first method uses the following facts:

$$f_1^+(z) = p_1 f_1(z) = (1 - \text{fdr}(z)) f(z) \tag{4.10}$$

$$p_1 = \int_{-\infty}^{+\infty} f_1^+(z) dz = \int_{-\infty}^{+\infty} (1 - \text{fdr}(z)) f(z) dz \tag{4.11}$$

so that

$$f_1(z) = \frac{(1 - \text{fdr}(z))f(z)}{p_1} \tag{4.12}$$

Using the notation above, we can approximate p_1 by

$$p_1 = \sum_{k=1}^{79}(1 - \text{fdr}_k)f_k \tag{4.13}$$

and $f_1(z_{(k)})$ by

$$f_{1k} = \frac{(1 - \text{fdr}_k)f_k}{p_1} \tag{4.14}$$

where p_1 is the proportion of non-null cases and f_{1k} is the non-null probability density at midpoint $z_{(k)}$ of bin k.

Estimating f_{1k} for every bin separately is how the non-null density $f_1(z)$ was drawn at Figure 6.6. However, this procedure involves the fitting to all 7680 cases, something that is not necessary if Poisson regression methods can be used to estimate $f_1(z)$.

Efron [4], uses Poisson regression methods to fit a regression curve to what he calls thinned counts. Thinned counts estimate the number of non-null cases in each histogram bin. The thinned count for bin k is equal to $s_{1k} = (1 - \text{fdr}_k)s_k$ which can be easily explained when we note that $(1 - \text{fdr}_k)$ is the probability of a non-null case in bin k and s_k is the number of cases in bin k. For each of the 79 bins there are $s_{11}, s_{12}, \ldots, s_{179}$ respective non-null cases which, as we explained before, are assumed to be independent Poisson counts

$$s_{1k} \sim P_0(\mu_{1k})$$

where μ_{1k} is the expected number of non-null cases, or the expected thinned counts in bin k, which is approximately $7680 \times 0.1 \times f_{1k}$, where 7680 is the total of z-values in the HIV experiment, 0.1 is the width of every bin and f_{1k} is the non-null density at midpoint $z_{(k)}$ of bin k. For the HIV experiment it is supposed that $\log \mu_{1k}$ is a cubic polynomial function, i.e $\log \mu_{1k} = \sum_{j=0}^{3} \beta_j z^j$. Using the Poisson regression methodology described in section 3.3, a regression curve is fitted to the thinned counts representing the over-expressed non-null genes; see Figure 6.6. The same methodology can be used to fit a regression curve to the under-expressed non-null genes.

4.3 Power estimation

An important issue is the power of a microarray experiment to identify non-null cases. In the HIV example we would expect that the proportion of non-null genes to be identified is $\hat{p}_1 = 0.083$. In other words, $0.083 \times 7680 = 637$

genes are estimated to come from $g_1(\mu)$. However, only 186 non-null genes are reported. This suggests poor power for the experiment.

Initially the low power of the experiment can be explained from the graphs in Figure 6.5 through the heavy curve of fdr and the dashed curve of $g_1(\mu)$. As it was said before, genes with fdr ≤ 0.2 are reported as non-null. In Figure 6.5 it can be observed that the total of non-null genes with fdr ≤ 0.2 are only 186. Although 637 genes are estimated to come from $g_1(\mu)$, the two modes of $g_1(\mu)$, the mode of under-expressed genes and the mode of over-expressed genes, correspond to fdr values exceeding 0.4, which is too large for identifying non-null genes.

The power of a microarray experiment can be assessed not only graphically but can also be estimated. A simple measure of the power of an experiment is "Efdr", the expectation of local false discovery rate for the non-null genes, given by

$$\text{Efdr} = \int_{-\infty}^{+\infty} \text{fdr}(z) f_1(z) dz \tag{4.15}$$

where, as before

$$f_1(z) = \int_{-\infty}^{+\infty} g_1(\mu) \phi_{\mu,\sigma_0}(z) d\mu \tag{4.16}$$

Large values of Efdr suggest low power while small values lead to the conclusion that the ability of the experiment to track non-null cases is satisfactory. If it is considered that fdr ≤ 0.2 indicates non-null genes, it can be quite clear why small values of Efdr indicate substantial power whereas large values of Efdr suggest low power. For example if Efdr $= 0.4$, forty percent of the genes that appear to be interesting are actually null genes. This value indicates the low ability of the experiment to identify "non-null" cases. We will discuss how Efdr may be estimated later in this section.

Another way to measure the power of the experiment is to check whether under-expressed and over-expressed genes appear on the list of genes having fdr ≤ 0.2. This can be done by estimating the probability of under-expressed genes having fdr ≤ 0.2 and separately estimating the same probability for the over-expressed genes, the non-null genes of the right mode of $g_1(\mu)$. For example in the HIV experiment the estimated probability of an under-expressed gene appearing on the list of genes having fdr ≤ 0.2 is

$$P(z \leq -2.34|\text{Non-null}) = \frac{\int_{-\infty}^{0} \phi\left(\frac{-2.34-\mu}{\sigma_0}\right) \hat{g}_1(\mu) d\mu}{\int_{-\infty}^{0} \hat{g}_1(\mu) d\mu} \tag{4.17}$$

which is equal to 0.210 and the respective probability for an over-expressed

gene is

$$P(z \geq 2.17|\text{Non-null}) = \frac{\int_0^{+\infty} \phi(\frac{2.17-\mu}{\sigma_0})\hat{g}_1(\mu)d\mu}{\int_0^{+\infty} \hat{g}_1(\mu)d\mu} \tag{4.18}$$

which is 0.43. Both results suggest the low power of the HIV experiment. The number 0.210 for the left mode of $g_1(\mu)$, suggests that only twenty-one percent of the under-expressed genes have fdr ≤ 0.2 and 0.43 for the right mode indicates that only forty-three percent of over-expressed genes have fdr ≤ 0.2. This is the reason that although the total of non-null genes is estimated to be 637, only 186 are identified. Observe that $(1 - (0.21 + 0.43))\,637 \approx 229$ which is close to the 186 identified as non-null.

Efron's 2005 research [4] applies improved methods to estimate the power of a microarray experiment and also investigates how an increase in the total number of subjects in the HIV study would improve its ability to track non-null cases. Both subjects are explained later on.

As discussed in section 4.2, f_{1k} is the non-null probability at midpoint $z_{(k)}$ of bin k and fdr_k is the local false discovery rate at $z_{(k)}$. Using the notation given in section 4.2, it can be concluded that the expected false discovery rate can be approximated as

$$\text{Efdr} = \sum_{k=1}^{79} \text{fdr}_k f_{ik} = \sum_{k=1}^{79} \text{fdr}_k \frac{(1 - \text{fdr}_k)f_k}{p_1} \tag{4.19}$$

by considering (4.5). As stated previously, small values of Efdr suggest substantial power and large values of Efdr indicate that the experiment's ability to identify non-null genes is not satisfactory.

Moreover, in order to answer to the question whether an increase in the number of subjects in the HIV experiment would improve its power, Efron assumes that for gene i, the mean and variance of $z_i|\mu$ is (μ_i, σ_0^2) where σ_0^2 does not vary with the true score μ (homoskedastic model). $z_i|\mu$ has expectation $\mu = 0$ for the null cases and consequently $z_i|\mu$ is $(0, \sigma_0^2)$. For the non-null cases μ has empirical mean and variance (α, β^2). The respective marginal mean and variance of z for non-null cases is $(A, B^2) = (\alpha, \beta^2 + \sigma_0^2)$, (The proof of this statement is given in the Appendix).

Suppose that l is the number of independent replicates of $z_i|\mu$. Then considering the new variable l the model for $z_i|\mu$ values will be

$$z_i^*|\mu = \sum_{j=1}^{l} \frac{z_{ij}}{\sqrt{l}} \tag{4.20}$$

with mean and variance $(\sqrt{l}\mu_i, \sigma_0^2)$. (This has been proved in Appendix). Considering l in the new model, the distribution of the null cases will still be $(0, \sigma_0^2)$. The marginal mean and variance of z^* for non-null cases will be $(A*, B*^2) = (\sqrt{l}\alpha, l\beta^2 + \sigma_0^2)$. The non-null marginal mean and variance of z^* can be estimated using Efron's formula

$$z^* = \sqrt{l}A + d(z - A), \qquad \left[d^2 = l - (l-1)\frac{\sigma_0^2}{B^2}\right] \tag{4.21}$$

(The proof of the above statement is given in Appendix). The marginal mean and variance (A, B^2) in the above formula are estimated using the thinned counts that we have explained in section 4.2. How the variable l affects non-null z-values on the right-side of the heavy curve $f_1(z)$ in Figure 6.6 and how it affects non-null z-values on the left-side of the $f_1(z)$ curve, is estimated using the thinned counts on the right-side and on the left-side respectively. To be more specific, the empirical mean and variance (A, B^2) are estimated separately for non-null z-values at the right-side of curve $f_1(z)$ in Figure 6.6 and for non-null z-values on the left-side of this curve, using the respective thinned counts. For example, using the right-side thinned counts, i.e the ones that express the number of over-expressed genes in bin k, the empirical mean and variance of the right-side non-null z-values are:

$$A = \frac{\sum z_{(k)} s_{1k}}{\sum s_{1k}} \qquad \text{and} \qquad B^2 = \frac{\sum z_{(k)}^2 s_{1k}}{\sum s_{1k}} - A^2 \tag{4.22}$$

Of course $\sum s_{1k}$ is the sum of the thinned counts which correspond only to the bins at the right-side of the f_1 curve. The same applies for the left-side calculations. The empirical mean and variance for the right-side of non-null cases in the HIV example are estimated to be $\hat{A} = 2.23$ and $\hat{B}^2 = 0.87^2$. Having estimated $\hat{A}$, $\hat{B}^2$ and assuming that $\sigma_0^2 = 0.735^2$ we can estimate the right-side non-null z^*-values for $l = 1, 1.5, 2 \ldots$ using the formula $z^* = \sqrt{l}A + d(z - A)$. The process is the same for estimating the left-side non-null z^*.

The above calculations allow us to assess how large an experiment should be to have effective power (see Efron [4]).

4.4 Application of parametric statistical analysis

Parametric statistical analysis is based on the same two group model as the nonparametric statistical analysis. The basic difference between the two

methods is that in the former both the observation component $z|\mu$ and the mean component μ have a known parametric form. The parameters' estimates are obtained either directly using maximum likelihood of the known mixture density or indirectly using EM algorithm.

Initially we used maximum likelihood to obtain estimates of the parameters in the HIV experiment, assuming that $z|\mu$ is $N(\mu, \sigma^2)$. For under-expressed genes μ is $N(\mu_1, \sigma_1^2)$, for null genes μ is $N(\mu_2, \sigma_2^2)$ and for over-expressed genes μ is $N(\mu_3, \sigma_3^2)$. The distribution of under-expressed genes is $f_1(z)$ which is $N(\mu_1, \sigma_1^2 + \sigma^2)$, for null genes is $f_2(z)$ which is $N(\mu_2, \sigma_2^2 + \sigma^2)$ and for over-expressed genes the distribution function is $f_3(z)$ which is $N(\mu_3, \sigma_3^2 + \sigma^2)$ (The proof of this statement is given in the Appendix). The mixture density at z is:

$$f(z) = p_1 f_1(z) + p_2 f_2(z) + p_3 f_3(z) \tag{4.23}$$

where p_1, p_2, p_3 are the proportions of under-expressed, null and over-expressed genes respectively.

The log-likelihood (based on J genes) for $\theta = (\mu_1, \mu_2, \mu_3, \sigma_1'^2, \sigma_2'^2, \sigma_3'^2)$ and $p = (p_1, p_2, p_3)$, where $\sigma_1'^2 = \sigma_1^2 + \sigma^2$, $\sigma_2'^2 = \sigma_2^2 + \sigma^2$ and $\sigma_3'^2 = \sigma_3^2 + \sigma^2$ is:

$$\log P(z|\theta, p) = \sum_{j=1}^{J} \log f(z_j) \tag{4.24}$$

The initial estimates we used for $\theta = (\mu_1, \mu_2, \mu_3, \sigma_1'^2, \sigma_2'^2, \sigma_3'^2)$ and for $p = (p_1, p_2, p_3)$ were $\theta^{(0)} = (-2, 0, 2, 1, 1, 1)$ and $p^{(0)} = (0.1, 0.8, 0.1)$ respectively. The function optim in R provided updated estimates that maximize the likelihood $P(z|\theta, p)$. The estimates of θ and p converged to: $(-1.5597, -0.077, 1.935919, 0.7388, 0.75288, 1.267787)$ and $(0.065649, 0.9010, 0.0333)$, respectively. However, by comparing Figure 6.7 with Figure 6.5 we see that the results using the above parametric model for the HIV data do not agree with Efron's results. According to Efron the probability of genes being null given the data (local fdr) is approximately 1 for z close to zero and the respective percentages of under-expressed and over-expressed genes in that area are approximately zero. This can be observed from the shape of the solid curve of fdr and the beaded curve of $g_1(\mu)$ around zero in Figure 6.5. In Figure 6.7 we notice that both the blue and red curves, which are the probability of under-expressed genes given the HIV data and the probability of over-expressed genes given the HIV data, respectively, are minimized around zero for a smaller area than the one that appears in Efron's Figure. Moreover, according to Efron's results, the probability of under-expressed genes given the data is much larger compared to the probability of over-expressed genes

given the data, as is evident when the mode of the former is compared with the latter and vice versa in Figure 6.5. In Figure 6.7 the probability of under-expressed and over-expressed genes given the data does not differ.

Assuming the same one group model for z-values, we apply EM algorithm to obtain parameters' estimates. The likelihood based on a single z-value is:

$$(p_1 f_1(z))^{\Phi_1}(p_2 f_2(z))^{\Phi_2}(p_3 f_3(z))^{\Phi_3} \tag{4.25}$$

where $\Phi = (\Phi_1, \Phi_2, \Phi_3)$, $\Phi_1 + \Phi_2 + \Phi_3 = 1$ and $\Phi_1, \Phi_2, \Phi_3 = 1$ when the gene j is under-expressed, null or over-expressed respectively. The full complete log-likelihood (based on J genes) for $\theta = (\mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma^2)$ and for $p = (p_1, p_2, p_3)$ is:

$$\begin{aligned}
&\log P(z, \Phi|\theta, p) = \\
&= \sum_{j=1}^{J}(\Phi_{1j} \log p_1 f_1(z) + \Phi_{2j} \log p_2 f_2(z) + \Phi_{3j} \log p_3 f_3(z)) \\
&= \sum_{j=1}^{J}(\Phi_{1j} \log f_1(z) + \Phi_{2j} \log f_2(z) + \Phi_{3j} \log f_3(z)) \\
&\quad + J(\bar{\Phi}_1 \log p_1 + \bar{\Phi}_2 \log p_2 + \bar{\Phi}_3 \log p_3)
\end{aligned} \tag{4.26}$$

Differentiating with respect to μ_1, μ_2, μ_3, σ_1^2, σ_2^2 and σ_3^2 separately, the estimates of μ_1, μ_2, μ_3 are:

$$\hat{\mu}_1 = \frac{\sum_{j=1}^{J} \Phi_{1j} z_j}{\sum_{j=1}^{J} \Phi_{1j}} \qquad \hat{\mu}_2 = \frac{\sum_{j=1}^{J} \Phi_{2j} z_j}{\sum_{j=1}^{J} \Phi_{2j}} \qquad \hat{\mu}_3 = \frac{\sum_{j=1}^{J} \Phi_{3j} z_j}{\sum_{j=1}^{J} \Phi_{3j}} \tag{4.27}$$

the estimates of $\sigma_1'^2$, $\sigma_2'^2$ and $\sigma_3'^2$ are:

$$\hat{\sigma}_1'^2 = \frac{\sum_{j=1}^{J} \Phi_{1j}(z_j - \mu_1)^2}{\sum_{j=1}^{J} \Phi_{1j}} \quad \hat{\sigma}_2'^2 = \frac{\sum_{j=1}^{J} \Phi_{2j}(z_j - \mu_2)^2}{\sum_{j=1}^{J} \Phi_{2j}} \quad \hat{\sigma}_3'^2 = \frac{\sum_{j=1}^{J} \Phi_{3j}(z_j - \mu_3)^2}{\sum_{j=1}^{J} \Phi_{3j}} \tag{4.28}$$

if the values of the $\bar{\Phi}_{ij}$ are known, which they are not. (The proofs of these results are given in the Appendix). As we have discussed at section 2.5, the conditional expectation of $\log P(z, \Phi|\theta, p)$ given z is computed (E-step) and maximum likelihood estimation is performed (M-step) in order to produce the estimates of Φ_1, Φ_2, Φ_3. Following the above procedure given current estimates $\theta^{(t)}$, $p^{(t)}$ of θ and p, the estimates of Φ_1, Φ_2, Φ_3 for gene j are:

$$\hat{\Phi}_{1j} = E(\Phi_{1j}|z_j, \theta^{(t)}, p^{(t)}) = P(\Phi_{1j} = 1|z_j, \theta^{(t)}) = \frac{p_1^{(t)} f_1(z_j|\theta^{(t)})}{f(z)}$$

$$\hat{\Phi}_{2j} = E(\Phi_{2j}|z_j, \theta^{(t)}, p^{(t)}) = P(\Phi_{2j} = 1|z_j, \theta^{(t)}) = \frac{p_2^{(t)} f_2(z_j|\theta^{(t)})}{f(z)}$$

$$\hat{\Phi}_{3j} = E(\Phi_{3j}|z_j, \theta^{(t)}, p^{(t)}) = P(\Phi_{3j} = 1|z_j, \theta^{(t)}) = \frac{p_3^{(t)} f_3(z_j|\theta^{(t)})}{f(z)} \qquad (4.29)$$

so the estimated proportions of under-expressed, null and over-expressed genes will be:

$$\hat{p}_1 = \frac{\sum_{j=1}^{J} \hat{\Phi}_{1j}}{J} \qquad \hat{p}_2 = \frac{\sum_{j=1}^{J} \hat{\Phi}_{2j}}{J} \qquad \hat{p}_3 = \frac{\sum_{j=1}^{J} \hat{\Phi}_{3j}}{J} \qquad (4.30)$$

The above procedure was repeated until there was convergence in the estimates.

The estimates of θ and p obtained using the EM algorithm were: $(-1.9034, -0.1159, -0.4693, 0.4184, 0.7364212, 1.6144)$ and $(0.0312, 0.8655, 0.1033)$. These estimates agree with Efron's results, except for the over-expressed genes. Similarly, the fdr and the probability of under-expressed genes given z agree with Efron's results, whereas the corresponding probabilities for over-expressed genes differ.

Assuming the same one group model, except for the null genes where we suppose that μ is $N(\mu_2, 0)$, the distributions of null, under-expressed and over-expressed genes are respectively $N(\mu_2, \sigma^2)$, $N(\mu_1, \sigma_1'^2)$ and $N(\mu_3, \sigma_3'^2)$. The assumption that the distribution of null true scores μ is $N(\mu_2, 0)$, expresses an ideal situation very close to the "zero assumption" in the non-parametric analysis. The reason for this is that, as in Efron's nonparameric analysis, all z-values close to μ_2 are considered as null cases with variation equal to zero.

Using the above model we applied maximum likelihood to the mixture density of z-values and also the EM algorithm starting from the initial estimates $\theta^{(0)} = (-1.9, 0, 2, 0.6324, 1, 0.9487)$ and $p^{(0)} = (0.1, 0.86, 0.04)$.

The function optim in R provided estimates that maximized the mixture density likelihood. The updated estimates of θ and p were respectively: $(-1.3177, -0.0659, 1.7658, 0.3276, 0.7482, 1.1952)$ and $(0.0893, 0.8721, 0.0386)$. Although there was convergence with the initial values, the optim results were clearly incorrect using the above model: the local fdr is approximately 1 not only for z close to 0 but also for z close to -3, something which is impossible if we consider the assumption that null genes correspond to z-values close to μ_2, which is taken to be equal to zero.

For the same one group model for z-values, we applied the EM algorithm to obtain the following estimates for θ and p
$(-0.9840, -0.088, 1.280, 0.5911, 0.8319, 1.0455)$ and $(0.0656, 0.9082, 0.0262)$

after a small number of iterations. However, as the number of iterations increased the estimates of θ and p got worse. Neither EM algorithm nor optim gave satisfactory results for the above model. These results indicate that the assumption that z-values close to μ_2 correspond to null genes is not acceptable.

Chapter 5

List of tables

Gene	1	1...	1	1	2	2...	2	2	tval	pval
1	-0.22	-0.13	-1.23	0.13	-0.80	-0.36	-0.31	0.38	1.550	0.128
3	-0.83	-0.01	-0.50	-1.69	-1.89	0.33	-1.12	-0.27	0.850	0.400
4	-0.14	0.69	-0.86	0.27	0.67	1.10	0.42	-0.96	-0.310	0.758
5	0.03	0.25	0.34	0.97	-0.43	0.10	0.03	-1.03	-1.852	0.070
6	0.66	0.68	0.22	0.58	-0.04	-0.09	-0.04	1.11	-2.226	0.031
7	-0.64	-0.36	0.66	0.01	0.18	0.31	0.57	-0.53	0.356	0.723
8	-0.02	-0.15	0.84	-0.13	-0.56	-0.24	-0.39	-0.43	-0.020	0.984
9	0.71	-0.29	0.48	-0.03	-0.56	-0.78	-0.34	0.27	0.460	0.648
10	0.16	-0.04	-0.55	-1.83	-0.90	-0.41	0.56	-0.04	1.914	0.062

Table 5.1: (From B Efron, [1]) Expression levels for the first 10 genes in the stomach cancer microarray example. The total number of genes is 2640, 48 microarrays are being used : 24 for Type 1 and 24 for Type 2 tumors. tval stands for the two sample t-statistic comparing Type 2 with Type 1; pval(p-value) is the two sided significance level of tval, with 46 degrees of freedom

3	7	6	1	2	3	3	1	7	5
4	4	1	3	3	5	0	1	0	0
2	2	0	0	0	1	0	0	1	1
1	0	0	0	0	0	0	0	0	0

Table 5.2: (After Efron and Tibshirani,[5]) Counts of the pain score data $y_i e[0,4]$ partitioned into 40 cells of length 0.1

Chapter 6

List of Figures

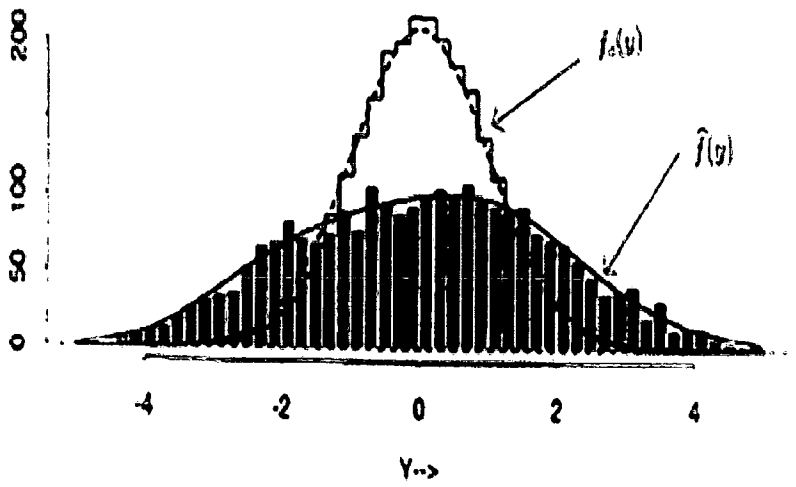

Figure 6.1: (After Efron,[2]) The histogram depicts the distribution of the 2640 two-sample statistics y_i; this is much wider than $f_0(y)$; solid curve $\hat{f(y)}$ is a smooth fit to the histogram and it is fitted using Poisson regression.

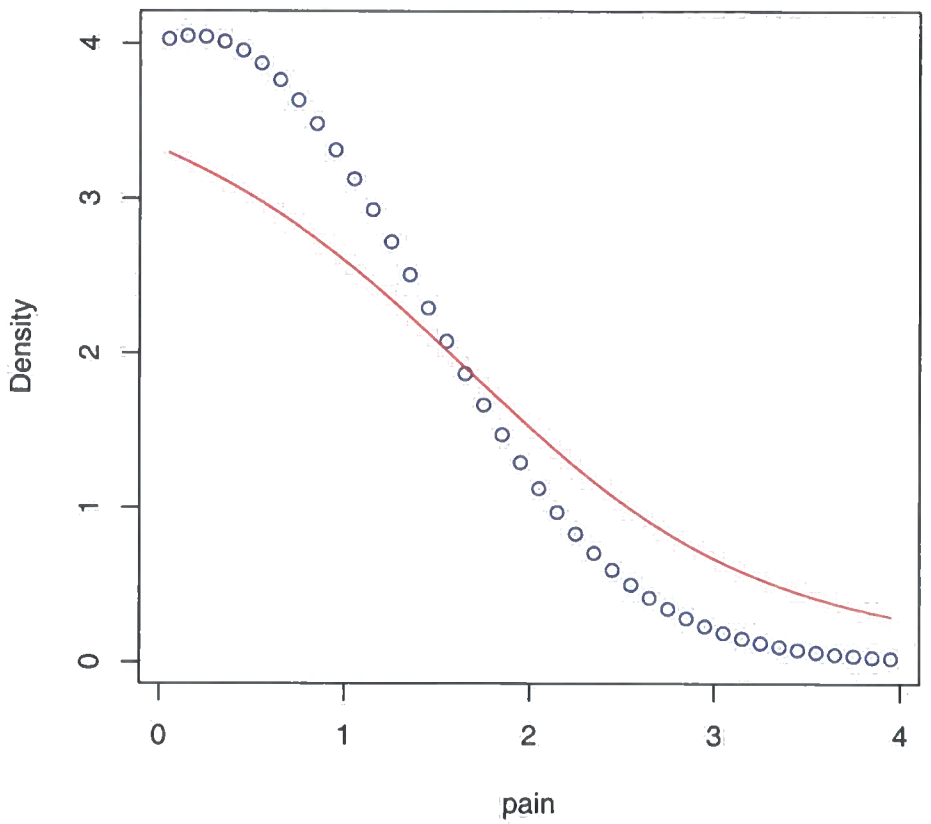

Figure 6.2: (After Efron and Tibshirani, [5]) The red curve is a normal kernel density estimator, $\hat{f}_0(y)$, with window width $\lambda = 1$; the dotted blue curve is the special exponential family density estimator, $\hat{f}(y)$, with base-line $\hat{f}_0(y)$ and sufficient statistic $t(y) = (y, y^2)$

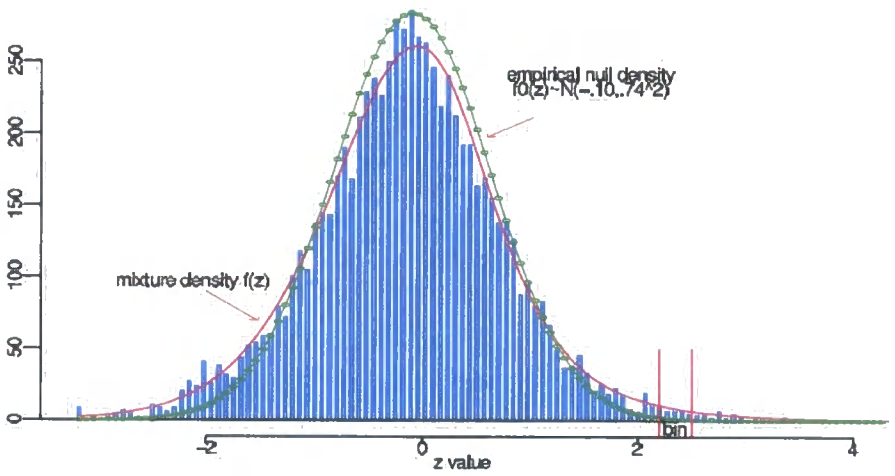

Figure 6.3: (After Efron, [1]) Histogram of 7680 z-values from an HIV microarray experiment. Solid curve estimates mixture density $f(z)$ and beaded curve estimates null density $f_0(z)$. The empirical null $f_0(z)$ is $N(-0.10, 0.74^2)$ with p_0 of null genes estimated as 0.917.

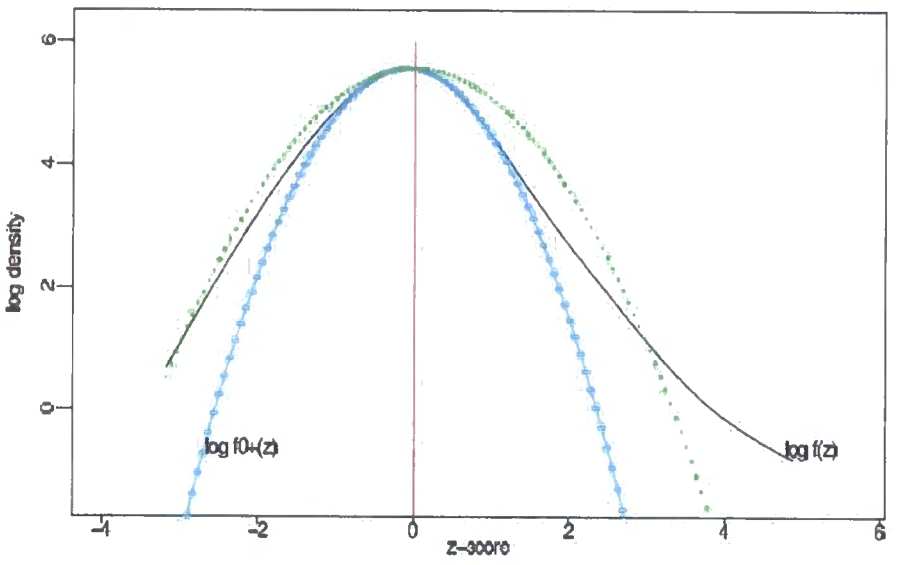

Figure 6.4: (After Efron, [1]) Estimating the empirical null for HIV data. Solid curve is $\log \hat{f}(z)$ and beaded curve a quadratic fit to $\log \hat{f}(z)$ near $z = 0$, gives estimate for $f_0^+(z)$. Dotted curve is the estimate of $f_0^+(z)$ when $f_0 \sim N(0,1)$.

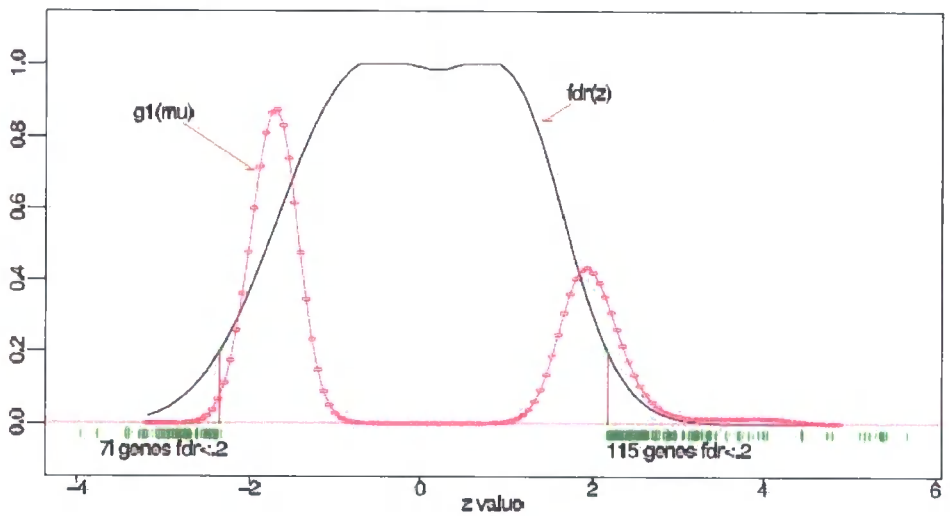

Figure 6.5: (After Efron, [1]) Solid curve is fdr estimated from the subdensity $f_0^+(z)$ and the estimate of the mixture density $f(z)$. 71 genes from the left have $fdr \leq 0.2$ and 115 genes from the right side of zero have $fdr \leq 0.2$. Only these genes are reported as under-expressed and over-expressed respectively. Beaded curve estimates the non-null effect size density $g_1(\mu)$.

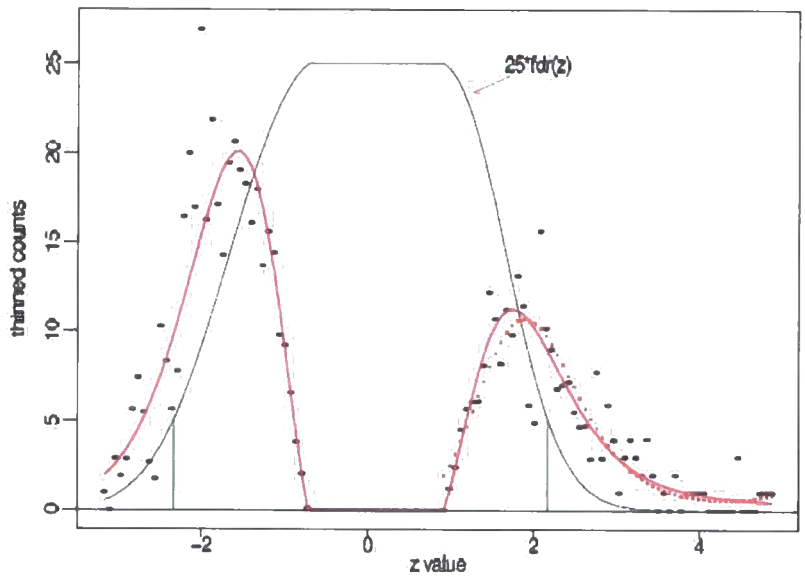

Figure 6.6: (After Efron, [4]) The heavy curve is the non-null density estimate $\hat{f}_1(z)$ for HIV study; light curve is the estimated $\text{fdr}(z)$. Points are what Efron calls thinned counts and the dotted curve is the regression curve that has been fitted to the thinned counts on the right

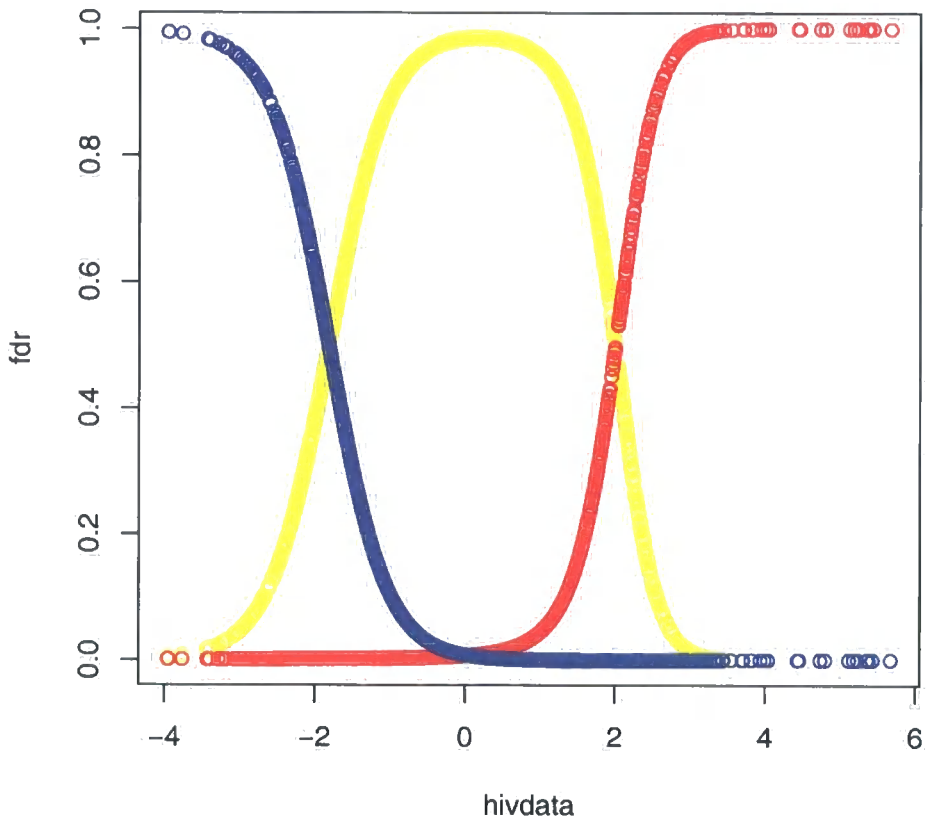

Figure 6.7: The yellow curve is the local fdr when $z|\mu$ is $N(\mu, \sigma^2)$, the distribution for null genes is $N(\mu_2, \sigma_2^2 + \sigma^2)$, the distribution of under-expressed genes is $N(\mu_1, \sigma_1^2 + \sigma^2)$ and the distribution of over-expressed genes is $N(\mu_3, \sigma_3^2 + \sigma^2)$. The left blue curve is the probability of under-expressed genes given the HIV data and the right red curve is the probability of over-expressed genes given the HIV data. These results were obtained using the function optim in R

Chapter 7

Conclusion

Empirical Bayes methodology is an efficient way of dealing with enormous data sets and that is one reason which makes it useful for statistical analysis of microarrays.

Although in individual hypothesis tests it is intended to reject the null, in large scale testing and particularly in microarrays the aim of the statistical analysis is to identify a small percentage of interesting cases. In microarrays, these interesting cases concern differential gene expression. This small percentage of interesting cases may give important information for the behaviour of these genes during a disease or during a drug application.

Empirical Bayes methods and false discovery rate are applied by dividing the data into two categories: genes that are equivalently expressed and genes that are differentially expressed. When the probability densities have been specified and they have an exact parametric form the process of calculating the posterior distribution of differential expressed genes is simple.

On the other hand, in nonparametric empirical Bayes analysis, Poisson regression fitting methodology is used to estimate the probability density for non-null cases and the mixture density of the whole data set. The density for null cases is based on the assumption that the proportion of such cases is expected to be close to one. After this estimation the small percentage of differentially expressed genes can be inferred.

A crucial part of both methods parametric and nonparametric is the local false discovery rate assessment, based on empirical Bayes analysis of the simple two group model described above. The local false discovery rate does not only help to identify non-null cases but also provides information about the power of the experiment and whether it is necessary to increase the experiment's size.

Bibliography

[1] Efron B. Selection and estimation for large-scale simultaneous inference. http://www-stat.stanford.edu/ brad/papers/Selection.pdf, 1986.

[2] Efron B. Robbins, empirical Bayes and microarrays. *The Annals of Statistics*, 31(2):366–378, 2003.

[3] Efron B. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *JASA*, 99:96–104, 2004.

[4] Efron B. Local false discovery rates. http://www-stat.stanford.edu/ brad/papers/Selection.pdf, 2005.

[5] Efron B. and Tibshirani R. Using specially designed exponential families for density estimation. *The Annals of Statistics*, 24(6):2431–2461, 1996.

[6] Efron B. and Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23:70–86, 2002.

[7] Stein C. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9:1135–1151, 1981.

[8] Morris C.N. Parametric empirical Bayes inference: Theory and applications. *American Statistical Assosiation*, 78:47–55, 1983.

[9] Lockhart D.J. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.

[10] Robbins H. Proceeding of the Third Berkeley Symposium of Mathematical Statistics. In *An empirical Bayes approach to statistics*, volume 1, pages 152–163, Berkeley, 1956. University of California Press.

[11] Brown L. Admissable estimators, recurrent diffusions, and insolvable boundary value problems. *The Annals of Mathematics and Statistics*, 42:855–903, 1971.

[12] Newton M.A. and Kendziorski C. *The analysis of gene expression data: methods and software*, chapter Parametric Empirical Bayes Methods for Microarrays. Springer Verlag, New York, 2003.

Appendix

Section 2.3

Proof for the Gamma Gamma model:

$$f_0(x_1, x_2, \ldots, x_I) =$$

$$= \int (\prod_{i=1}^{I} (\frac{\lambda^{\alpha} x^{\alpha-1} \exp(-\lambda x)}{\Gamma(\alpha)}) \frac{v^{\alpha_0} \lambda^{\alpha_0 - 1} \exp(-\lambda v)}{\Gamma(\alpha_0)} d\lambda$$

$$= \frac{v^{\alpha_0}}{(\Gamma(\alpha))^I \Gamma(\alpha_0)} \int (\prod_{i=1}^{I} x_i)^{\alpha-1} \lambda^{I\alpha+\alpha_0-1} \exp(-\lambda(\sum_{i=1}^{I} x_i + v)) \, d\lambda$$

$$= \frac{v^{\alpha_0} \Gamma(I\alpha + \alpha_0)(\prod_{i=1}^{I} x_i)^{\alpha-1}}{(\Gamma(\alpha))^I \Gamma(\alpha_0)(\sum_{i=1}^{I} x_i + v)^{I\alpha+\alpha_0}}$$

$$\cdot \int \frac{(\sum_{i=1}^{I} x_i + v)^{I\alpha+\alpha_0} \lambda^{I\alpha+\alpha_0-1} \exp(-\lambda(\sum_{i=1}^{I} x_i + v))}{\Gamma(I\alpha + \alpha + 0)} d\lambda$$

$$= \frac{v^{\alpha_0} \Gamma(I\alpha + \alpha_0)}{(\Gamma(\alpha))^I \Gamma(\alpha_0)} \frac{(\prod_{i=1}^{I} x_i)^{\alpha-1}}{(\sum_{i=1}^{I} x_i + v)^{I\alpha+\alpha_0}}$$

$$= k \frac{(\prod_{i=1}^{I} x_i)^{\alpha-1}}{(\sum_{i=1}^{I} x_i + v)^{I\alpha+\alpha_0}}$$

Section 2.4

Proof for the Log-normal Normal model:

$$f_0(y_1, y_2, \ldots, y_n) = \int \left(\prod_{i=1}^{n} \frac{\exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right)}{\sigma\sqrt{2\pi}} \right) \frac{\exp\left(-\frac{(\mu-\mu_0)^2}{2{\tau_0}^2}\right)}{\tau_0\sqrt{2\pi}} d\mu$$

$$= \frac{1}{(\sigma\sqrt{2\pi})^n \tau_0\sqrt{2\pi}} \int \exp\left\{ -\tfrac{1}{2}\left(\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{{\tau_0}^2}\right)\right\} d\mu$$

Superscript:

$$\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{{\tau_0}^2} = \frac{\sum_{i=1}^{n}(\mu-y_i)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{{\tau_0}^2}$$

$$= \frac{\sum_{i=1}^{n}(\mu-\bar{y}-y_i+\bar{y})^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{{\tau_0}^2}$$

$$= \frac{n(\mu-\bar{y})^2}{\sigma^2} + \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{{\tau_0}^2}$$

$$= \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sigma^2} + \frac{n\mu^2 - 2n\mu\bar{y} + n\bar{y}^2}{\sigma^2} + \frac{\mu^2 - 2\mu\mu_0 + {\mu_0}^2}{{\tau_0}^2}$$

$$= \frac{1}{2}\left(\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sigma^2} + \mu^2\left(\frac{n}{\sigma^2} + \frac{1}{{\tau_0}^2}\right) - 2\mu\left(\frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{{\tau_0}^2}\right) + \left(\frac{n\bar{y}^2}{\sigma^2} + \frac{{\mu_0}^2}{{\tau_0}^2}\right)\right)$$

$$= \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sigma^2} + \left(\frac{n}{\sigma^2} + \frac{1}{{\tau_0}^2}\right)\left(\mu^2 - 2\mu\left(\frac{\frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{{\tau_0}^2}}{\frac{n}{\sigma^2} + \frac{1}{{\tau_0}^2}}\right) + \left(\frac{\frac{n\bar{y}^2}{\sigma^2} + \frac{{\mu_0}^2}{{\tau_0}^2}}{\frac{n}{\sigma^2} + \frac{1}{{\tau_0}^2}}\right)\right)$$

$$\frac{\frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{{\tau_0}^2}}{\frac{n}{\sigma^2} + \frac{1}{{\tau_0}^2}} = \frac{n{\tau_0}^2\bar{y} + \sigma^2\mu_0}{\sigma^2 + n{\tau_0}^2} = \omega_1\bar{y} + \omega_2\mu_0$$

where :

$$\omega_1 = \frac{n{\tau_0}^2}{\sigma^2 + n{\tau_0}^2} \omega_2 = \frac{\sigma^2}{\sigma^2 + n{\tau_0}^2}$$

$$\frac{\frac{n\bar{y}^2}{\sigma^2}+\frac{{\mu_0}^2}{{\tau_0}^2}}{\frac{n}{\sigma^2}+\frac{1}{{\tau_0}^2}}=\frac{n{\tau_0}^2\bar{y}^2+\sigma^2{\mu_0}^2}{\sigma^2+n{\tau_0}^2}=\omega_1\bar{y}^2+\omega_2{\mu_0}^2$$

Then :

$$\frac{\sum_{i=1}^n(y_i-\mu)^2}{\sigma^2}+\frac{(\mu-\mu_0)^2}{{\tau_0}^2}$$
$$=\frac{\sum_{i=1}^n(y_i-\bar{y})^2}{\sigma^2}+(\frac{n}{\sigma^2}+\frac{1}{{\tau_0}^2})(\mu-\omega_1\bar{y}+\omega_2\mu_0)^2$$
$$+\frac{1}{2}(\frac{n}{\sigma^2}+\frac{1}{{\tau_0}^2})(\omega_1\bar{y}^2+\omega_2{\mu_0}^2-(\omega_1\bar{y}+\omega_2\mu_0)^2)$$

So:

$$f_0(y_1,y_2,,y_n)=\int\frac{\exp\left\{(-\frac{1}{2}(\frac{n}{\sigma^2}+\frac{1}{{\tau_0}^2})(\mu-\omega_1\bar{y}+\omega_2\mu_0)^2)\right\}}{\sqrt{(\frac{n}{\sigma^2}+\frac{1}{{\tau_0}^2})2\pi}}d\mu$$
$$\exp\left\{(-\frac{1}{2}(\frac{\sum_{i=1}^n(y_i-\bar{y})^2}{\sigma^2}+(\frac{n}{\sigma^2}+\frac{1}{{\tau_0}^2})(\omega_1\bar{y}^2+\omega_2{\mu_0}^2-(\omega_1\bar{y}+\omega_2\mu_0)^2))\right\}$$
$$\cdot\frac{\sqrt{2\pi(\frac{n}{\sigma^2}+\frac{1}{{\tau_0}^2})}}{(\sigma\sqrt{2\pi})^n\tau_0\sqrt{2\pi}}$$
$$=\frac{1}{(\sigma\sqrt{2\pi})^n\sigma_n\tau_0}\exp\left\{(-\frac{1}{2}(\frac{\sum_{i=1}^n(y_i-\bar{y})^2}{\sigma^2}+\omega_1\omega_2\frac{(\bar{y}-\mu_0)^2}{{\sigma_n}^2})\right\}$$

Section 3.3

Proof of $X'[s - \mu(\hat{\beta})] = 0$:

$$l(s_1, s_2, \ldots, s_{40}|\beta_0, \beta_1, y_{(1)}, y_{(2)}, \ldots, y_{(40)}) = \prod_{j=1}^{J} \frac{\mu_j^{s_j}}{s_j!} e^{\mu_j}$$

$$s_j \sim P(\mu_j(\beta)), \mu_j(\beta) = \mu_j^0 exp\{\beta_0 + \beta_1 t_j\}$$

To make the solution more obvious it is presumed that $\mu_j^0 = 1$. Then

$$\begin{aligned} l(s_1, \ldots, s_{40}|\beta_0, \beta_1, y_{(1)}, \ldots, y_{(40)}) &= \prod_{j=1}^{J} \frac{(e^{\beta_0+\beta_1 t_j})^{s_j}}{s_j!} e^{-(e^{\beta_0+\beta_1 t_j})} \\ &= \frac{e^{\sum_{j=1}^{J} s_j(\beta_0+\beta_1 t_j)}}{\prod_{j=1} J s_j!} e^{-(e^{\beta_0+\beta_1 t_j})} \end{aligned}$$

$$\begin{aligned} \Rightarrow L(s_1, \ldots, s_{40}|\beta_0, \beta_1, y_{(1)}, \ldots, y_{(40)}) &= \sum_{j=1}^{J} s_j(\beta_0 + \beta_1 t_j) \\ &\quad - \sum_{j=1}^{J} e^{\beta_0+\beta_1 t_j} - \log \prod_{j=1}^{J} s_j! \end{aligned}$$

$$\Rightarrow \begin{cases} \dfrac{dL}{d\beta_0} = \sum_{j=1}^{J} s_j - \sum_{j=1}^{J} e^{\beta_0+\beta_1 t_j} = 0 \\ \dfrac{dL}{d\beta_1} = \sum_{j=1}^{J} s_j t_j - \sum_{j=1}^{J} t_j e^{\beta_0+\beta_1 t_j} = 0 \end{cases}$$

$$X'[s - \mu(\hat{\beta}) = \begin{bmatrix} \sum_{j=1}^{J} s_j \\ \sum_{j=1}^{J} s_j t_j \end{bmatrix} - \begin{bmatrix} \sum_{j=1}^{J} e^{\beta_0+\beta_1 t_j} \\ \sum_{j=1}^{J} t_j e^{\beta_0+\beta_1 t_j} \end{bmatrix} = 0$$

Which means that the maximum likelihood equations can be summarized using:

$$X'[s - \mu(\hat{\beta})] = 0$$

Section 4.2

Proof of Brown's [11] and Stein's [7] formula:

$$\mu|z \sim (z + \sigma_0^2 l'(z), \sigma_0^2(1 + \sigma_0^2 l''(z))$$

$$f(z) = \int_{-\infty}^{+\infty} \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(\frac{-(z-\mu)^2}{2\sigma_0^2}\right) g(\mu) d\mu$$

$$f'(z) = \int_{-\infty}^{+\infty} -\frac{(z-\mu)}{\sigma_0^2} \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2\sigma_0^2}\right) g(\mu) d\mu$$

$$\frac{f'(z)}{f(z)} = -\frac{z}{\sigma_0^2} + \frac{\int_{-\infty}^{+\infty} \frac{\mu}{\sigma_0\sqrt{2\pi}} \exp\left(\frac{1}{2\sigma_0^2}(z-\mu)^2\right) g(\mu) d\mu}{f(z)\sigma_0^2}$$

$$\Rightarrow E(\mu|z) = z + \sigma_0^2 l'(z)$$

$$l''(z) = \frac{f''(z)f(z) - f'(z)^2}{f(z)^2} = \frac{f''(z)}{f(z)} - \left(\frac{f'(z)}{f(z)}\right)^2$$

$$\frac{f''(z)}{f(z)} = -\frac{1}{\sigma_0^2} \int_{-\infty}^{+\infty} \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_0^2}(z-\mu)^2\right) g(\mu) d\mu$$

$$+ \int_{-\infty}^{+\infty} \frac{(z-\mu)^2}{\sigma_0^4} \frac{\exp\left(-\frac{1}{2\sigma_0^2}(z-\mu)^2\right)}{\sigma_0\sqrt{2\pi}} g(\mu) d\mu$$

$$\Rightarrow \frac{f''(z)}{f(z)} = -\frac{1}{\sigma_0^2} + \frac{z^2}{\sigma_0^4} - \frac{2z}{\sigma_0^4} E(\mu|z) + \frac{1}{\sigma_0^4} E(\mu^2|z)$$

$$\left(\frac{f'(z)}{f(z)}\right)^2 = \frac{z^2}{\sigma_0^4} - \frac{2z}{\sigma_0^4} E(\mu|z) + \frac{E(\mu^2|z)^2}{\sigma_0^4}$$

$$V(\mu|z) = \sigma_0^2(1 + \sigma_0^2 l''(z))$$

Section 4.3

1. The marginal mean and variance for $\mu \sim (\alpha, \beta^2)$, $z|\mu \sim (\mu, \sigma_0^2)$
   $z \sim (\alpha, \beta^2 + \sigma_0^2)$.

The mean and variance of z_i^* given μ_i is $(\sqrt{l}\mu_i, \sigma_0^2)$.

$$z_i^* = \frac{\sum_{j=1}^{l} z_{ij}}{\sqrt{l}}$$

$$E(z_i^*|\mu_i) = \frac{\sum_{j=1}^{l} E(z_{ij})}{\sqrt{l}} = \sqrt{l}\mu_i$$

$$V(z_i^*|\mu_i) = \frac{\sum_{j=1}^{l} V(z_{ij})}{l} = \sigma_0^2$$

2. The marginal mean and variance of z^* can be estimated using Efron's formula in (4.8).

$$E(z^*) = \sqrt{l}A$$

$$\begin{aligned} V(z^*) = d^2V(z) &= d^2B^2 \\ &= lB^2 - (l-1)\sigma_0^2 \\ &= l\beta^2 + l\sigma_0^2 + \sigma_0^2 \\ &= l\beta^2 + \sigma_0^2 \end{aligned}$$

Section 4.4

(a) Assuming that $z|\mu$ is $N(\mu, \sigma^2)$ and that for under-expressed genes μ is $N(\mu_1, \sigma_1^2)$, for null genes μ is $N(\mu_2, \sigma_2^2)$ and for over-expressed genes μ is $N(\mu_3, \sigma_3^2)$, the distribution of under-expressed genes is $N(\mu_1, \sigma_1^2 + \sigma^2)$, for null genes is $N(\mu_2, \sigma_2^2 + \sigma^2)$ and for over-expressed genes the distribution function is $N(\mu_3, \sigma_3^2 + \sigma^2)$.

(1) $z|\mu$ is $N(\mu, \sigma^2)$ and for under-expressed genes μ is $N(\mu_1, \sigma_1^2)$. The distribution of under-expressed genes is $N(\mu_1, \sigma_1^2 + \sigma^2)$:

$E(z) = E(E(z|\mu)) = E(\mu) = \mu_1$
$V(z) = E(V(z|\mu)) + V(E(z|\mu)) = \sigma^2 + \sigma_1^2$
The fact that Normality is preserved is a standard argument in statistics.

(2) $z|\mu$ is $N(\mu, \sigma^2)$ and for null genes μ is $N(\mu_2, \sigma_2^2)$. The distribution of null genes is $N(\mu_2, \sigma_2^2 + \sigma^2)$:

$E(z) = E(E(z|\mu)) = E(\mu) = \mu_2$
$V(z) = E(V(z|\mu)) + V(E(z|\mu)) = \sigma^2 + \sigma_2^2$
The fact that Normality is preserved is a standard argument in statistics.

(3) $z|\mu$ is $N(\mu, \sigma^2)$ and for over-expressed genes μ is $N(\mu_3, \sigma_3^2)$. The distribution of over-expressed genes is $N(\mu_3, \sigma_3^2 + \sigma^2)$:

$E(z) = E(E(z|\mu)) = E(\mu) = \mu_3$
$V(z) = E(V(z|\mu)) + V(E(z|\mu)) = \sigma^2 + \sigma_3^2$
The fact that Normality is preserved is a standard argument in statistics.

(b) Differentiating with respect to μ_1, μ_2, μ_3, σ_1^2, σ_2^2 and σ_3^2 separately, we estimate:

$$\hat{\mu}_1 = \frac{\sum_{j=1}^{J} \Phi_{1j} z_j}{\sum_{j=1}^{J} \Phi_{1j}} \qquad \hat{\mu}_2 = \frac{\sum_{j=1}^{J} \Phi_{2j} z_j}{\sum_{j=1}^{J} \Phi_{2j}} \qquad \hat{\mu}_3 = \frac{\sum_{j=1}^{J} \Phi_{3j} z_j}{\sum_{j=1}^{J} \Phi_{3j}}$$

and $\sigma_1'^2$, $\sigma_2'^2$ and $\sigma_3'^2$:

$$\hat{\sigma}_1'^2 = \frac{\sum_{j=1}^{J} \Phi_{1j}(z_j - \mu_1)^2}{\sum_{j=1}^{J} \Phi_{1j}} \quad \hat{\sigma}_2'^2 = \frac{\sum_{j=1}^{J} \Phi_{2j}(z_j - \mu_2)^2}{\sum_{j=1}^{J} \Phi_{2j}} \quad \hat{\sigma}_3'^2 = \frac{\sum_{j=1}^{J} \Phi_{3j}(z_j - \mu_3)^2}{\sum_{j=1}^{J} \Phi_{3j}}$$

$$\frac{\mathrm{d} \log P(z, \Phi|\theta, p)}{\mathrm{d}\mu_1} = \frac{\sum_{j=1}^{J} \Phi_{1j}(z_j - \mu_1)}{\sigma_1^{2\prime}} = 0 \implies \hat{\mu}_1 = \frac{\sum_{j=1}^{J} \Phi_{1j} z_j}{\sum_{j=1}^{J} \Phi_{1j}}$$

$$\frac{\mathrm{d} \log P(z, \Phi|\theta, p)}{\mathrm{d}\mu_2} = \frac{\sum_{j=1}^{J} \Phi_{2j}(z_j - \mu_2)}{\sigma_2^{2\prime}} = 0 \implies \hat{\mu}_2 = \frac{\sum_{j=1}^{J} \Phi_{2j} z_j}{\sum_{j=1}^{J} \Phi_{2j}}$$

$$\frac{\mathrm{d} \log P(z, \Phi|\theta, p)}{\mathrm{d}\mu_3} = \frac{\sum_{j=1}^{J} \Phi_{3j}(z_j - \mu_3)}{\sigma_3^{2\prime}} = 0 \implies \hat{\mu}_3 = \frac{\sum_{j=1}^{J} \Phi_{3j} z_j}{\sum_{j=1}^{J} \Phi_{3j}}$$

$$\begin{aligned} \frac{\mathrm{d} \log P(z, \Phi|\theta, p)}{\mathrm{d}\sigma_1^2} &= \sum_{j=1}^{J} \Phi_{1j}(z_j - \mu_1)^2 - \sum_{j=1}^{J} \Phi_{1j}\sigma_1^{2\prime} = 0 \\ \implies \hat{\sigma}_1^{2\prime} &= \frac{\sum_{j=1}^{J} \Phi_{1j}(z_j - \mu_1)^2}{\sum_{j=1}^{J} \Phi_{1j}} \end{aligned}$$

$$\begin{aligned} \frac{\mathrm{d} \log P(z, \Phi|\theta, p)}{\mathrm{d}\sigma_2^2} &= \sum_{j=1}^{J} \Phi_{2j}(z_j - \mu_2)^2 - \sum_{j=1}^{J} \Phi_{2j}\sigma_2^{2\prime} = 0 \\ \implies \hat{\sigma}_2^{2\prime} &= \frac{\sum_{j=1}^{J} \Phi_{2j}(z_j - \mu_2)^2}{\sum_{j=1}^{J} \Phi_{2j}} \end{aligned}$$

$$\begin{aligned} \frac{\mathrm{d} \log P(z, \Phi|\theta, p)}{\mathrm{d}\sigma_3^2} &= \sum_{j=1}^{J} \Phi_{3j}(z_j - \mu_3)^2 - \sum_{j=1}^{J} \Phi_{3j}\sigma_3^{2\prime} = 0 \\ \implies \hat{\sigma}_3^{2\prime} &= \frac{\sum_{j=1}^{J} \Phi_{3j}(z_j - \mu_3)^2}{\sum_{j=1}^{J} \Phi_{3j}} \end{aligned}$$

(c) $z|\mu$ is $N(\mu, \sigma^2)$ and for null genes μ is $N(\mu_2, 0)$. The distribution of null genes is $N(\mu_2, \sigma^2)$:

$E(z) = E(E(z|\mu)) = E(\mu) = \mu_2$
$V(z) = E(V(z|\mu)) + V(E(z|\mu)) = \sigma^2 + 0$
$\implies V(z) = \sigma^2$