Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN
Metodologia Statistica per la Ricerca Scientifica

XXVII Ciclo

Settore Concorsuale di afferenza: 13/D1

Settore Scientifico disciplinare: SECS-S/01

Differential expression analysis for sequence count data via mixtures of negative binomials

Presentata da
Elisabetta Bonafede

Coordinatore Dottorato
Prof.ssa Angela Montanari

Relatori
Prof.ssa Cinzia Viroli
Prof. Stéphane Robin

Esame finale anno 2015

Abstract

The recent advent of Next-generation sequencing technologies has revolutionized the way of analyzing the genome. This innovation allows to get deeper information at a lower cost and in less time, and provides data that are discrete measurements. One of the most important applications with these data is the differential analysis, that is investigating if one gene exhibit a different expression level in correspondence of two (or more) biological conditions (such as disease states, treatments received and so on). As for the statistical analysis, the final aim will be statistical testing and for modeling these data the Negative Binomial distribution is considered the most adequate one especially because it allows for "over dispersion". However, the estimation of the dispersion parameter is a very delicate issue because few information are usually available for estimating it. Many strategies have been proposed, but they often result in procedures based on plug-in estimates, and in this thesis we show that this discrepancy between the estimation and the testing framework can lead to uncontrolled first-type errors. We propose a mixture model that allows each gene to share information with other genes that exhibit similar variability. Afterwards, three consistent statistical tests are developed for differential expression analysis. We show that the proposed method improves the sensitivity of detecting differentially expressed genes with respect to the common procedures, since it is the best one in reaching the nominal value for the first-type error, while keeping elevate power. The method is finally illustrated on prostate cancer RNA-seq data.

Acknowledgments

It is very hard to express my whole grateful in few lines: these three years have been very precious and enriching.
First of all, thanks to my supervisor Prof. Cinzia Viroli: thank you for all the patient and scrupulous support, and thanks for giving me not only significant statistical teachings but also a very precious human example.
A great thanks is also for Prof. Stéphane Robin: thank you for all the interesting suggestions and for your helpfulness; working with you has been really stimulating, and the parisian months have been a very nice experience that I will never forget.
Thanks to Prof. Angela Montanari, who has always been present in my academic life, and who has given to me the opportunity of helping her with teaching.
Thanks to my thesis committee Prof. Marco Alfò, Prof. Livio Finos and Prof. Luigi Ippoliti, who gave me interesting suggestions to improve this work.
Thanks also to Dr. Franck Picard, for his contribution at this work.
Thanks to my Ph.D. colleagues Linda, Giovanni, Sara, Elena and Arianna: it has been nice to share with you these intense years! Thanks also to the guys of the AgroParisTech, and especially to Eleanna, for all the moments that we have enjoyed together.
Finally, last but not least, a special thanks to my family, to Giacomo and to all my friends: if this experience has been so special, it is certainly also thanks to you.

Contents

Chapter 1

Introduction

In the last decade the next-generation sequencing (NGS) assays, such as RNA-seq or ChIP-seq, have been revolutionizing the depth of understanding of the genome structure and the DNA or RNA interaction regions, due to the higher resolution of the data provided by these technologies [Soon et al., 2013, Wang et al., 2009]. From a statistical point of view, this innovation has gone along with a change in the nature of the data. Indeed, whilst the past mostly-used microarray technologies measured the abundance of a particular transcript as a fluorescence intensity expressed as continuous real data, the NGS experiments give *read counts* assigned to target genome regions, measuring the expression level or the abundance of the target transcript.

When the purpose of the assay is to perform *differential analysis*, that is comparing the counts of a given region between conditions, the statistical task is then to provide an appropriate model to account for biological and technical variations, as well as a testing framework to test the hypothesis of no difference. Here we deal with the case where regions of interest are given *a priori*, contrarily to analysis where the regions themselves have to be discovered [Frazee et al., 2014].

Generalized linear models based on count distributions now constitute a consensus framework for the analysis, with the original Poisson distribution [Marioni et al., 2008, Wang et al., 2010] being replaced by the Negative Binomial model [Robinson and Smyth, 2008, Anders and Huber, 2010, Robinson et al., 2010]. Indeed, the simplest choice of the Poisson distribution was rapidly identified as the cause of uncontrolled first-type errors, due to a poor adjustment to the larger observed variability compared with the equal mean-variance specification of

the Poisson model (see, for a discussion, Anders and Huber [2010]). Since then, the correct modeling and estimation of this observed overdispersion has been a key issue in differential analysis.

Taking perspective from our past experience in micro-array analysis, the proper modeling of the dispersion parameter has long been a subject of debate in differential analysis, with a difficult trade-off between a common variance for every genes and gene-specific variances. Given the limited number of replicates, the first strategy provides robust estimates, but the testing procedure lacks of power and the model is not realistic, whereas the second is more sensitive at the price of increased first-type errors. Actually, the debate is still ongoing with the Negative Binomial framework, but the problem is much more difficult to solve due to this complex (and unknown) mean-variance relationship.

Several contributions have been proposed to find a trade-off between the common overdispersion and the gene-specific overdispersion frameworks, and we will describe the three mostly used strategies in the next Chapter.

Despite the rapidly increasing diffusion of these statistical procedures, also thanks to the availability of several well documented Bioconductor packages, the estimation of the dispersion in NGS data remains a crucial and tricky issue because of the limited number of available observations for each gene. However, less attention has been focused on the consistency between the estimation and the testing frameworks. Indeed, many strategies consider the use of plug-in estimators, but an important drawback of this choice is that the expected variations of the test statistics are no longer controlled under the null hypothesis, which may result in an un-controlled level of the test. We will illustrate this point by a simulation study, showing that most proposed methods do not reach the nominal level of the test, whereas it is precisely what is expected to be controlled when performing standard hypothesis testing.

Our contribution is to explore and discuss a mixture model approach [McLachlan and Peel, 2000, Fraley and Raftery, 2002] based on the idea of sharing information among genes that exhibit similar dispersion. More specifically, mixtures of negative binomial distributions are investigated as a way to get more accurate estimations for the dispersion parameter of each gene, exploiting also the information provided by the others. Such an approach has already been considered in the same context for the differential analysis of microarray data [Delmar et al., 2005]. A consistent statistical testing procedure is then developed within the unified model based clustering framework. The proposed method improves the sen-

sitivity of detecting differentially expressed genes with small replicates, and we show that our method controls the nominal level of the test by simulation.

A description of the procedures that provide these data together with a review of the most used methods that have been proposed in last years for performing differential analysis will be presented in Chapter 2. In Chapter 3 the negative binomial distribution and mixtures of negative binomials will be introduced. The novel method, together with the derivation of three statistical tests for performing the differential analysis will be described in Chapter 4. We will show through a large simulation study in Chapter 5 that the proposed statistical test procedure outperforms the mostly used strategies in the literature, because it is the best one in reaching the nominal value for the first-type error, while keeping elevate power, thus indicating its inferential reliability. The method will be applied on prostate cancer data in Chapter 6. A final discussion is presented in Chapter 7.

Chapter 2

NGS technologies and differential analysis

2.1 NGS technologies and RNA-Seq data

The advent of the Next-Generation Sequencing (NGS) technologies has led to the production of sequencing platforms that allow to obtain high-throughput genomic data. These technologies can be used for many kinds of experiment, and in this work we will focus on those that lead to the analysis and the quantification of the transcriptome (that is the set of all RNA molecules), namely RNA-Seq data. The procedure for getting this kind of data can be summarized in three main steps [Oshlack et al., 2010]:

1. **Sequencing**: the studied transcriptome has to be preliminarily split into millions of fragments. The *sequencing* process produces the *short reads*, that represent the sequence of the nucleotide basis that compose each fragment.

   Many types of *NGS sequencing platforms* have been produced for this fundamental procedure (among the others, *454 Genome Sequencer* by Roche, *Genome Analyzer* by Illumina and *SOLiD* by Applied Biosystems). Each machine has its specific characteristics. We will not go into technical details, but as regards the mostly used one, the Illumina *Genome Analyzer*, we can say that the samples are put on the *flow-cell* that is a sequencing plate composed by several (usually 8) *lanes* (independent regions on the support). This makes possible to sequence different samples at once.

2. **Read mapping**, or **alignment**: mapping the *reads* means to find a unique

location where a short read is identical (or, in practice, almost identical) to a reference genome or transcriptome.

3. **Summarizing**: after that as more reads as possible have been mapped on the genome, the data can be summarized simply by counting the number of *reads* overlapping the exons (the codifying regions) in a gene.

As such, they are discrete measurements. It is important to underline that different lanes could be characterized by different *sequencing depth* (or *library size*), and potentially also by other technical effects. This makes *normalization* procedures mandatory for comparative purpose (see, for instance, Bullard et al. [2010], Tarazona et al. [2011], Risso et al. [2011] and Dillies et al. [2013]), but in this thesis we will not go into details of normalization methods.
The statistical procedures used for the analysis have to account for the features of count data, for the limited number of available information for each gene (due to high costs in sequencing procedures) and for the fact that NGS data are very often characterized by excess of zeros (inactivated regions).

2.2 Differential Analysis

Preliminaries: notation

RNA-seq data consist of nonnegative counts indicating the number of reads observed for each gene. Suppose we analyze p genes in d different conditions, and on each of them observations are taken over n_j replicates.
We denote Y_{ijr} the random variable that expresses the counts of reads mapped to gene i (i=1, ..., p), in condition j (j=1, ..., d; in this work, d=2 w.l.g.), in sample r (r= $1, ..., n_j$).

Differential analysis

One of the most important and largely studied applications for NGS data is *differential analysis*, that is comparing the expression level of a specific gene (or exon) between samples observed in correspondence of two different biological conditions such as tissue types, disease states or treatments. Identifying differentially expressed genes could be a first step in detecting possible connections between one of these situations and the expression level of a specific gene [Kvam et al., 2012].

From the statistical viewpoint, the differential analysis implies to perform statistical testing to decide whether, for a given gene, an observed difference in read counts between two biological conditions is significant under a specific *discrete* probabilistic distribution or if it is just due to natural random variability [Anders and Huber, 2010]. The final aim consists in testing the null-hypothesis:

$$H_0 : \text{expression levels}_{condition\ 1} = \text{expression levels}_{condition\ 2}.$$

The benchmark distribution for count data is the Poisson [Cameron and Trivedi, 1998], but it can be too restrictive because it implies that variance and mean are equal ("equi-dispersion" property), while the RNA-seq data could be characterized by higher dispersion, leading to the so-called "over-dispersion problem", and therefore the resulting statistical test would be unreliable.

The quasi-Poisson model had been proposed as alternative [Ismail and Zamani, 2013], which is fitted on the basis of a quasi-likelihood function, specifying a relationship between the mean and the variance. However the provided estimators do not have great properties.
A more appropriate distribution is the negative binomial (NB), that is characterized by two parameters (a mean and a dispersion one) where the variance is a function of both oh them. For a deeper description of the NB distribution we refer to Section 3.1. A severe issue associated with this probabilistic framework is the reliable estimation of the *dispersion* parameter, reinforced by the limited number of replicates generally observable for each gene.

2.3 State of the art

Differential expression analysis is a largely studied application of RNA-Seq data, and many works have been published about it; here we report a brief overview of the mostly known ones. Bloom et al. [2009] have proposed to use the Fisher's exact test for comparing the proportion of reads mapped to each gene in correspondence of different conditions.
Several strategies arose considering the Poisson distribution as reference for modeling the data. Among the others, Marioni et al. [2008] have fitted a Poisson model, thus performing a χ^2 goodness of fit test; Bullard et al. [2010] have explored the likelihood ratio test possibility. Wang et al. [2010] have assumed the normality

distribution for the log ratios of the counts and thus have computed a z-score, providing an R package that is called *DEGseq* (available on Bioconductor); Li et al. [2011] proposed a score statistic on the basis of a Poisson log-linear model, providing an R package that is called *PoissonSeq* (available on CRAN). Various approaches based on the negative binomial distribution have been studied, for handling the overdispersion problem. Hardcastle and Kelly [2010] proposed to iteratively estimate the dispersion using the quasi-likelihood approach, thus providing a ranking of the genes on the basis of the posterior probabilities of being DE instead of the classical p-values; they published the R package *baySeq*. Zhou and Wright [2011], with their R package *BBSeq*, modeled the dispersions on the means, thus computing a Wald test. Other three strategies, probably the mostly used at all, are described in next sections with much more detail. These strategies have lead to largely used R packages, available and well documented from Bioconductor: *DESeq*, *edgeR* and *DSS*. We will present each method assuming that we are comparing just $d = 2$ different biological conditions.

2.3.1 edgeR

Robinson and Smyth (Robinson and Smyth [2007], Robinson and Smyth [2008]) proposed to estimate a common dispersion parameter for all genes expressed as a quadratic combination of the mean, and then, by making use of a weighted likelihood procedure, they provide an estimation of each dispersion parameter as a weighted combination of the common and of the individual ones, assuming empirical weights. Then an approximation is introduced in order to develop an exact test. This procedure is available in the R package *edgeR* [Robinson et al., 2010].

The model

Let us assume a NB distribution for the random variable Y_{ijr} that describes the counts for gene i in the $r-th$ sample of condition j: $Y_{ijr} \sim NB(\mu_{ijr}, \phi)$ where ϕ is the dispersion parameter such that $E(Y_{ijr}) = \mu_{ijr}$ and $Var(Y_{ijr}) = \mu_{ijr}(1 + \mu_{ijr}\phi)$. Let us suppose that $E(Y_{ijr}) = \mu_{ijr} = s_{jr}\lambda_{ij}$ such that λ_{ij} describes the real abundance of transcripts for gene i in condition j and s_{jr} is the *size factor*. Performing differential expression analysis means to test the null hypothesis

$$H_0 : \lambda_{i1} = \lambda_{i2}, \text{for } i = 1, \ldots, p.$$

Robinson and Smyth [2008] have proposed a new way for estimating the dispersion parameter with this kind of data, that consists in making use of the information provided from all genes in order to estimate a common dispersion ϕ, maximizing the common likelihood function $l_C(\phi)$.

Afterwards, they applied a quantile adjustment for avoiding problems due to differences in sample sizes.
The assumption of common dispersion is a good way for gain more stable results but it is not realistic for this kind of data, and therefore an Empirical Bayesian strategy has been proposed. Such procedure had already been applied to microarray data [Smyth, 2004]. The idea is to use a weighted conditional log-likelihood for estimating each gene-wise dispersion ϕ_i (WL(ϕ_i)):

$$WL(\phi_i) = l_i(\phi_i) + \alpha l_C(\phi_i) \tag{2.1}$$

where we can recognize a special case of weighted likelihood [Wang, 2006], where the common likelihood rules as a prior for ϕ_i and α plays the rule of the prior precision. The value for α has to be chosen accordingly to the strength of the similarity between the different dispersion parameters: the greater is α, the stronger is the effect of the common component.

As regards the selection of an appropriate value for α, first of all they have introduced their strategy considering a hierarchical model assuming (ideally) that the gene-specific estimators $\widehat{\phi}_i$ were normally distributed: $\widehat{\phi}_i|\phi_i \sim N(\phi_i, \tau_i^2)$ and $\phi_i \sim N(\phi_0, \tau_0^2)$. The Bayes posterior mean estimator of ϕ_i would be:

$$\widehat{\phi}_i^B = E(\phi_i|\widehat{\phi}_i) = \frac{\widehat{\phi}_i/\tau_i^2 + \phi_0/\tau_0^2}{1/\tau_i^2 + 1/\tau_0^2}. \tag{2.2}$$

ϕ_0 and τ_0 can be estimated from the marginal distribution of $\widehat{\phi}_i$ to get a EB rule. Under this idealistic model, we could derive:

$$\widehat{\phi}_i^{WL} = \frac{\widehat{\phi}_i/\tau_i^2 + \alpha \sum_{j=1}^{d} \phi_j/\tau_j^2}{1/\tau_i^2 + \alpha \sum_{j=1}^{d} 1/\tau_j^2}, \tag{2.3}$$

that coincides with the (2.2) for $\phi_0 = \widehat{\phi}_0 = \frac{\sum_{i=1}^{p} \widehat{\phi}_i/\tau_i^2}{\sum_{i=1}^{p} 1/\tau_i^2}$ and $1/\alpha = \sum_{i=1}^{p} \tau_0^2/\tau_i^2$. As for the estimation of τ_0^2, under the normal model we would have that $(\widehat{\phi}_i - \phi_0)^2/(\tau_i^2 + \tau_0^2) \sim \chi_1^2$, so that a consistent estimator for τ_0^2 could be

computed by solving:

$$\sum_{i=1}^{p}\left(\frac{(\widehat{\phi}_i-\widehat{\phi}_0)^2}{\tau_i^2+\tau_0^2}-1\right)=0. \tag{2.4}$$

In practice the individual estimators $\widehat{\phi}_i$ are not normally distributed and we do not know their variances, but since the score statistics converge to normality more rapidly than maximum likelihood (ML) estimators and (2.4) can be written in terms of the score likelihood function and the expected information, they have proposed the following algorithm: first of all they have estimated the common dispersion $\widehat{\phi}_0$, maximizing l_C; then they evaluated, for each gene, the score function $S_i(\widehat{\phi}_0) = \partial l_i(\widehat{\phi}_0)/\partial(\widehat{\phi}_0)$ and the expected information $I_i(\widehat{\phi}_0) = E(-\partial^2 l_i(\widehat{\phi}_0)/\partial\widehat{\phi}_0^2)$. Afterwards they estimated τ_0 by solving $\sum_{i=1}^{p}\left(\frac{S_i^2}{I_i(1+I_i\tau_0^2)}-1\right)=0$ and they set $1/\alpha=\tau_0^2\sum_{i=1}^{p}I_i$ and finally they got weighted likelihood estimators $\widetilde{\phi}_i$ by maximizing $WL(\phi_i)$.
This way, if $\phi_i=\phi_0$ for $i=1,\ldots,p$ then $E(S_i^2)=I_i$ so that τ_0 will be estimated close to 0 and α will be large. Conversely, if the gene-specific dispersion parameters are dissimilar, the algorithm will account for that proposing greater values for τ_0 (and therefore weakening the shrinkage effect).

Testing for differential expression

After computing the dispersion parameters through the Empirical Bayes estimator, Robinson and Smyth [2007] have proposed an exact test analogous to the Fisher's one. It is adapted for this kind of data, replacing the hypergeometric probabilities with negative binomial ones and conditioning on the sum of all the reads that are mapped to gene i. For doing so, they had needed to introduce an approximation considering the normalized data as identically distributed. Finally they computed the exact p-values as the probabilities of observing counts as or more extreme than the observed.

2.3.2 DESeq

Anders and Huber [2010] proposed to use a mean-dependent local regression to smooth the gene-specific dispersion estimates, related to the idea that genes that share a similar mean expression level have also a similar variance, and therefore they can contribute to the estimation of the respective parameters. The method is

implemented in the *DESeq* R package, available from Bioconductor.

The model

We denote Y_{ijr} the number of reads that have been mapped to gene i in condition j, for sample r. *DESeq* assumes $Y_{ijr} \sim NB(\mu_{ijr}, \sigma^2_{ijr})$ where μ_{ijr} is the mean and σ^2_{ijr} is the variance. The mean μ_{ij} is supposed to be equal to $\lambda_{ij} s_{jr}$, that is the product of two terms: the first one is proportional to the real abundance of transcripts for gene i in condition j, and the latter is the *size factor*, that is lane-dependent. The variance σ^2_{ijr} is the sum of two components: $\sigma^2_{ijr} = \mu_{ijr} + s^2_{jr}\nu_{ij}$. The first one is called *shot noise*, and it is the variance that would be computed assuming a Poisson model for the data. The second is the *raw variance term* and it is supposed to be a smooth function of λ_{ij}. The novel aspect of this strategy consists indeed in considering that the estimation of the *overdispersion* component can be gained pooling information among genes that exhibit a similar expression level.
This model requires the estimation of three sets of parameters:

- the size factors s_{jr} (for each sample r in condition j),
- the expression strength parameters λ_{ij} (for each gene i in condition j),
- the smooth functions $\nu_j : \Re^+ \to \Re^+$ (for each condition j), for modeling the dependence of the raw variances ν_{ij} on the expectations λ_{ij} .

As regards the first set of parameters, Anders and Huber [2010] have proposed a new way for computing the library sizes s_{jr}, that is:

$$\widehat{s}_{jr} = median_i \frac{y_{ijr}}{(\prod_j \prod_r y_{ijr})^{1/n}},$$

where n is the total number of samples ($n = \sum_{j=1}^{2} n_j$), and the denominator is a geometric mean of the counts computed on the n information available for gene i. For the estimation of the expression level,

$$\widehat{\lambda}_{ij} = \frac{1}{n_j} \sum_{r=1}^{n_j} \frac{y_{ijr}}{\widehat{s}_{jr}}. \tag{2.5}$$

As regards the ν_j, first of all they have computed the sample variances for the normalized counts:

$$\omega_{ij} = \frac{1}{n_j - 1} \sum_{r=1}^{n_j} \left(\frac{y_{ijr}}{\widehat{s}_{jr}} - \widehat{\lambda}_{ijr} \right)^2, \tag{2.6}$$

and they have defined:

$$z_{ij} = \frac{\widehat{\lambda}_{ij}}{n_j} \sum_{r=1}^{n_j} \frac{1}{\widehat{s}_{jr}}. \tag{2.7}$$

It can be proved that $\nu_{ij} = \omega_{ij} - z_{ij}$ is an unbiased estimator for the raw variance. Nevertheless, the limitedness of the number of observations that are usually collected make the ω_{ij} very variable, leading to unreliable estimates. Therefore the authors have suggested to fit a local regression model on $(\widehat{\lambda}_{ij}, \omega_{ij})$ to get a smooth function $\nu_j(\lambda)$ with $\widehat{\nu}_j(\widehat{\lambda}_{ij}) = \omega_j(\widehat{\lambda}_{ij}) - z_{ij}$ as estimation of the raw variance.

Testing for differential expression

The differential expression analysis consists in testing the null hypothesis $\lambda_{i1} = \lambda_{i2}$, for $i = 1, \ldots, p$. Anders and Huber [2010] have derived a test statistic that is the total counts in each condition: $Y_{i1} = \sum_{r=1}^{n_1} Y_{i1r}$ and $Y_{i2} = \sum_{r=1}^{n_2} Y_{i2r}$; we define also the overall sum $Y_{i+} = Y_{i1} + Y_{i2}$. For each pair of values (a, b), with $a + b = y_{i+}$, they needed to compute the probability of the events: $Y_{i1} = a$ and $Y_{i2} = b$ and we denote it as $p(a, b)$. The p-value of a pair of observed count sums (y_{i1}, y_{i2}) has been defined as:

$$p_i = \frac{\sum_{\substack{a+b=y_{i+} \\ p(a,b) \leq p(y_{i1}, y_{i2})}} p(a, b)}{\sum_{a+b=y_{i+}} p(a, b)}. \tag{2.8}$$

They assume that, under the null hypothesis, the samples are independents: $p(a, b) = p(Y_{i1} = a)p(Y_{i2} = b)$.
Y_{i1} and Y_{i2} are sums of NB random variables, and they suggested to approximate their distribution by a NB. For the derivation of the parameters, first of all they computed $\widehat{\lambda}_{i0} = \sum_{j=1}^{d} \sum_{r=1}^{n_j} \frac{y_{ijr}}{s_{jr}}$ (pooling the counts of all conditions), that is an average on all the normalized information for gene i, considering the hypothesis $\lambda_{i1} = \lambda_{i2}$ as true. The final mean and variance parameters of the resulting NB distribution are: $\widehat{\mu}_{ij} = \sum_{r=1}^{n_j} s_{jr} \widehat{\lambda}_{i0}$ and $\widehat{\sigma}^2_{ij} = \sum_{r=1}^{n_j} \widehat{s}_{jr} \widehat{\lambda}_{i0} + \widehat{s}^2_{jr} \widehat{\nu}_j(\widehat{\lambda}_{i0})$ for $j = 1, 2$.

2.3.3 DSS

Wu et al. [2013] introduced an empirical Bayes shrinkage approach choosing a log-normal prior distribution on the dispersion parameters and therefore imposing a negative binomial likelihood. Then the estimations are plugged-in the Wald statistic to perform the statistical test. The method is implemented in the *DSS* `R` library.

The model

Wu et al. [2013] have assumed a hierarchical model:

$$\begin{gathered}\phi_i \sim log - Normal(m_0, \tau^2) \\ \downarrow \\ \theta_{ij}|\phi_i \sim Gamma(\lambda_{ij}, \phi_i) \\ \downarrow \\ Y_{ijr}|\theta_{ij} \sim Poisson(\theta_{ij} s_{jr})\end{gathered}$$

The marginal distribution of Y_{ijr} given λ_{ij} and ϕ_i is a NB with mean $\mu_{ijr} = \lambda_{ij} s_{jr}$ (where λ_{ij} describes the real abundance of transcripts for gene i in condition j, s_{jr} is the library size), and dispersion ϕ_i, that is variance equal to $\lambda_{ij} + \phi_i \lambda_{ij}^2$. As it is well known there is not a conjugate prior for ϕ_i, and they imposed a prior that seems to reflect the empirical behavior of the dispersion parameters, that is the log-normal distribution.
It is possible to derive a conditional posterior distribution of ϕ_i given all observed counts and means:

$$\begin{aligned}log\big(p(\phi_i|Y_{ijr}, \mu_{ijr}, j = 1, 2; r = 1, \ldots, n_j)\big) \propto \\ \sum_{j,r} \psi(\phi_i^{-1} + Y_{ijr}) - n\psi(\phi_i^{-1}) - \phi_i^{-1} \sum_{j,r} log(1 + \mu_{ijr}\phi_i) \\ + \sum_i Y_{ijr}\big(log(\mu_{ijr}\phi_i) - log(1 + \mu_{ijr}\phi_i)\big) \\ - \frac{(log(\phi_i) - m_0)^2}{2\tau^2} - log(\phi_i) - log(\tau) \quad (2.9)\end{aligned}$$

where $n = \sum_j n_j$ is the number of samples. We could consider the posterior mean as a good estimate for $\widehat{\phi_i}$, but it would be too computationally intensive. Therefore

they propose to compute the posterior mode by maximizing an approximation of (2.9). In practice, they substitute μ_{ijr} by $\widehat{\mu}_{ijr} = \widehat{\lambda}_{ij} s_{jr}$ where $\widehat{\lambda}_{ij} = \frac{\sum_r Y_{ijr}/s_{jr}}{n_j}$, and they plug-in the two hyper-parameters m_0 and τ^2 by pooling the data provided by all genes. Finally they maximized the approximated equation (2.9) using the Newton-Raphson method. The estimated $\widehat{\phi}$ is an empirical Bayes estimator, shrunken toward the common prior.

Testing for differential expression

Performing differential analysis means testing the null hypothesis $\lambda_{i1} = \lambda_{i2}$, for $i = 1, \ldots, p$.
For this aim, they simply proposed to use the estimated parameters to plug-in the Wald test:

$$t_i = \frac{\widehat{\lambda}_{i1} - \widehat{\lambda}_{i2}}{\widehat{\sigma}^2_{i1} + \widehat{\sigma}^2_{i2}} \tag{2.10}$$

where $\widehat{\sigma}^2_{ij}$ is the estimated variance for λ_{ij}: $\widehat{\sigma}^2_{ij} = 1/n_j^2\big(\widehat{\lambda}_{ij}(\sum_{r=1}^{n_j} 1/s_{jr}) + n_j \widehat{\lambda}^2_{ij}\phi_i\big)$.

Chapter 3

Mixtures of Negative Binomials

As it has already been explained, the Negative Binomial (NB) distribution is an appropriate choice for fitting RNA-Seq data. The analysis of these data requires to work with few replicates for each gene in each condition. Statistical strategies involving mixture models could provide additional flexibility and could be useful for sharing information among genes that exhibit similar features, thus leading to more reliable estimations.
In this chapter we are going to describe the NB and mixtures of NB distributions, and in the next chapter we will illustrate the usefulness of this strategy in performing differential analysis.

3.1 The Negative Binomial distribution

As reported in Hilbe 2011 (Hilbe [2011]), the *Negative Binomial* probability density function can be defined according to several parameterizations. Among these, the so-called Negative Binomial 2 (NB2) is convenient because it is characterized by two parameters that are particularly related to its moments: the first one corresponds to the expectation and the second one has a particular role to determine the variance.
Let us define a random variable $y \sim NegBin(\lambda, \alpha)$ with λ, $\alpha > 0$ and

$$E(y) = \lambda,$$

$$Var(y) = \lambda \left(1 + \frac{1}{\alpha}\lambda\right);$$

Y has density:

$$f(y|\lambda,\alpha) = \binom{y+\alpha-1}{\alpha-1}\left(\frac{\lambda}{\lambda+\alpha}\right)^{y}\left(\frac{\alpha}{\lambda+\alpha}\right)^{\alpha}. \tag{3.1}$$

It is possible to show that (3.1) reduces to the Poisson for $\alpha \to \infty$.

An interesting characteristic of this parametrization is that it can be derived from a Poisson-Gamma mixture model, by considering a random variable having a Poisson distribution with parameter depending on a second random variable distributed according to a Gamma distribution with mean equal to 1.

3.1.1 Poisson-Gamma mixture model

Let us define:

- a random variable u following a $Gamma(\alpha,\beta)$ distribution (with α, $\beta > 0$) such that
$$f(u;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}u^{\alpha-1}e^{-\beta u}$$
with $E(u) = \frac{\alpha}{\beta}$ and $Var(u) = \frac{\alpha}{\beta^2}$.
We have $E(u) = 1$ for $\alpha = \beta$ and
$$f(u;\alpha,\alpha) = \frac{\alpha^{\alpha}}{\Gamma(\alpha)}u^{\alpha-1}e^{-\alpha u}$$

- a random variable y that conditional on u follows a $Poisson(\lambda u)$ distribution (with $\lambda > 0$) such that
$$f(y;\lambda u) = \frac{e^{-\lambda u}(\lambda u)^{y}}{y!}$$
with $E(y) = Var(y) = \lambda u$.

Then if we consider the following structure:

$$u \sim Gamma(\alpha,\alpha)$$
$$\downarrow$$
$$y|u \sim Pois(\lambda u)$$

it can be proved that Y is marginally distributed according to:

$$y \sim NegBin(\lambda, \alpha).$$

This means that the parameter λ of the Poisson component rules the expectation of the negative binomial distribution, and the parameter α of the Gamma component controls the heterogeneity, allowing overdispersion.

3.2 Finite mixture models

The history of finite mixture models goes back to the end of the XIX century, when the famous biometrician Karl Pearson (Pearson [1894]) fitted a mixture of two normal probability density functions with different means and variances to some data about measurements on the ratio of forehead to body length of 1000 crabs sampled from the Bay of Naples, suggesting that there were two subspecies present.

> *"When a series of measurements gives rise to a normal curve, we may probably assume something approaching a stable condition; there is production and destruction impartially round the mean. In the case of certain biological, sociological, and economic measurements there is, however, a well-marked deviation from this normal shape, and it becomes important to determine the direction and amount of such deviation. The asymmetry may arise from the fact that the units grouped together in the measured material are not really homogeneous. It may happen that we have a mixture of 2, 3, ... n homogeneous groups, each of which deviates about its own mean symmetrically and in a manner represented with sufficient accuracy by the normal curve. Thus an abnormal frequency-curve may be really built up of normal curves having parallel but not necessarily coincident axes and different parameters."* (Pearson [1894])

We denote $y = \{y_1, ..., y_i, ..., y_p\}$ a random sample of size p where $\mathbf{y}_i$ is a n-dimensional random vector (with $n = \sum_j n_j$) with probability density function $f(\mathbf{y}_i)$. If we suppose a K-component mixture model we can write it in parametric form as:

$$f(\mathbf{y}_i; \theta) = \sum_{k=1}^{K} w_k f_k(\mathbf{y}_i; \theta), \tag{3.2}$$

where the vector θ contains all the unknown parameters in the mixture model, i.e. the parameters of the K probability density functions f_k of the K components and the K weights w_k, i.e. the mixing proportions (with the constraints $0 \leq w_k \leq 1$ and $\sum_{k=1}^{K} w_k = 1$).
The data matrix y is called "incomplete" since we actually do not know at which one of the mixture components each unit belongs. A new random variable $z = \{z_1, ..., z_i, ..., z_p\}$ is then introduced, called "allocation variable"; $\mathbf{z}_i$ will be a $K-$ dimensional vector of zeros except from the element that is in the position corresponding to the mixture-component at which the $i - th$ unit belongs.
We will denote $z_{ik} = 1$ when

$$\mathbf{z}_i = \begin{bmatrix} 0 & 0 & \dots & 1 & \dots & 0 \end{bmatrix}$$

$$\uparrow$$

$$\text{k}^{\text{th}} \text{ position}$$

Thus $\mathbf{z}_i$ is assumed to be distributed according to a Multinomial distribution, and in particular:

$$\mathbf{z}_i \sim Multin(1; w_1, ..., w_K) \tag{3.3}$$

We consider now y and z all together we have the "complete data", where y has been observed whilst z is hidden. If the label z of each observation was known, it would be easy to estimate the mixture parameters. However z is a hidden variable, hence it is very difficult to reach closed form formulas for the parameters and it is often complicated to estimate them. We will see in the next section one of the available algorithms to do it.

3.3 Mixtures of Negative Binomials

We consider a random variable $\mathbf{y}_i$ that follows a K-component mixture model of NB distributions:

$$f(\mathbf{y}_i; \theta) = \sum_{k=1}^{K} w_k f_k(\mathbf{y}_i; \theta_k) = \sum_{k=1}^{K} w_k NB_k(\lambda_k, \alpha_k), \tag{3.4}$$

that is:

$$f(\mathbf{y}_i;\theta) = \sum_{k=1}^{K} w_k \binom{y_i + \alpha_k - 1}{\alpha_k - 1} \left(\frac{\lambda_k}{\lambda_k + \alpha_k}\right)^{y_i} \left(\frac{\alpha_k}{\lambda_k + \alpha_k}\right)^{\alpha_k} \tag{3.5}$$

and therefore $\mathbf{y}_i|\mathbf{z}_i \sim NegBin(\lambda_k, \alpha_k)$ hence $E(\mathbf{y}_i|\mathbf{z}_i) = \lambda_k$ and $Var(\mathbf{y}_i|\mathbf{z}_i) = \lambda_k \left(1 + \frac{1}{\alpha_k}\lambda_k\right)$.

3.3.1 Estimation issues

One of the mostly used algorithms for estimating mixture models is the Expectation-Maximization (EM) algorithm (Dempster et al. [1977]). It is able to fit models in presence of missing data and it provides maximum-likelihood estimates, that have useful inferential properties. Nevertheless, as it is well known, the estimation of the whole parameters θ is meaningful only if θ is identifiable. Otherwise, the maximum likelihood estimator would not be consistent, i.e. the estimators would not converge to the true parameters values as the amount of information increases.

Identifiability

We can define a parametric family of densities $f(\mathbf{y}_i, \theta)$ with $\theta \in \Theta$, where Θ is the parameter space; $f(\mathbf{y}_i, \theta)$ is identifiable if distinct values for the parameter θ determine distinct members of the family of densities, that is:

$$f(\mathbf{y}_i, \theta) = f(\mathbf{y}_i, \theta') \text{ if and only if } \theta = \theta'$$

for each θ, $\theta' \in \Theta$.
Identifiability for mixture distributions [McLachlan and Peel, 2000] has to take into account also the possible permutations of the component labels. If we consider $f(\mathbf{y}_i, \theta) = \sum_{k=1}^{K} w_k f_k(\mathbf{y}_i, \theta_k)$ and $f(\mathbf{y}_i, \theta') = \sum_{k=1}^{K'} w'_k f_k(\mathbf{y}_i, \theta'_k)$, we must require that $f(\mathbf{y}_i, \theta) = f(\mathbf{y}_i, \theta')$ if and only if $K = K'$ and we can permute the component labels so that $w_i = w'_i$ and $f_k(\mathbf{y}_i, \theta_k) = f_k(\mathbf{y}_i, \theta'_k)$ $(k = 1, \ldots, K)$ for almost all $\mathbf{y}_i$. The lack of identifiability of θ due to labels permutation can be handled by imposing appropriate constraints, such as $w_1 \leq w_2 \leq \ldots \leq w_K$.

As regards the identifiability of mixtures of NB distributions, we can refer to Yakowitz and Spragins [1968]. In this work, the authors have defined and demon-

strated the following proposition:

> *"The family $\mathfrak{F}$ of all non-degenerate negative binomial distributions induces an identifiable set of finite mixtures".*

The identifiability of mixture models is a largely studied issue, and other more recent works have confirmed the statement above. Among the others we can cite Sapatinas [1995] and, because of the specific structure of the data that will be illustrated in the next chapter, also Allman et al. [2009].

The Expectation-Maximization algorithm

The mostly used algorithm for estimating the parameters of a finite mixture model is the Expectation-Maximization (EM) algorithm. First of all, we have to define the density of the complete data, which can be written

$$f((y,z);\theta) = f(y;\theta)f(z|y;\theta), \tag{3.6}$$

where $f(y;\theta)$ is the density of the observed data and $f(z|y;\theta)$ is the conditional density of the hidden variable given the data.
Therefore we will have to define two different likelihoods:

- the "incomplete-likelihood" is the one that is referred just to the observed data, i.e $L(y;\theta)$;
- the "complete-likelihood" $L_c((y,z);\theta)$ is the one that involves both the observed data and the hidden variable.

Taking the logarithm we have:

$$l_c((y,z);\theta) = l(y;\theta) + logf(z|y;\theta), \tag{3.7}$$

where:

- $l_c((y,z);\theta)$ is the "complete" log-likelihood,
- $l(y;\theta)$ is the "incomplete" likelihood, i.e. referred only to the observed data,
- $logf(z|y;\theta)$ is the log-likelihood of the hidden ("allocation") variable, given the observed data.

The EM algorithm is used in this context by treating the allocation variable z as missing data. It proceeds iteratively in two steps; at the $h-th$ iteration:

at the E-step (Expectation step) we can handle the hidden variable z, computing the expectation of the complete-data log likelihood given the observed data y, using the current estimates for θ (noted $\theta^{(h-1)}$; at the first iteration, it will correspond to the initialization values $\theta^{(0)}$) . The E step requires the computation of

$$E_{z|y;\theta^{(h-1)}}(l_c(\theta)) \tag{3.8}$$

where the subscript $z|y;\theta^{(h-1)}$ means that we are conditioning on y and we are using the estimates of the parameters obtained at the previous iteration;

at the M-step (Maximization step) of each iteration we update the estimates of the parameters, through the computation of the roots of the partial derivatives of the conditional expectation of the complete log-likelihood given y, with respect to each parameter. Thus, at the end of the $h-th$ iteration we have $\widehat{\theta}^{(h)}$ and we will use it for al the computations at the E-step of the $(h+1)-th$ iteration.

The E-step and the M-step are alternately repeated until convergence is reached.

One of the classical criteria to detect convergence consists in computing

$$\frac{l^{(h)} - l^{(h-1)}}{\mid l^{(h-1)} \mid} < \epsilon$$

where $l^{(h)}$ is the likelihood at the h-th iteration, and ϵ is an arbitrary small value.

Dempster et al. [1977] proved that the likelihood function is monotonically not decreasing during the EM iteration. Nevertheless, it has to be noted that it could get trapped in local maxima and the initialization is a crucial point. One possible procedure for overcoming this problem and getting good approximations for the global maximum has been proposed by Böhning [2003]. This strategy is based on the combination of the EM algorithm with a gradient function update. At least for the moment, we will not go into details of such procedure because the EM algorithm seems to provide consistent results, as it will be shown in simulation studies.

Chapter 4

The proposed method

In this chapter we are going to propose a new strategy for estimating the variances of RNA-seq data, based on mixtures of NB distributions (Section 4.1). Afterwards three consistent statistical procedures for performing differential analysis are developed (Section 4.2).

4.1 The proposal

As it has been described in Section 3.1, an interesting characteristic of the Negative Binomial distribution is that it can be derived from a Poisson-Gamma mixed process, defined by an heterogeneity component u that follows a Gamma distribution, and a random variable y that, conditioning on u, follows a Poisson distribution.
The innovative idea of the proposed method consists in considering the heterogeneity component u as a random variable distributed as a mixture of Gamma distributions, with the purpose of getting a more reliable estimation of the dispersion parameters:

$$\begin{gathered}\mathbf{z}_i \sim Multinom(1, \mathbf{w}) \text{ where } \mathbf{w} = (w_1, ..., w_K) \\ \downarrow \\ \mathbf{u}_i | z_{ik} = 1 \sim Gamma(\alpha_k, \alpha_k) \\ \downarrow \\ \mathbf{y}_i | \mathbf{u}_i \sim Pois(\lambda_i u_i)\end{gathered} \tag{4.1}$$

and therefore $\mathbf{y}_i|\mathbf{z}_i \sim NegBin(\lambda_i, \alpha_k)$ hence $E(\mathbf{y}_i|\mathbf{z}_i) = \lambda_i$ and

$Var(\mathbf{y}_i|\mathbf{z}_i) = \lambda_i \left(1 + \frac{1}{\alpha_k}\lambda_i\right)$.

In so doing $\mathbf{y}_i$ follows a particular Negative Binomial mixture distribution:

$$\mathbf{y}_i \sim \sum_k w_k NegBin\left(\lambda_i, \alpha_k\right) \tag{4.2}$$

Proof.

$$\begin{aligned}
f(y) &= \int_0^{+\infty} f(y,u)du \\
&= \int_0^{+\infty} f(y|u)f(u)du \\
&= \int_0^{+\infty} \frac{e^{-(\lambda u)}(\lambda u)^y}{y!} \sum_k w_k \frac{\alpha_k^{\alpha_k}}{\Gamma(\alpha_k)} u^{\alpha_k - 1} e^{-\alpha_k u} du \\
&= \frac{1}{y!}\lambda^y \sum_k w_k \frac{\alpha_k^{\alpha_k}}{\Gamma(\alpha_k)} \int_0^{+\infty} u^y u^{\alpha_k - 1} e^{-\alpha_k u} e^{-\lambda u} du \\
&= \sum_k w_k \frac{\lambda^y \alpha_k^{\alpha_k}}{\Gamma(y+1)\Gamma(\alpha_k)} \underbrace{\int_0^{+\infty} u^{y+\alpha_k - 1} e^{-u(\alpha_k + \lambda)} du}_{\text{kernel of a } Gamma(y+\alpha_k, \alpha_k + \lambda)} \\
&= \sum_k w_k \frac{\lambda^y \alpha_k^{\alpha_k}}{\Gamma(y+1)\Gamma(\alpha_k)} \frac{\Gamma(y+\alpha_k)}{(\lambda + \alpha_k)^{y+\alpha_k}} \\
&= \sum_k w_k \binom{y + \alpha_k - 1}{\alpha_k - 1} \left(\frac{\lambda}{\lambda + \alpha_k}\right)^y \left(\frac{\alpha_k}{\lambda + \alpha_k}\right)^{\alpha_k}
\end{aligned} \tag{4.3}$$

and we can recognize that this is the probability density function of a mixture of K Negative Binomial distributions each one with expectation equal to λ and dispersion parameter equal to α_k. □

It is interesting to underline that in this particular mixture model the expectation λ_i is not component-varying, and the estimation of just K dispersion parameters (where $K \ll p$) is required.

During the last years, many methods to estimate the dispersion parameter have been proposed. The two extreme possibilities are: “a specific estimation per gene”, that is the most realistic but is based on few information and therefore is not reliable, and “one estimation for all the genes”, that is based on more data but is too restrictive.

We consider the use of mixture models to share information between genes with

similar heterogeneity in order to estimate the dispersion parameter on the basis of a larger number of observations.

4.1.1 Modeling RNA-Seq data

The NGS data we will analyze have a hierarchical structure. Borrowing the terminology of multilevel models we have:

1. first-level units are the replicates ($r = 1, ..., n_j$);
2. at the second level we have the conditions ($j = 1, ..., d$);
3. at the third level there are the genes ($i = 1, ..., p$).

Therefore this is the hierarchical structure of our model:

$$\begin{gathered}\mathbf{z}_i \sim Multinom(1, \mathbf{w}) \text{ where } \mathbf{w} = (w_1, ..., w_K) \\ \downarrow \\ \{u_{j,r}\}_i | z_{ik} = 1 \sim Gamma(\alpha_k, \alpha_k) \\ \downarrow \\ \{y_{j,r}\}_i | \{u_{j,r}\}_i \sim Pois(\lambda_{ij} u_{ijr})\end{gathered}$$

and $\{y_{j,r}\}_i | \{z_{j,r}\} \sim NegBin(\lambda_{ij}, \alpha_k)$ from which $E(y_{ijr}|z_{ijr}) = \lambda_{ij}$ and $Var(y_{ijr}|z_{ijr}) = \lambda_{ij}\left(1 + \frac{1}{\alpha_k}\lambda_{ij}\right)$
and marginalizing with respect to z, $\{y_{j,r}\}_i$ follows a Negative Binomial mixture distribution:

$$\{y_{j,r}\}_i \sim \sum_k w_k NegBin(\lambda_{ij}, \alpha_k). \tag{4.4}$$

Therefore we consider the model:

$$f(\mathbf{y}_i) = \sum_{k=1}^{K} w_k f(\mathbf{y}_i | z_{ik} = 1), \tag{4.5}$$

where

$$f(\mathbf{y}_i | z_{ik} = 1) = \prod_{j=1}^{d} f(\mathbf{y}_{ij} | z_{ik} = 1), \tag{4.6}$$

and

$$f(\mathbf{y}_{ij} | z_{ik} = 1) = \prod_{r=1}^{n_j} f(y_{ijr} | z_{ik} = 1), \tag{4.7}$$

where $y_{ijr}|z_{ik} = 1$ is the $r - th$ realization from a Negative Binomial distribution with mean equal to λ_{ij} and Variance equal to $\lambda_{ij}\left(1 + \frac{1}{\alpha_k}\lambda_{ij}\right)$.

Hence

$$f(\mathbf{y}_i) = \sum_{k=1}^{K} w_k \prod_{j=1}^{d} \prod_{r=1}^{n_j} f(y_{ijr}|z_{ik} = 1), \tag{4.8}$$

and, for $\theta = \{\lambda_{ij}, w_k, \alpha_k\}_{i=1,...p;j=1,...,d;k=1,...,K}$ the (incomplete) likelihood function will be:

$$L(\theta) = \prod_{i=1}^{p} \sum_{k=1}^{K} w_k \prod_{j=1}^{d} \prod_{r=1}^{n_j} f(y_{ijr}|z_{ik} = 1). \tag{4.9}$$

4.1.2 Estimation

Given the features of our model, we have to derive an EM algorithm with two hidden layers (one for the Gamma component u, that rules the heterogeneity, and one for the Multinomial component z, that rules the mixture).

The complete Log-Likelihood

The complete log-likelihood for our model is:

$$\begin{aligned} l_c(\theta) &= \sum_{i=1}^{p} \sum_{j=1}^{d} \sum_{r=1}^{n_j} \ln f(y_{ijr}, u_{ijr}, \mathbf{z}_i) \\ &= \sum_{i} \sum_{j} \sum_{r} \ln(f(y_{ijr}|u_{ijr}, \mathbf{z}_i)) + \sum_{i} \sum_{j} \sum_{r} \ln(f(u_{ijr}|\mathbf{z}_i)) + \sum_{i} \ln(f(\mathbf{z}_i)) \end{aligned} \tag{4.10}$$

The initialization

One of the most crucial points concerning the EM algorithm is the initialization. For $h = 1$, that is at the first iteration, we initialize the parameters as follows:

- for the w_k we draw K values from K $Uniform(0, 1)$ distributions and then we normalize these values so as to respect the constraint $\sum_{k=1}^{K} w_k = 1$;
- for the α_k we generate a regular sequence of K values from 5 and 700. This range has been chosen considering plausible values from the empirical situation, and considering a regular sequence can be a good way for

exploring the whole interval;

- for the λ_{ij} we compute the sample mean (that is also the ML estimate), i.e. initial λ_{ij}= $\frac{\sum_{r=1}^{n_j} y_{ijr}}{n_j}$.

The algorithm

The EM algorithm consists of two steps (the E-step and the M-step) that are repeated alternately until the stop criterion is satisfied.

The E-step

Considering the $h-th$ iteration, at the E-step we have to compute the $E_{z,u|y,\theta^{(h-1)}}(l_c(\theta))$, i.e. the Expectation of the complete-likelihood assuming that $\theta = \theta^{(h-1)}$ and conditioning on y but considering both the two hidden variables u and z as random.

Therefore we have to calculate:

$$
\begin{aligned}
&E_{z,u|y,\theta^{(h-1)}}(l_c(\theta)) = \\
&\quad \int_0^{+\infty} \sum_{k=1}^{K} \sum_{i=1}^{p} \sum_{j=1}^{d} \sum_{r=1}^{n_j} \ln(f(y_{ijr}|\mathbf{z}_i, u_{ijr};\theta)) f(u_{ijr}, \mathbf{z}_i|\mathbf{y}_i;\theta^{(h-1)}) du_{ijr} + \\
&+ \int_0^{+\infty} \sum_{k} \sum_{i} \sum_{j} \sum_{r} \ln(f(u_{ijr}|z_{ik}=1;\theta)) f(u_{ijr}, z_{ik}=1|\mathbf{y}_i;\theta^{(h-1)}) du_{ijr} + \\
&\qquad + \sum_{k} \sum_{i} \ln(f(z_{ik}=1|\theta)) f(z_{ik}=1|\mathbf{y}_i;\theta^{(h-1)}) \quad (4.11)
\end{aligned}
$$

Where we have:

$$f(y_{ijr}|\mathbf{z}_i, u_{ijr}) = Pois(\lambda_{ij} u_{ijr}) \tag{4.12}$$

$$f(u_{ijr}, \mathbf{z}_i|\mathbf{y}_i) = f(u_{ijr}|\mathbf{y}_i, \mathbf{z}_i) f(\mathbf{z}_i|\mathbf{y}_i) \tag{4.13}$$

with:

$$\begin{aligned} f(u_{ijr}|\mathbf{y}_i, \mathbf{z}_i) &= \frac{f(\mathbf{y}_i|u_{ijr}, \mathbf{z}_i) f(u_{ijr}|\mathbf{z}_i) f(\mathbf{z}_i)}{f(\mathbf{y}_i|\mathbf{z}_i) f(\mathbf{z}_i)} \\ &= \frac{f(u_{ijr}|\mathbf{z}_i) \prod_j \prod_r f(y_{ijr}|u_{ijr})}{\prod_j \prod_r f(y_{ijr}|\mathbf{z}_i)} \\ &\propto \prod_j \prod_r e^{-u_{ijr}\lambda_{ij}} (\lambda_{ij} u_{ijr})^{y_{ijr}} u_{ijr}^{\alpha_k - 1} e^{-\alpha_k u_{ijr}} \\ &= \prod_j \prod_r e^{-u_{ijr}(\lambda_{ij} + \alpha_k)} u_{ijr}^{\alpha_k - 1 + y_{ijr}} \end{aligned} \tag{4.14}$$

and since we are considering just the single u_{ijr}, we can recognize that it is simply the kernel of a $Gamma(y_{ijr} + \alpha_k, \lambda_{ij} + \alpha_k)$ since all the factors of the products that concerned $i' \neq i$ and $j' \neq j$ can be viewed as constant terms;

$$f(\mathbf{z}_i|\mathbf{y}_i) = \frac{f(\mathbf{y}_i|\mathbf{z}_i) f(\mathbf{z}_i)}{f(\mathbf{y}_i)} = \frac{w_k \prod_j \prod_r f(y_{ijr}|\mathbf{z}_i)}{\sum_k w_k \prod_j \prod_r f(y_{ijr}|z_{ik} = 1)} \tag{4.15}$$

$$f(u_{ijr}|\mathbf{z}_i) = Gamma(\alpha_k, \alpha_k) \tag{4.16}$$

$$f(\mathbf{z}_i) = Multin(1, \mathbf{w}) \tag{4.17}$$

The M-step

The M step consists in maximizing the conditional expectation of the complete-likelihood with respect to each one of the parameters that have to be estimated. For the estimation of λ_{ij} we can focus on the first term of (4.11) given that it is the only one addend that involves the parameters λ_{ij}.

$$
\begin{aligned}
&\frac{\partial}{\partial \lambda_{ij}} E_{z,u|y,\theta^{(h-1)}}(l_c(\theta)) = \\
&= \frac{\partial}{\partial \lambda_{ij}} \int_0^{+\infty} \sum_{k=1}^{K} \sum_{i=1}^{p} \sum_{j=1}^{d} \sum_{r=1}^{n_j} \ln f(y_{ijr}|\mathbf{z}_i, u_{ijr}) f(u_{ijr}, \mathbf{z}_i|\mathbf{y}_i; \theta^{(h-1)}) du_{ijr} \\
&= \frac{\partial}{\partial \lambda_{ij}} \int_0^{+\infty} \sum_{k} \sum_{i} \sum_{j} \sum_{r} \ln \underbrace{f(y_{ijr}|z_{ik}=1, u_{ijr})}_{Pois(\lambda_{ij} u_{ijr})} f(u_{ijr}|\mathbf{y}_i, z_{ik}=1) \cdot \\
&\qquad\qquad \cdot f(z_{ik}=1|\mathbf{y}_i) du_{ijr} \\
&= \frac{\partial}{\partial \lambda_{ij}} \int_0^{+\infty} \sum_{k} \sum_{i} \sum_{j} \sum_{r} (-\lambda_{ij} u_{ijr} + y_{ijr} \ln \lambda_{ij} - \ln y_{ijr}!) f(u_{ijr}|y_{ijr}, z_{ik}=1) \cdot \\
&\qquad\qquad \cdot f(z_{ik}=1|\mathbf{y}_i) du_{ijr} \\
&= \int_0^{+\infty} \sum_{k} \sum_{r} \left(-u_{ijr} + \frac{y_{ijr}}{\lambda_{ij}} \right) f(u_{ijr}|y_{ijr}, z_{ik}=1) f(z_{ik}=1|\mathbf{y}_i) du_{ijr} \\
&= \sum_{r} \frac{y_{ijr}}{\lambda_{ij}} - \sum_{k} \sum_{r} E(u_{ijr}|y_{ijr}, z_{ik}=1) f(z_{ik}=1|\mathbf{y}_i)
\end{aligned}
\tag{4.18}
$$

and if we set it equal to 0 we have that

$$
\widehat{\lambda_{ij}} = \frac{\sum_r y_{ijr}}{\sum_k f(z_{ik}=1|\mathbf{y}_i) \sum_r E(u_{ijr}|\mathbf{y}_i, \mathbf{z}_i)} \tag{4.19}
$$

where $E(u_{ijr}|\mathbf{y}_i, \mathbf{z}_i) = \frac{y_{ijr}+\alpha_k}{\lambda_{ij}+\alpha_k}$. It can be proved (see Appendix A.1) that (4.19) simply reduces to: $\widehat{\lambda_{ij}} = \frac{\sum_r y_{ijr}}{n_j}$

With regards to α_k we can focus on the second term of (4.11) and we need an optimization algorithm, since

$$
\begin{aligned}
&\frac{\partial}{\partial \alpha_k} E_{z,u|y,\theta^{(h-1)}}(l_c(\theta)) = \\
&= \frac{\partial}{\partial \alpha_k} \int_0^{+\infty} \sum_{k=1}^{K} \sum_{i=1}^{p} \sum_{j=1}^{d} \sum_{r=1}^{n_j} \ln f(u_{ijr}|\mathbf{z}_i) f(u_{ijr}, \mathbf{z}_i|\mathbf{y}_i; \theta^{(h-1)}) du_{ijr} \\
&= \frac{\partial}{\partial \alpha_k} \int_0^{+\infty} \sum_{k} \sum_{i} \sum_{j} \sum_{r} (\alpha_k \ln \alpha_k + (\alpha_k - 1) \ln u_{ijr} - \alpha_k u_{ijr} - \ln \Gamma(\alpha_k)) \\
&f(u_{ijr}|y_{ijr}, \mathbf{z}_i) f(z_{ik} = 1|\mathbf{y}_i) du_{ijr} \\
&= \alpha_k \ln \alpha_k - \ln \Gamma(\alpha_k) + (\alpha_k - 1) E(\ln u_{ijr}|y_{ijr}, z_{ik} = 1) - \\
&\qquad - \alpha_k E(u_{ijr}|y_{ijr}, z_{ik} = 1) f(\mathbf{z}_i|\mathbf{y}_i)
\end{aligned}
\tag{4.20}
$$

and this does not have close-form solution. We can make use of a quasi-Newton algorithm that is called `L-BFGS`. This strategy has been proposed by Byrd et al. [1995] for solving large nonlinear optimization problems, when bounds of the searching interval are provided by the user. This procedure provides solutions on the basis of the second-order Taylor expansion of the function that has to be maximized, and an approximation of the Hessian matrix [Byrd et al., 1994] is used in order to overcome many issues related to its exact computation. It also makes use of the gradient projection method for detecting useful constraints at each iteration.

With regards to $E(\ln u_{ijr}|y_{ijr}, \mathbf{z}_i)$ we can use an already-known results that states that, given a random variable $X \sim Gamma(\alpha, \beta)$, $E(\ln X) = \psi(\alpha) - \ln(\beta)$ where ψ is the digamma function. Thus, for the (4.14) we have:

$$
E(\ln u_{ijr}|y_{ijr}, \mathbf{z}_i) = \psi(y_{ijr} + \alpha_k) - \ln(\lambda_{ij} + \alpha_k) \tag{4.21}
$$

Finally for w_k we can take advantage of an already known result, since the third addend of (4.11) is usually present in all the mixture models. Introducing Lagrange multipliers (considering the constraints that $w_k \geq 0$ and $\sum_{k=1}^{K} w_k = 1$) we obtain:

$$
\widehat{w_k} = \frac{\sum_i f(\mathbf{z}_i|\mathbf{y}_i)}{p} \tag{4.22}
$$

Hence $\theta^{(h)} = \{\lambda_{ij}, w_k, \alpha_k\}_{i=1,\ldots p; j=1,\ldots,d; k=1,\ldots,K}$ and we can proceed with the $(h+1) - th$ iteration until convergence.

The EM algorithm has been implemented in R. A discussion about how to choose the number of components K is defer to Section 5.1, and in Section 5.2 an analysis of the properties of the estimations that we get is presented.

4.2 The proposed test statistics

Our final goal is to identify the genes ($i = 1, ..., p$) that differentially express under two ($j = 1, 2$) different conditions.
Hence, now that we have defined and estimated our model we want to make use of three different statistical testing procedures with the aim of comparing the expression levels of a gene in the two conditions:

1. $H_0 : \lambda_{i1} - \lambda_{i2} = 0$
2. $H_0 : \frac{\lambda_{i1}}{\lambda_{i2}} = 1$
3. $H_0 : \ln \frac{\lambda_{i1}}{\lambda_{i2}} = \ln(\lambda_{i1}) - \ln(\lambda_{i2}) = 0$

In order to perform testing, we can use the Wald test (Wald [1943]) since the estimations provided by the EM algorithm are maximum likelihood (ML); as it is known, the maximum-likelihood estimators have many asymptotic properties as consistency, normality, efficiency, (i.e., they achieve the Cramér–Rao lower bound). Although we have few observations for each gene, these properties can be considered as approximately true even for our data, as it will be shown by the results of our simulation studies.

Therefore we can propose the following alternative test-statistics:

1. "Difference": $\frac{\widehat{\lambda}_{i1} - \widehat{\lambda}_{i2}}{\sqrt{Var(\widehat{\lambda}_{i1} - \widehat{\lambda}_{i2})}} | H_0 \sim N(0, 1)$
2. "Ratio": $\frac{\frac{\widehat{\lambda}_{i1}}{\widehat{\lambda}_{i2}} - 1}{\sqrt{Var\left(\frac{\widehat{\lambda}_{i1}}{\widehat{\lambda}_{i2}}\right)}} | H_0 \sim N(0, 1)$
3. "Log Ratio": $\frac{\ln \widehat{\lambda}_{i1} - \ln \widehat{\lambda}_{i2}}{\sqrt{Var(\ln \widehat{\lambda}_{i1} - \ln \widehat{\lambda}_{i2})}} | H_0 \sim N(0, 1)$

where $\widehat{\lambda}_{i1}$ and $\widehat{\lambda}_{i2}$ are the EM-estimators.

4.2.1 Variances for each test statistic

For computing the test statistics we need to estimate the variances at each denominator. First of all we can make some observations:

1. $Var(\widehat{\lambda}_{i1} - \widehat{\lambda}_{i2}) = Var(\widehat{\lambda}_{i1}) + Var(\widehat{\lambda}_{i2})$
2. To compute $Var\left(\frac{\widehat{\lambda}_{i1}}{\widehat{\lambda}_{i2}}\right)$ we can use Delta method (van der Vaart [2000]), and as reported in Cox 1990 (Cox [1990]) we have:

$$f(\lambda_{i1}, \lambda_{i2}) = \frac{\lambda_{i1}}{\lambda_{i2}}$$

$$\frac{\partial f(\lambda_{i1}, \lambda_{i2})}{\partial \lambda_{i1}} = \frac{1}{\lambda_{i2}}$$

$$\frac{\partial f(\lambda_{i1}, \lambda_{i2})}{\partial \lambda_{i2}} = -\frac{\lambda_{i1}}{{\lambda_{i2}}^2}$$

hence

$$Var\left(\frac{\widehat{\lambda}_{i1}}{\widehat{\lambda}_{i2}}\right) \approx \frac{Var(\widehat{\lambda}_{i1})}{E(\widehat{\lambda}_{i2})^2} + \frac{E(\widehat{\lambda}_{i1})^2}{E(\widehat{\lambda}_{i2})^4} Var(\widehat{\lambda}_{i2})$$

3. $Var(\ln \widehat{\lambda}_{i1} - \ln \widehat{\lambda}_{i2}) = Var(\ln \widehat{\lambda}_{i1}) + Var(\ln \widehat{\lambda}_{i2})$

Computing $E(\widehat{\lambda}_{ij})$

As regards the $E(\widehat{\lambda}_{ij})$ we can say that

$$E(\widehat{\lambda}_{ij}) = \lambda_{ij}$$

since MLE are asymptotically correct.

Computing $Var(\widehat{\lambda}_{ij})$

$Var(\widehat{\lambda}_{ij})$ is the variance of a function of y_{ijr}; in particular, noting $\sum_r y_{ijr}$ as y_{ij+}:

$$\widehat{\lambda}_{ij} = f(y_{ij+}) = \frac{y_{ij+}}{n_j} \tag{4.23}$$

Therefore:

$$Var(\widehat{\lambda}_{ij}) = \frac{1}{n_{ij}^2} Var(y_{ij+})$$

Computing $Var(y_{ij+})$

For the law of total variance,

$$Var(y_{ij+}) = E[Var(y_{ij+}|z_{ik}=1)] + Var[E(y_{ij+}|z_{ik}=1)] \tag{4.24}$$

where $Var[E(y_{ij+}|z_{ik}=1)]$ = 0 because the expectation is not component varying, and as regards $E[Var(y_{ij+}|z_{ik}=1)]$ we can consider the conditional expectation given the observed data because of the multilevel structure of the data, and therefore:

$$\begin{aligned} Var(y_{ijr}) &= E_{\mathbf{z}_i|\mathbf{y}_i}[Var(y_{ijr}|z_{ik}=1)] = \\ &\widehat{\lambda}_{ij}\left(1+\widehat{\lambda}_{ij}\sum_k \frac{f(z_{ik}|\mathbf{y}_i)}{\widehat{\alpha}_k}\right). \end{aligned} \tag{4.25}$$

This formula enlightens the effect of the mixture model we propose: the over-dispersion term is a weighted average of the (estimated) over-dispersion terms λ_{ij}/α_k one would get in each component of the mixture. These terms are weighted according to the posterior probability for observation i to belong to each component k: $f(z_{ik}|\mathbf{y}_i)$.

Computing Var($\ln \lambda_{ij}$**)**

$$\begin{aligned} Var(\ln\lambda_{ij}) &= Var(\ln(f(y_{ij+}))) \\ &= Var\left(\ln\left(\frac{y_{ij+}}{\sum_k f(z_{ik}=1|\mathbf{y}_i)\sum_r E(u_{ijr}|\mathbf{y}_i,\mathbf{z}_i)}\right)\right) \\ &= Var(\ln(y_{ij+})) \end{aligned} \tag{4.26}$$

and using Delta method we can state that

$$\begin{aligned} Var(\ln y_{ij+}) = Var(g(y_{ij+})) &\approx Var(y_{ij+})\left(\frac{\partial}{\partial y_{ij+}}g(y_{ij+})\right)^2 \\ &= \frac{1}{y_{ij+}^2}Var(y_{ij+}) \end{aligned} \tag{4.27}$$

Chapter 5

A simulation study

The performance of the proposed strategy is evaluated by a large simulation study comprising several data generating processes.

First of all, in a first simulation we have studied the behavior of the proposed strategy and its capability in estimating the variances as the number of components of the mixture model varies. Afterwards in a second simulated experiment we have evaluated the properties of the estimates provided by the EM algorithm as the number of replicates n_j increases, together with the accuracy of the three statistical test procedures in terms of power and first-type error. It is important to underline that for each data-set we have to test p null-hypothesis H_0^i ($i = 1, \ldots, p$) for the p genes, each one by itself.
In particular, a statistical test procedure can be considered as well working and reliable if the empirical first type error (i.e. the proportion of times in which we reject the null-hypothesis for the genes that have been drawn as not differentially expressed) reaches the nominal value. Otherwise, if this does not occur, it means that the statistical distribution that has been considered as the reference one for the statistic under the null hypothesis is not correct.

We have compared the results of the three proposed statistical tests with the procedures of Robinson et al. [2010], Anders and Huber [2010] and Wu et al. [2013], implemented in the `R` packages *edgeR*, *DESeq* and *DSS*, respectively.

5.1 Simulation A

In the first simulation study, we evaluated the capability of the proposed mixture model to estimate the variances of the genes as the number of components, K, increases. We also computed some conventional information criteria in order to select the optimal number of components. We have simulated $H = 100$ data-sets, each one drawing:

- $p = 300$ genes, of which:
  - $\frac{1}{3}$ genes (= 100 genes) are differentially expressed ($\lambda_{i1} \neq \lambda_{i2}$)
  - $\frac{2}{3}$ genes (= 200 genes) are not differentially expressed ($\lambda_{i1} = \lambda_{i2}$)
- $d = 2$ conditions
- $n_j = 5$ replicates for each condition;
- as regards λ_{ij}:
  - for the 100 differentially expressed genes $\lambda_{i1} \sim Unif(0, 250)$; $\lambda_{i2} = \frac{\lambda_{i1}}{e^{\phi_i}}$ where ϕ_i is randomly drawn from a $N(\mu = 1, \sigma = 0.125)$
  - for the 200 non-differentially expressed genes $\lambda_{i1} \sim Unif(0, 250)$ and $\lambda_{i1} = \lambda_{i2}$
- for all the genes, the α_i ($i = 1, \ldots, p$) are randomly drawn from a $Unif(0.5, 600)$.

All the values for the parameters have been chosen to be consistent with the empirical situation.

On each data-set we fitted the proposed mixture model for K ranging from 1 to 6, and we have computed the relative distances in absolute values across the 100 data-sets between the estimated variances $\widehat{Var(y_{ijr})}$ and the true ones $Var(y_{ijr})$ as K varies, as follows:

$$\text{distance}_{ij}^{(h)} = \frac{|Var(y_{ijr}) - \widehat{Var(y_{ijr})}^{(h)}|}{Var(y_{ijr})} \tag{5.1}$$

for $j = 1, 2$, $i = 1, \ldots, p$ and $h = 1, \ldots, H$; for the computation of the variances, since the number of replicates is limited we applied the factor $n_j/(n_j - 1)$ to the variance in (4.25), in order to obtain the corresponding correct estimator. Figure 5.1 shows the average distances and the standard error bands (mean $\pm$ 2$\cdot$ standard

error), and looking at the graph we could say that from $K = 3$ components the gain of fitting more complex mixture models becomes irrelevant. In other terms, it seems that $K = 2$ and $K = 3$ components well describe the variability of the p genes.

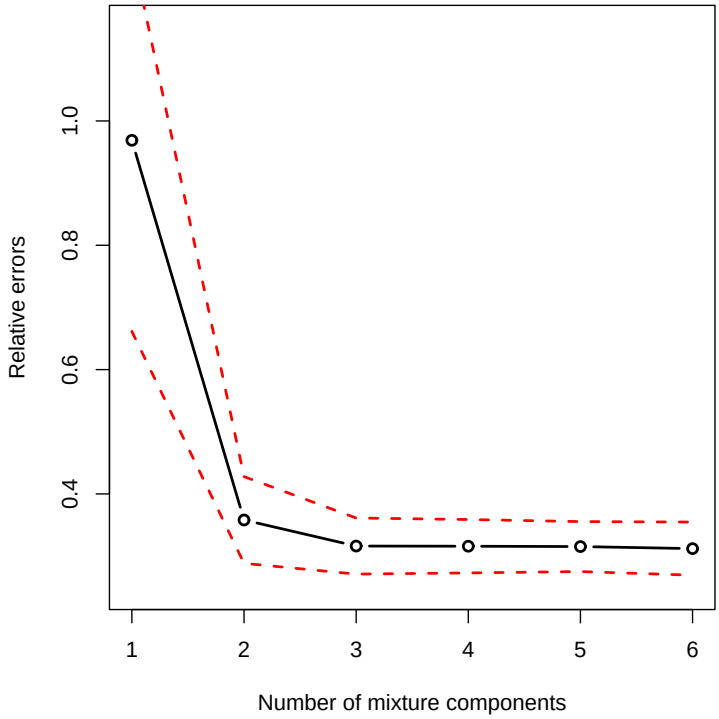

Figure 5.1: Simulation A: Relative distances between the estimated variances and the true ones as K varies. The dashed lines depict the standard error bands (mean $\pm 2 \cdot$ se).

This insight is also confirmed by the information criteria. More specifically, we have considered the Akaike's Information Criterion (Akaike [1974]), AIC $= -2 \log \max L + 2b$, where b is the total number of required parameters and the more conservative Bayesian Information Criterion (Schwarz et al. [1978]), BIC $= -2 \log \max L + b \log p$. In addition, we have also computed the so-called Integrated Classification Likelihood Criterion (Biernacki et al. [2000]) that combines the BIC penalty term with the entropy of the posterior classification. As a result, ICL-BIC is characterized by a heavier penalty term and it tends to favor simpler models against mixture models with more components.

In Table 5.1 the number of times each criterion suggests a specific number of components K is shown. These results recommend that $K = 3$ mixture components are enough to give a good description of the data.

Table 5.1: Simulation A: number of times each information criterion suggests a specific value for K.

K	AIC	BIC	ICL-BIC
1	0	0	0
2	0	2	76
3	76	86	24
4	6	4	0
5	6	4	0
6	12	4	0

In Figure 5.2 the classic trend that we can observe for these information criteria is shown, and it is interesting to note that it is very similar to the one that describes the relative errors in the computation of the variances.

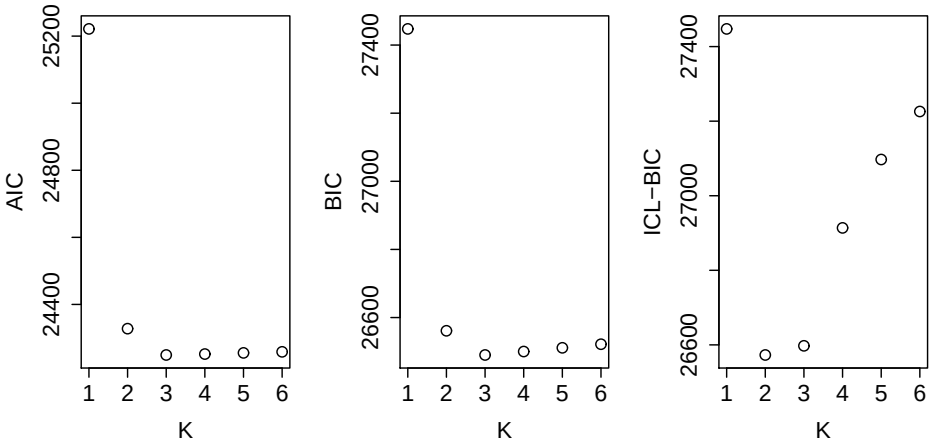

Figure 5.2: Simulation A: Information criteria as K varies.

The relative errors of the variances have been computed also for the other already-known methods, using the Formula (5.1). In Figure 5.3 the boxplots that describe the distribution of these measures of bias for all the genes are presented. It is clear from this graph that the proposed method actually improves the estimation of the variances.

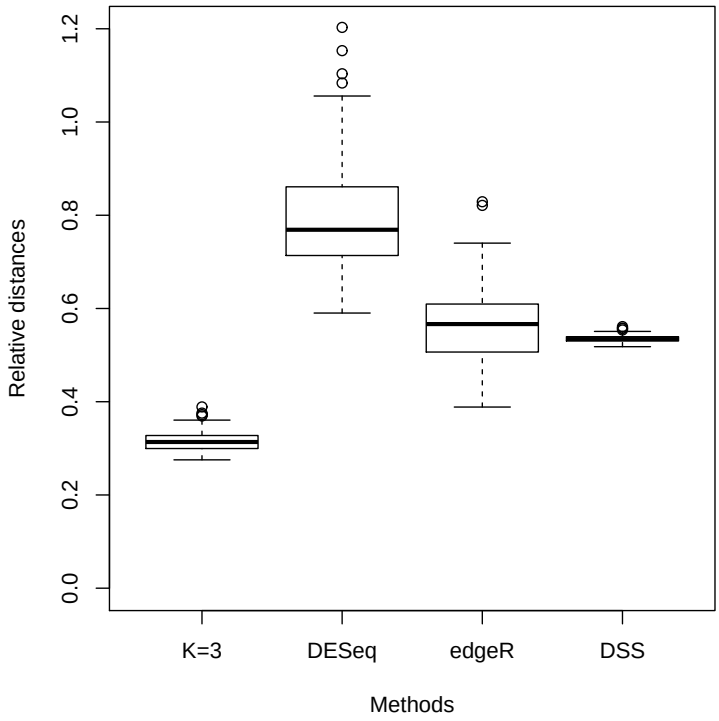

Figure 5.3: Simulation A: Relative distances between the estimated variances and the true ones for all the methods.

5.2 Simulation B

In this simulation study we considered the same data-generating scheme presented above, with a varying number of replicates $n_j = 3, 5, 10$. For each case we have generated $H = 1000$ data-sets, and we have estimated the mixture model with $K = 3$ components, accordingly to the results provided by Simulation A. The properties of the ML estimators that we get through the EM algorithm have been studied, and afterwards the three proposed test statistics have been computed. For comparative purposes, we have performed all the analyses also through *DESeq*, *edgeR* and *DSS*.

5.2.1 Properties of the EM algorithm estimates

For the analysis of the properties of the EM algorithm estimates for the proposed Negative Binomial mixture model we have drawn inspiration from the paper by Nityasuddhi and Böhning [2003], and for each number-of-replicates simulation

scheme ($n_j = 3, 5, 10$) we have computed the bias of the λ estimators:

$$\text{BIAS}(\widehat{\lambda}_{ij}) = \frac{\sum_{h=1}^{H} \widehat{\lambda}_{ij}^{(h)}}{H} - \lambda_{ij}, \tag{5.2}$$

where $\widehat{\lambda}_{ij}^{(h)}$ is the estimation for λ_{ij} that we obtain from the analysis of the $h - th$ dataset.

We have computed also the mean square errors as:

$$\text{MSE}(\widehat{\lambda}_{ij}) = \frac{\sum_{h=1}^{H} (\widehat{\lambda}_{ij}^{(h)} - \lambda_{ij})^2}{H}. \tag{5.3}$$

The same error-measures have been computed also for the variances $Var(y_{ijr})$. The analysis of these indicators offers a global idea of the ability of the algorithm, since for obtaining consistent estimates for the variances it is necessary to have consistent estimates of all the parameters (see (4.25)).

It is interesting to study the distribution of these quantities as n_j varies, for assessing the capability of the mixture model to improve the estimates as the number of observations increases.

In Figures 5.4, 5.5, 5.6 and 5.7 we can see the box-plots describing their behavior. As expected, both the BIAS and the MSE suggest that the EM algorithm provides reasonable estimates of the parameter values, and they get better as n_j increases, thus revealing consistency.

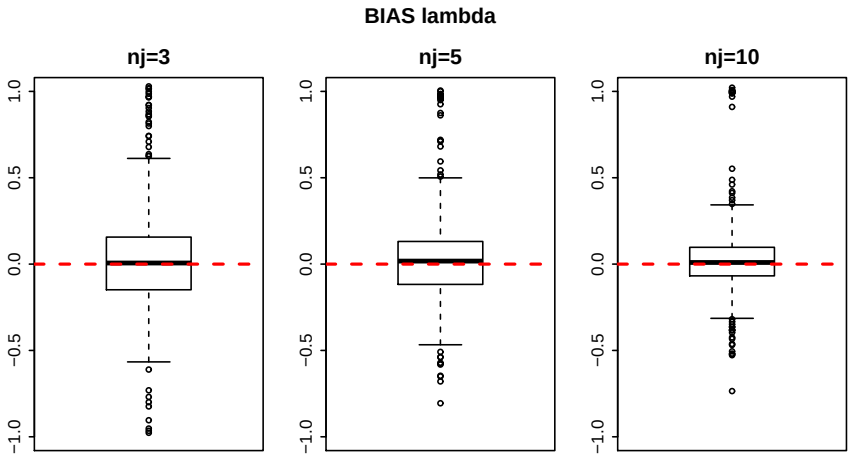

Figure 5.4: Box-plots of the distribution of the BIAS for the λ_{ij} estimators.

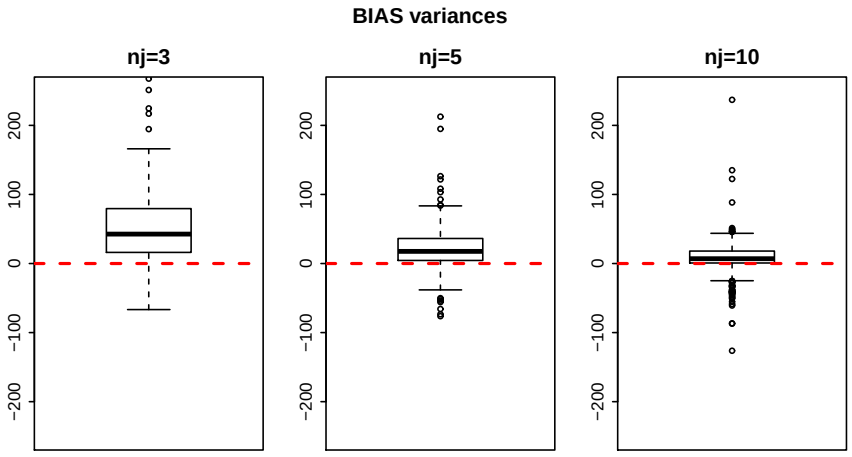

Figure 5.5: Box-plots of the distribution of the BIAS for the $Var(y_{ijr})$ estimators.

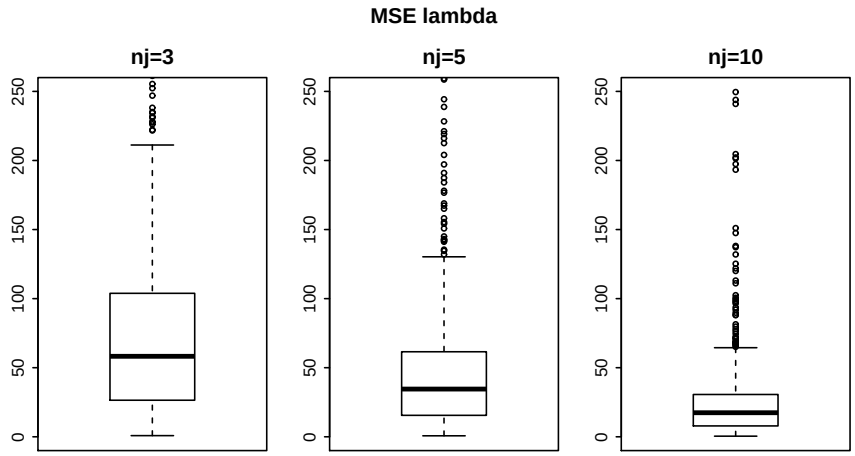

Figure 5.6: Box-plots of the distribution of the mean square errors for the λ_{ij} estimators.

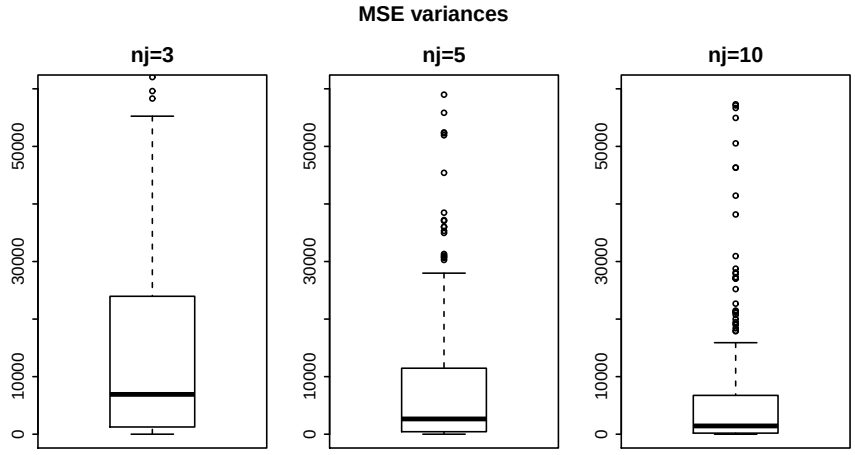

Figure 5.7: Box-plots of the distribution of the mean square errors for the $Var(y_{ijr})$ estimators.

5.2.2 The first-type error

The adequateness of the statistical procedures can be evaluated by observing the approximation towards the nominal significance level under the null hypothesis as the number of replicates increases. For each of the 200 not-differentially expressed genes, we have computed the empirical first-type errors across the 1000 data-sets for all the considered strategies.

Figures 5.8, 5.9 and 5.10 report the box-plots of the empirical first-type errors that we get through each one of the different methods as the number of replicates varies; different levels of the test (0.05, 0.01 and 0.001) have been considered.

It can be useful to underline that we have a value for each gene, describing the empirical first-type error for the specific null hypothesis H_0^i ($i = 1, \ldots, p$). The three proposed statistical tests fast converge to the nominal level (indicated by the red dashed line) as the number of the replicates increases, while the *DESeq*, *edgeR* and *DSS* based tests are always under the nominal level. It is clear from these graphs that the proposed test statistics are the only ones that actually reach the nominal value for the first-type error. The distribution of the first-type errors that have been obtained from the estimations provided by *edgeR* and *DSS* crosses the nominal values only with the upper whisker, and *DESeq* distribution does not cross the nominal value at all. From a first observation of these plots it could seem that the variability of the first-type errors gets larger as n_j increases, thus indicating some possible problems in the stability of the estimates provided by the EM algorithm. Actually, as it has been shown in the simulation experiment presented above, this problem does not occur. We must be careful in interpreting these plots, because comparing the variability of the first-type errors just looking directly at the inter-quartile range (that is the difference between the third quartile Q_3 and the first quartile Q_1) could lead to mistakes. We have to analyze an appropriate measure of the variability that takes into account also of the differences in the orders of magnitude, and therefore we can check that actually the variability of the first type errors decreases as n_j increases, as expected. We refer to Appendix A.2.1 for more details, and in particular to Table A.1.

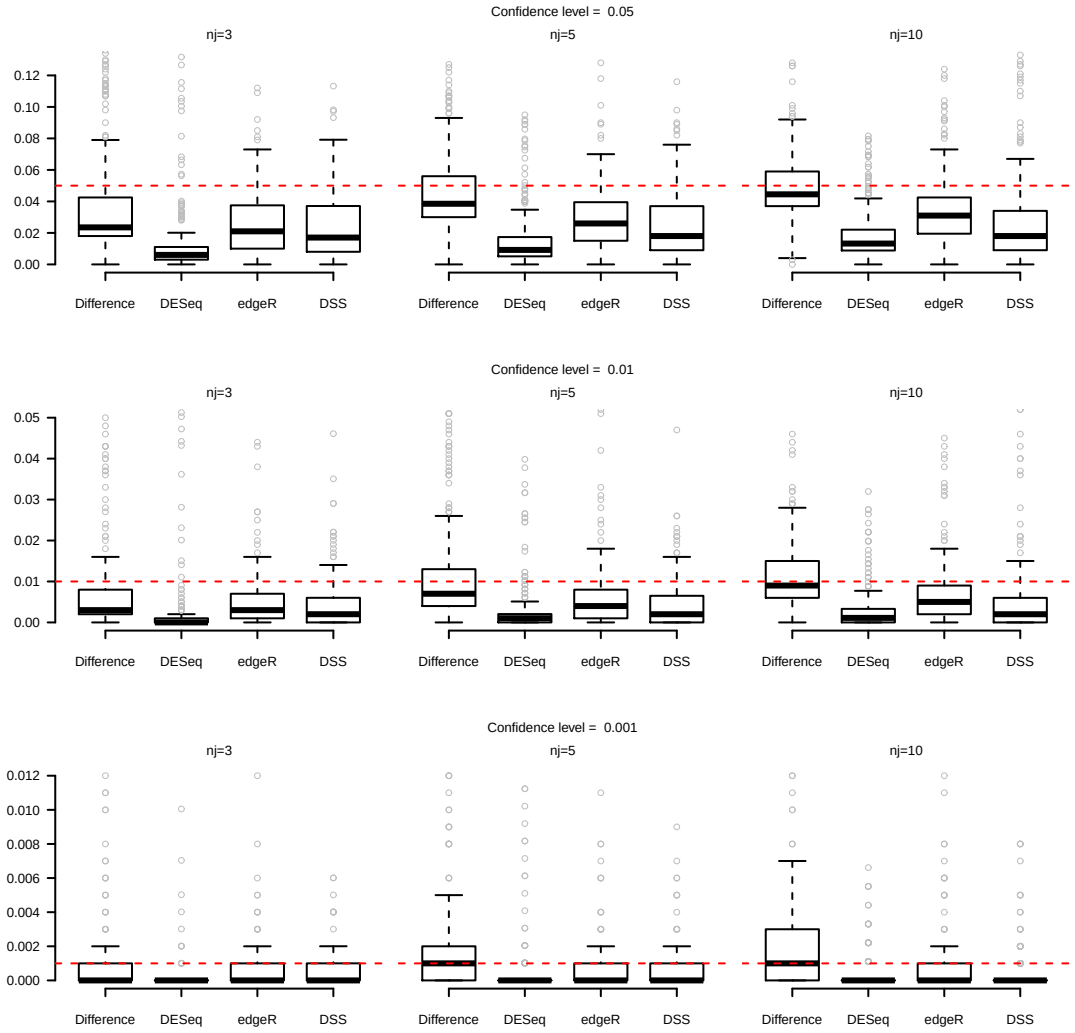

Figure 5.8: Box-plots of the distribution of the first-type errors computed on the null genes. Comparison between the performances of the proposed "Difference" test statistic, DESeq, edgeR and DSS as n_j varies. The dashed line indicates the nominal value that has been considered.

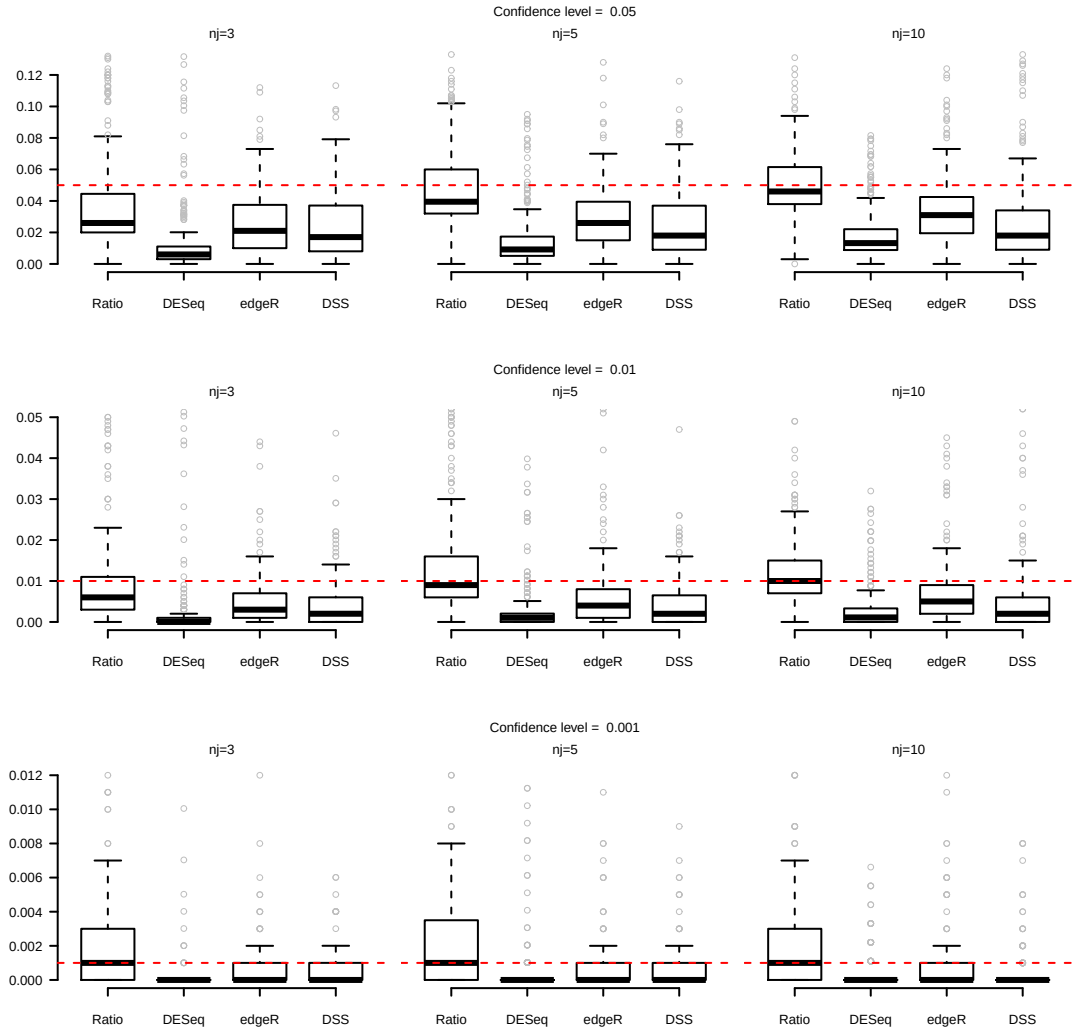

Figure 5.9: Box-plots of the distribution of the first-type errors computed on the null genes. Comparison between the performances of the proposed "Ratio" test statistic, DESeq, edgeR and DSS as n_j varies. The dashed line indicates the nominal value that has been considered.

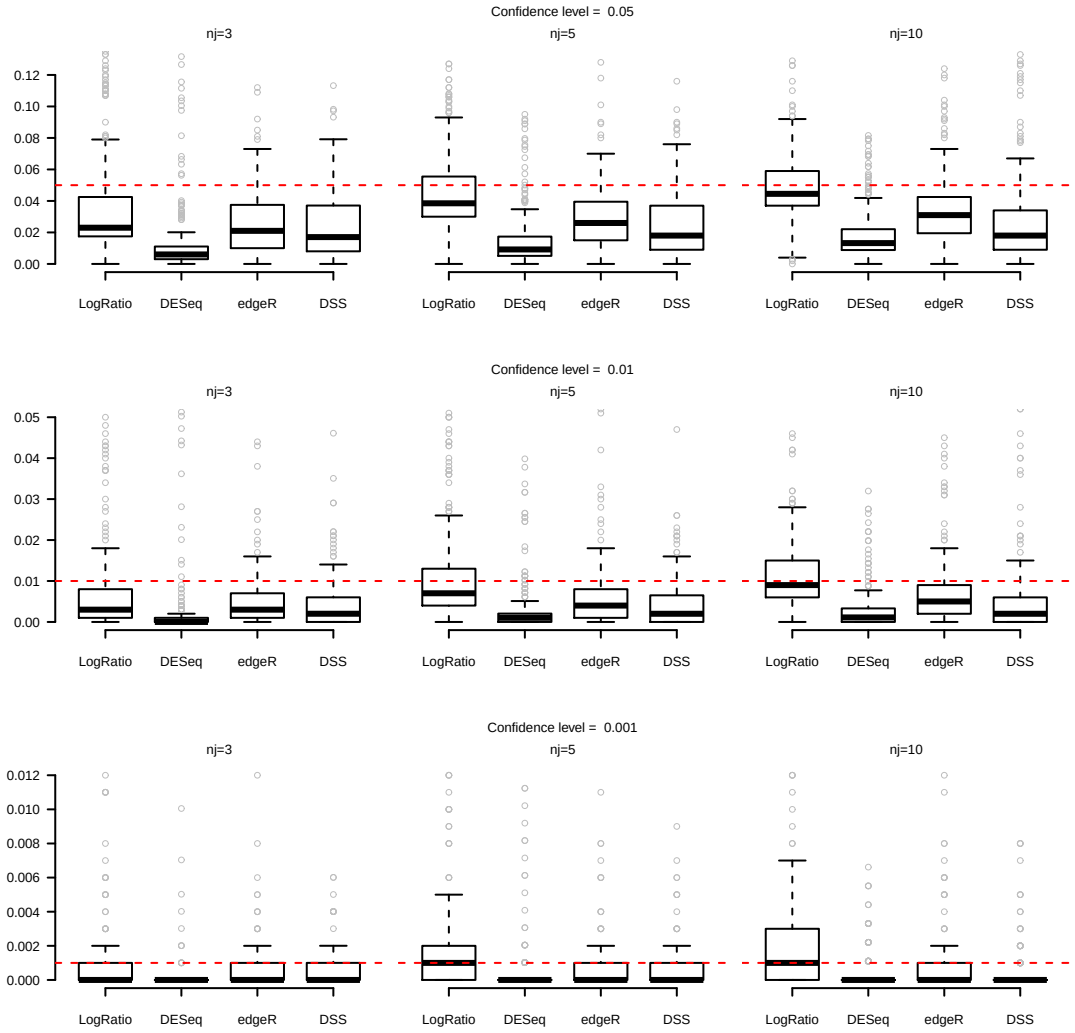

Figure 5.10: Box-plots of the distribution of the first-type errors computed on the null genes. Comparison between the performances of the proposed "Log Ratio" test statistic, DESeq, edgeR and DSS as n_j varies. The dashed line indicates the nominal value that has been considered.

The average values for the first type errors are reported in Table 5.2 together with the correspondent standard errors. The median values are shown in the Appendix A.2.1, Table A.2. It is important to underline that the distribution of the errors is skewed, and therefore the average values should be analyzed together with the quartiles information.

Table 5.2: Simulation B: means and standard errors for first-type errors, at different confidence levels.

Statistic	$n_j = 3$	$n_j = 5$	$n_j = 10$
Confidence level= 0.05			
Difference	0.0392 (0.0356)	0.0483 (0.0273)	0.0505 (0.0213)
Ratio	0.0418 (0.0351)	0.0501 (0.0267)	0.0516 (0.0211)
Log Ratio	0.0395 (0.0366)	0.0485 (0.0278)	0.0506 (0.0217)
DESeq	0.0143 (0.0242)	0.0172 (0.0206)	0.0201 (0.0187)
edgeR	0.0337 (0.0454)	0.0333 (0.0335)	0.0346 (0.0229)
DSS	0.0380 (0.0624)	0.0352 (0.0499)	0.0293 (0.0318)
Confidence level= 0.01			
Difference	0.0107 (0.0179)	0.0121 (0.0134)	0.0119 (0.0098)
Ratio	0.0135 (0.0197)	0.0146 (0.0142)	0.0131 (0.0104)
Log Ratio	0.0110 (0.0190)	0.0123 (0.0138)	0.0120 (0.0100)
DESeq	0.0036 (0.0111)	0.0034 (0.0072)	0.0037 (0.0061)
edgeR	0.0102 (0.0252)	0.0085 (0.0155)	0.0074 (0.0085)
DSS	0.0128 (0.0382)	0.0102 (0.0260)	0.0066 (0.0125)
Confidence level= 0.001			
Difference	0.0031 (0.0086)	0.0025 (0.0047)	0.0021 (0.0032)
Ratio	0.0045 (0.0105)	0.0037 (0.0063)	0.0026 (0.0039)
Log Ratio	0.0033 (0.0092)	0.0027 (0.0051)	0.0021 (0.0034)
DESeq	0.0012 (0.0053)	0.0007 (0.0023)	0.0005 (0.0012)
edgeR	0.0032 (0.0126)	0.0018 (0.0058)	0.0012 (0.0024)
DSS	0.0048 (0.0211)	0.0032 (0.0117)	0.0013 (0.0038)

The capability of controlling the first-type error can be checked also looking at the empirical cumulative distribution function (ECDF) of the null p-values; the more their distribution is close to the diagonal, the more they can be considered as actually uniformly distributed, as requested by the probability integral transform theorem.

In Figure 5.11 the ECDFs for the null p-values obtained through the proposed test statistics, *DESeq*, *edgeR* and *DSS* as n_j varies are shown. It is clear that the proposed test statistics behave better than the others already in correspondence of $n_j = 3$, then the correspondent ECDFs become closers and closers to the diagonal as the number of replicates increases and for $n_j = 10$ the ECDFs for the null p-values of the proposed procedures even overlap the diagonal, whereas *edgeR*,

DESeq and *DSS* reveal curves that lie behind the diagonal for all the three scenarios.

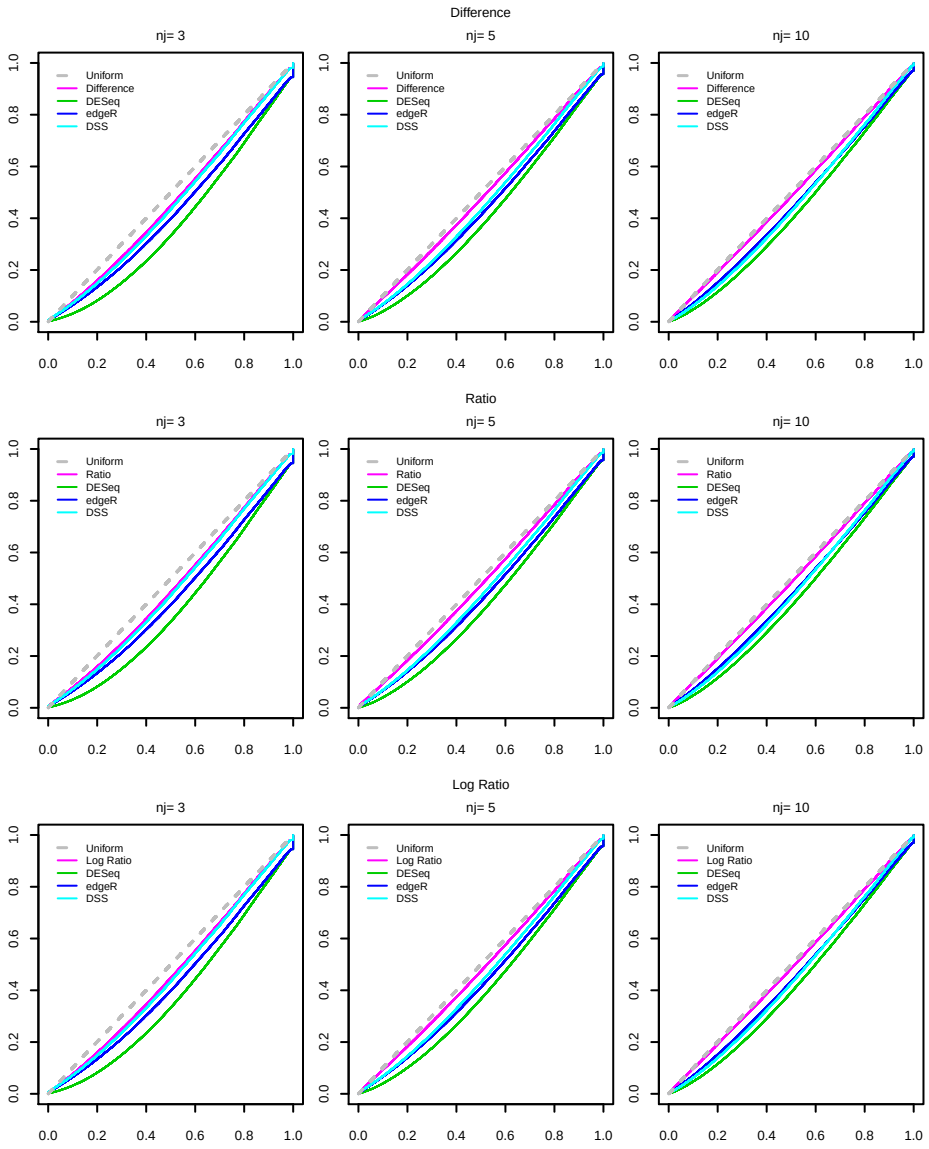

Figure 5.11: Empirical cumulative distribution functions for the null p-values that are obtained by the Log Ratio test statistic, DESeq, edgeR and DSS. The dashed line indicates the ECDF of the uniform distribution, that is the target one.

Finally, it could be interesting to look at how the capability of controlling the first type error is influenced by the real dispersion parameter from which each gene has been generated. Figures 5.12 and 5.13 show for each of the 200 null-genes the empirical first-type errors (at a confidence level of 0.05) in correspondence of

the true levels of dispersion α_i and for different numbers of replicates, for all the compared methods (the proposed test statistics, based on the mixture of NB are presented in Figure 5.12, and the others in Figure 5.13). For the other confidence levels we refer to Appendix A.2.1, Figures A.1, A.2, A.3 and A.4. These plots highlight that, as expected, controlling the first-type error is more difficult for the genes that have been generated with a lower α_i, that is with a greater variance, and this behavior holds for all the method. These figures also confirm the already explained results about the effective capability of reaching the nominal level (indicated by the red dashed line) as n_j increases, that is valid only for the three proposed test statistics.

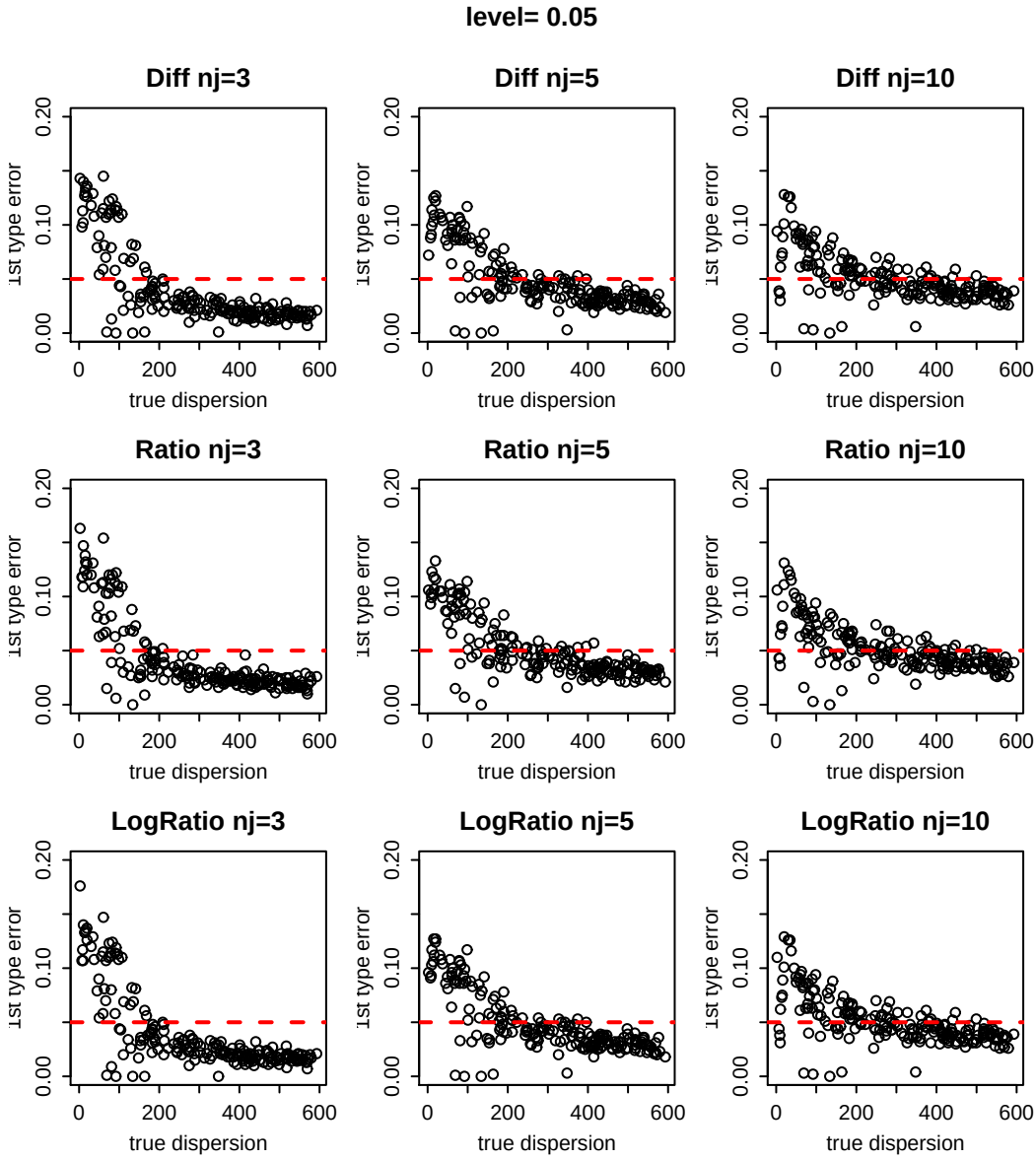

Figure 5.12: Empirical first-type errors at a confidence level of 0.05 as a function of the true dispersion parameters α_i, computed on the 200 null genes for the proposed test statistics, as n_j varies. Red dashed lines indicate the nominal levels.

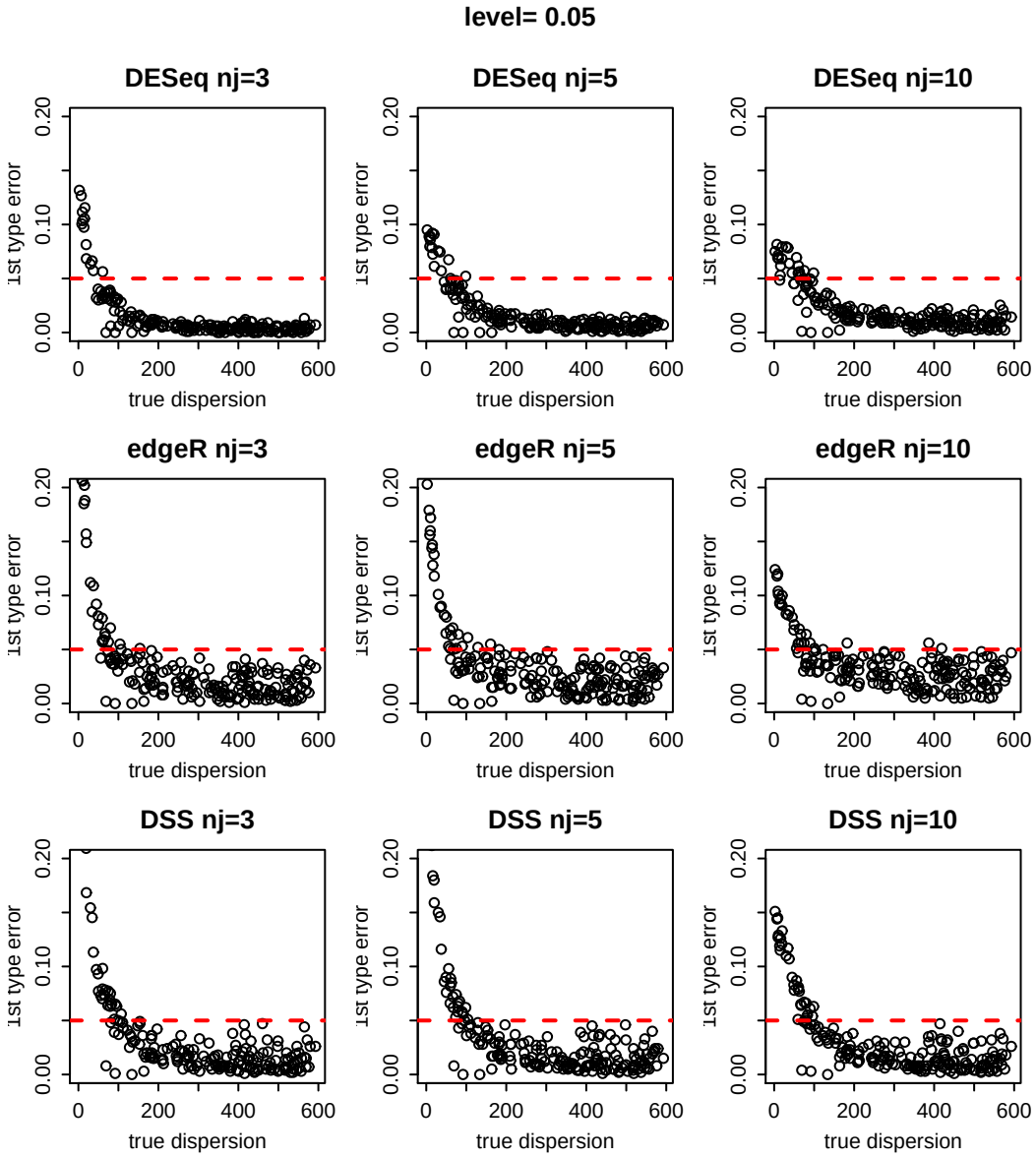

Figure 5.13: Empirical first-type errors at a confidence level of 0.05 as a function of the true dispersion parameters α_i, computed on the 200 null genes for the already-known strategies, as n_j varies. Red dashed lines indicate the nominal levels.

5.2.3 The second-type error

After that we have checked the reliability of the proposed strategy in terms of controlling the level of the tests, we have computed the second-type errors for each of the 100 differentially expressed genes across the 1000 data-sets, again for the

three proposed test statistics and for the other already known methods. The lower is the second-type error, the more the method is able to detect the differentially expressed genes. In Table 5.3 are reported the average second-type errors together with the correspondent standard errors. See the appendix A.2.2 for the boxplots of the distributions of the second-type errors, and for the median values.

Table 5.3: Simulation B: means and standard errors for second-type errors, at different confidence levels.

Statistic	$n_j = 3$	$n_j = 5$	$n_j = 10$
Confidence level= 0.05			
Difference	0.1582 (0.2738)	0.1002 (0.2267)	0.0543 (0.1455)
Ratio	0.2112 (0.3259)	0.1304 (0.2812)	0.0764 (0.2046)
Log Ratio	0.1569 (0.2726)	0.0991 (0.2246)	0.0534 (0.1443)
DESeq	0.1987 (0.3007)	0.1196 (0.2568)	0.0642 (0.1809)
edgeR	0.1444 (0.2526)	0.0945 (0.2197)	0.0529 (0.1533)
DSS	0.1354 (0.2449)	0.0892 (0.2109)	0.0513 (0.1526)
Confidence level= 0.01			
Difference	0.2341 (0.3289)	0.1442 (0.2867)	0.0874 (0.2199)
Ratio	0.3336 (0.3874)	0.1897 (0.3334)	0.1146 (0.2775)
Log Ratio	0.2331 (0.3278)	0.1430 (0.2845)	0.0856 (0.2167)
DESeq	0.3141 (0.3472)	0.1755 (0.3102)	0.0980 (0.2462)
edgeR	0.2268 (0.2997)	0.1384 (0.2740)	0.0815 (0.2170)
DSS	0.2159 (0.3014)	0.1357 (0.2710)	0.0813 (0.2181)
Confidence level= 0.001			
Difference	0.3441 (0.3703)	0.2037 (0.3345)	0.1228 (0.2834)
Ratio	0.5075 (0.3996)	0.2889 (0.3847)	0.1545 (0.3260)
Log Ratio	0.3433 (0.3693)	0.2026 (0.3333)	0.1212 (0.2799)
DESeq	0.4873 (0.3635)	0.2620 (0.3572)	0.1382 (0.3016)
edgeR	0.3609 (0.3359)	0.2066 (0.3193)	0.1166 (0.2753)
DSS	0.3508 (0.3471)	0.2061 (0.3230)	0.1176 (0.2758)

As it has already been noticed for the first-type errors, we have to recall that the distribution of the errors is very skewed and thus the average information is not completely explicative of the behavior of each test. From Table 5.3 we can see that the proposed "Difference" and "Log Ratio" test statistics are competitive with the *edgeR* and *DSS* ones, and looking at this information together with observing

the quartiles we can conclude that the proposed strategy is the most powerful. The "Ratio" one performs worse, even if it is anyway preferable than *DESeq* in terms of detecting DE genes.
This is a very good result because it means that the estimation of the parameters through the Mixture of Negative Binomial distributions together with the proposed test statistics (especially the "Difference" and the "Log Ratio" ones) are the only ones that are actually able to control the first-type error and they are the best ones also in terms of power. We can conclude taht they are the most reliable procedures for differential analysis.

Chapter 6

Application to Prostate Cancer Data

We have analyzed RNA-Seq data on prostate cancer cells collected in correspondence of two different conditions: a group of patients has been treated with androgens, and the second one with an inactive compound. The data have been sequenced and analyzed by Li et al. [2008]. It is well known that androgen hormones stimulate some genes, and they also have a positive effect in curing prostate cancer cells. Therefore the connection between these stimulated genes and the survival of these cells is a largely studied issue. Seven biological replicates of prostate cancer cells (three for the androgen-treated condition and four for the control-group) for 37435 genes have been sequenced using the Illumina 1G Genome Analyzer. Then they have been mapped to the NCBI36 build of the human genome using Bowtie (allowing up to two mismatches) and then the number of reads that corresponded to each Ensembl gene (version 53) was counted. The resulting read count table is available from `https://sites.google.com/site/davismcc/useful-documents`. For the analysis we have considered the $p = 16424$ genes with mean count greater than 1, because they provide sufficient statistical information on the differential analysis.

6.1 Normalization and explorative analyses

Normalization procedures

In order to account for the bias introduced by the different lanes of the experiment and the eventual effect the gene length, we preliminarily normalized the data using quantile-based normalization scheme implemented in the R package *EDASeq* [Risso et al., 2011].

As it has been described in Chapter 2, the total number of reads could depend on the sequencing depth and it could vary between lanes or samples, or it could be influenced by the gene length, thus affecting the differential analysis procedure. Therefore data have to be preliminarily normalized before performing the analysis (see, for instance, Tarazona et al. [2011] and Dillies et al. [2013]). Bullard et al. [2010] evaluated a variety of normalization procedures in order to detect the sensitivity of the differential expression detection. We used the quantile-based normalization scheme implemented in the R package *EDASeq* that consists in matching counts in term of quantiles, scaling them within and between lanes by a center moment, such as their median, so that to remove the bias due to the gene length and the different lanes. In Figure 6.1 the boxplots of the read counts distribution for each lane before and after normalization are shown. They give evidence how the distribution of the read counts strongly depends on the lane and normalization is necessary in order to make samples sequenced by different lanes comparable.

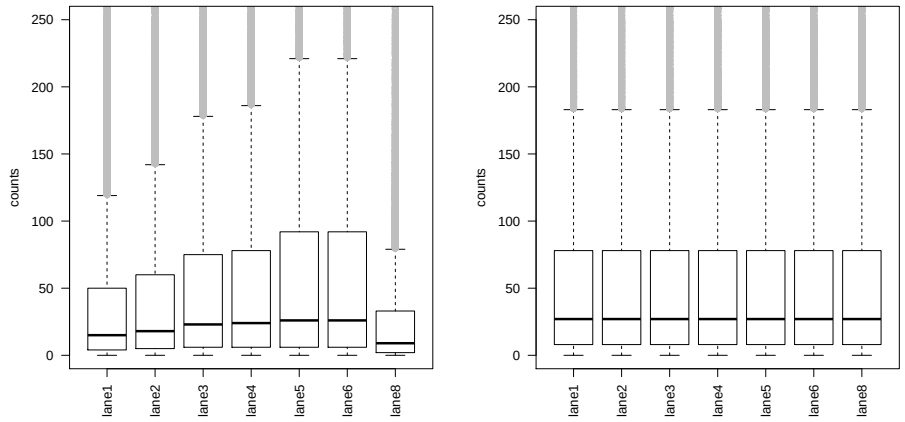

Figure 6.1: In the first plot, the boxplots of the counts per lane without the between-lane normalization. In the second plot, the boxplots of the counts per lane with the normalization.

Moreover, the correlation between the genome region lengths and the read counts was significant (0.1193 with p-value $<$ 2.2e-16) on the raw data; after the normalization procedure, we obtained a residual correlation of 0.0076 (p-value=0.333).

Explorative analyses

After that the normalization procedures have been applied to the data, we can consider the samples sequenced through different lanes comparable, and we can compound the information provided by different genes for estimating the variances. First of all we can have a look at the first rows of the dataset:

Genes	Control group				Treatment group		
	lane1	lane2	lane3	lane4	lane5	lane6	lane8
ENSG00000124208	766	934	698	782	392	651	560
ENSG00000182463	19	12	13	12	20	23	26
ENSG00000124201	192	205	223	203	215	167	130
$\vdots$							

The data matrix **y** is constituted by $p = 16424$ rows (one for each gene) and $n = 7$ columns (one for each sample). The first four columns contain the counts that have been got for people to whom the inactive compound was given, and the last three columns regard the androgen-treated group of patients.

In Figure 6.2 the empirical relationship between the mean and the variance for all the genes is plotted; the black line describes the theoretical relationship that would be observed under a Poisson model; the red line indicates the locally weighted scatter plot smooth (lowess) fit. It is clear from this graph that the Poisson assumption would not be adequate for these data.

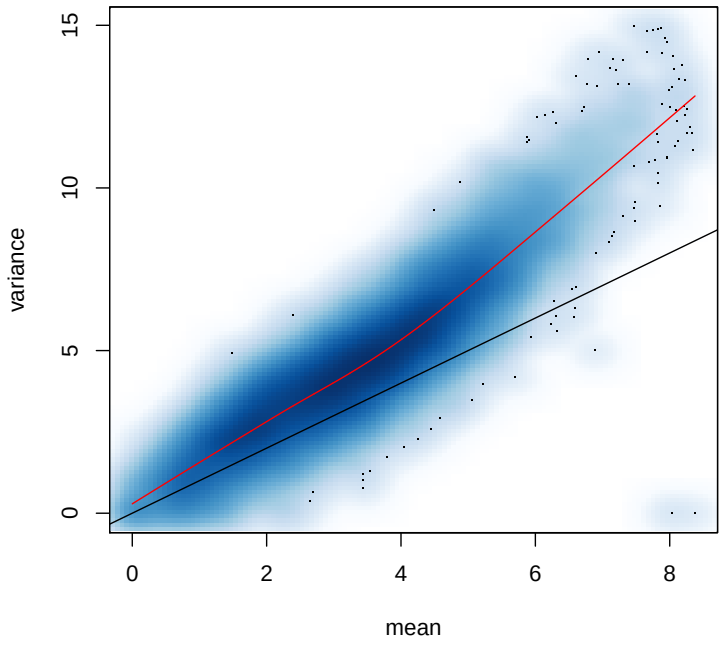

Figure 6.2: Relationship between mean and variance, computed on the logarithm of the counts.

6.2 Analysis and Results

The proposed NB mixture model has been fitted on the data with a number of components K ranging from 1 to 6. As it is reported in Figure 6.3, the BIC and AIC criteria suggested $K = 3$ components while, as expected, the ICL-BIC suggested that even just $K = 2$ components would well fit the data. Indeed, the latter information criterion usually supports simpler models, giving a greater weight to the number of parameters. We have chosen to consider $K = 3$ components. Convergence has been obtained with 50 iterations of the EM-algorithm at the log-likelihood of -384886 (BIC= 1088660, AIC= 835479, ICL-BIC= 1107665).

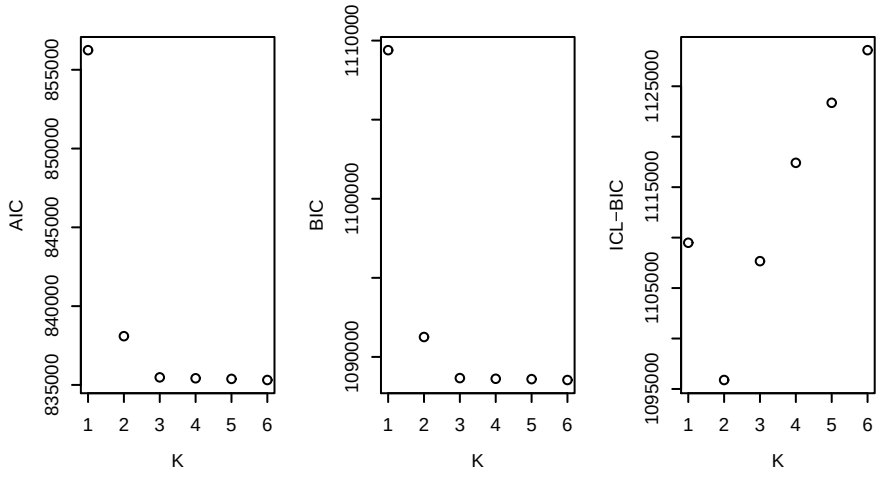

Figure 6.3: Information criteria as K varies.

Differential expression analysis has been conducted by computing the three proposed test statistics. For comparative purposes we have performed differential analysis also using the *DESeq*, *edgeR* and *DSS* methods implemented in `R` using the default settings.

All the obtained p-values have been adjusted following the procedure of Benjamini and Hochberg [1995] in order keep under control the total first error in multiple comparison testing. Indeed, even if basically we have to perform p statistical testings, one for each gene under investigation, in practice it could be more prudent applying some adjustments in order to get a global first-type error that approximates the level of each test.

In Table 6.1 the number of genes declared differentially expressed (DE) by each method at the confidence levels of 0.05, 0.01 and 0.001 is shown. The different methods detect a proportion of DE genes ranging from about 10% to 25%.

Table 6.1: Number of genes declared DE for all the compared methods at different confidence levels (adjusted p- values)

Statistic	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
Difference	3167	2146	1360
Ratio	3538	2591	1914
Log Ratio	4254	2941	2024
DESeq	2695	1828	1271
edgeR	3918	2774	1886
DSS	4215	2737	1737

In order to investigate the degree of accordance between two methods, we measured the proportion between the number of genes declared DE jointly by both methods and the average number of the genes declared DE marginally at a certain confidence level.

The first panel of Figure 6.4 shows the pairwise comparison between the proposed "Difference" test statistic and the *DESeq*, *edgeR* and *DSS* methods. The other two pictures of Figure 6.4 show the same results for the "Ratio" and "Log Ratio" test statistics respectively.

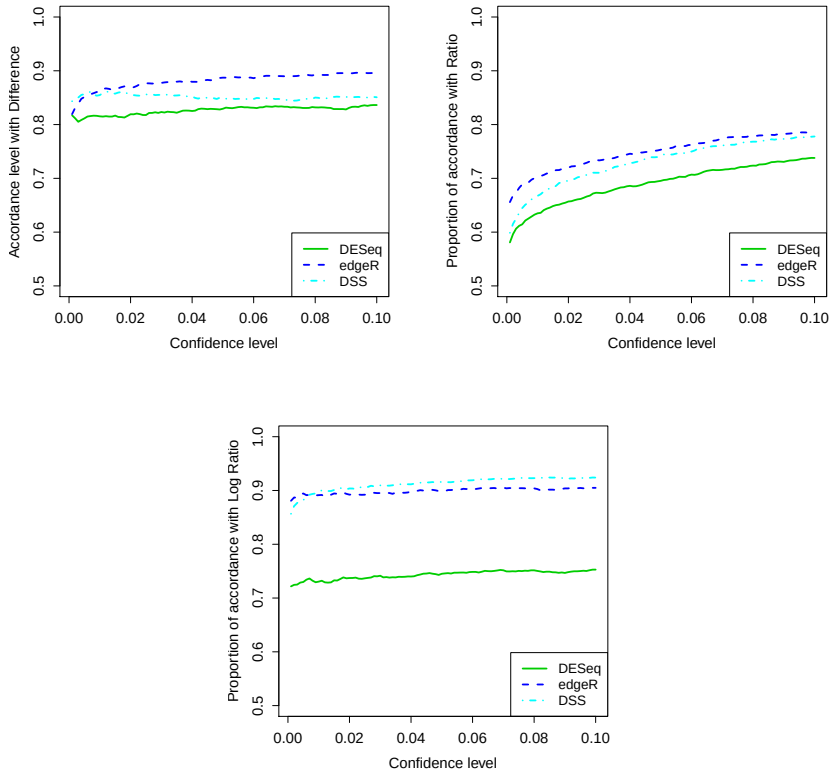

Figure 6.4: Proportion of genes declared DE as the confidence level increases for the different methods where the proposed "Difference" (panel a), "Ratio" (panel b) and "Log Ratio" test statistic (panel c) are taken as baseline.

It is clear from these graphs that the proposed test statistics provide results that are strongly consistent with the ones obtained by *edgeR* and *DSS*, with a degree of accordance of about 90% when the "Difference" or the "Log Ratio" is used. The set of DE genes detected by *DESeq* seems to be slightly different by the ones selected by all the other methods, even if the accordance level is between 60% and

80%.

Moreover, since often people are interested in grading the strength of the difference in the expression levels, another way for comparing the results is shown in Figures 6.5, 6.6 and 6.7, where the accord between the rankings of the adjusted p-values between each proposed test statistic and the other considered methods are plotted. These graphs confirm the results that has been described above.

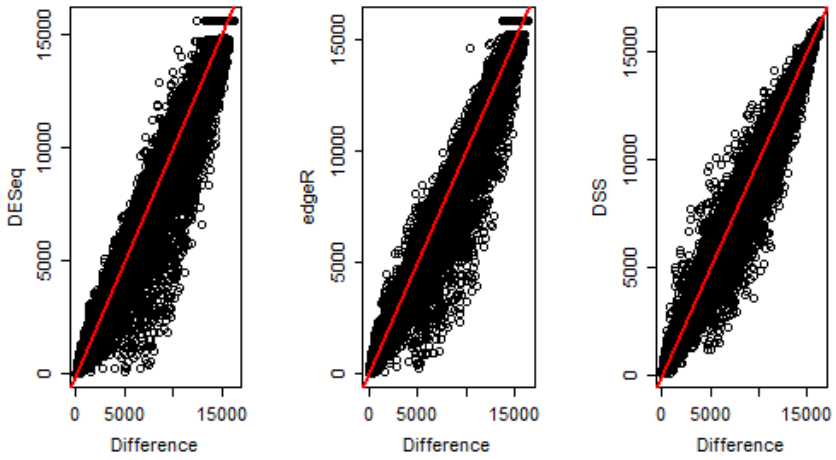

Figure 6.5: Rankings of the adjusted p-values - Comparison between the "Difference" test statistic and the other methods.

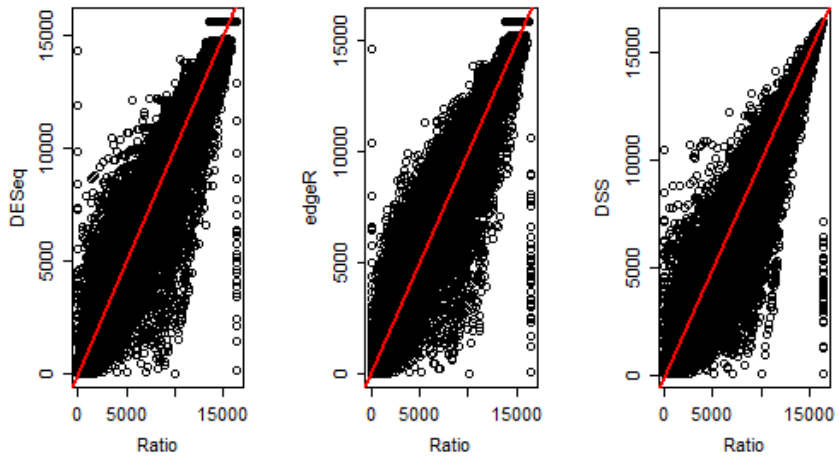

Figure 6.6: Rankings of the adjusted p-values - Comparison between the "Ratio" test statistic and the other methods.

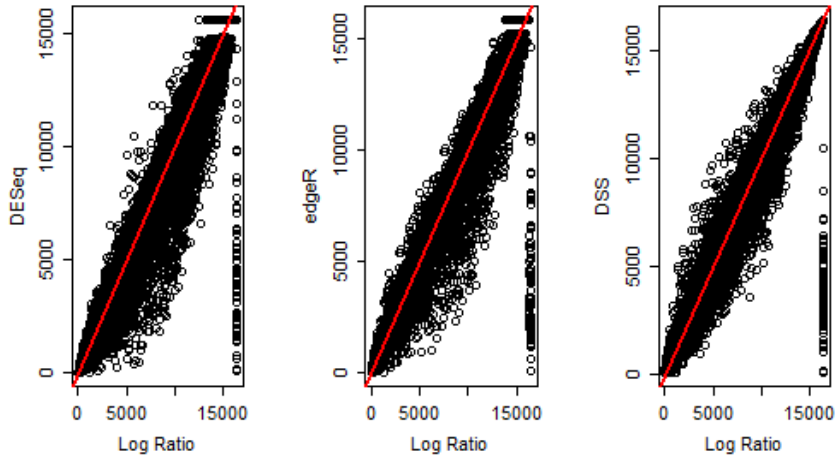

Figure 6.7: Rankings of the adjusted p-values - Comparison between the "Log Ratio" test statistic and the other methods.

Chapter 7

Concluding remarks

Massive parallel sequencing has deeply changed our understanding of gene expression thanks to a higher resolution, and this revolution has implied also the need of new statistical methods for detecting genes that are differentially expressed in correspondence of diverse biological conditions. This procedure is called *differential expression analysis*, and it is notable because it could reveal connections between the expression of specific genes and, for example, the presence of some pathologies or the effect of some treatments.

In this thesis we have focused on RNA-Seq data, that provide information about the sequencing of the transcriptome. As it has been described in Chapter 2, analyzing these data is a complex and delicate issue because they are discrete measurements, they are characterized by overdispersion and they have a hierarchical structure that requires some attention. Moreover, usually few information are available for each gene in each condition, because of costs in producing these data.

The major difficulty regards the estimation of the variances, and in Chapter 4 we have illustrated a new strategy for accounting for the heterogeneity of the dispersion parameter across genes. We have shown that the mixture of Negative Binomial distributions can be an appropriate new way for sharing information among genes about their dispersion levels, and to gain a more accurate estimation of the variances. We have also developed three test statistics, and the proposed approach is fully consistent in terms of parameter estimation and hypothesis testing. As a result, the simulation studies that have been illustrated in Chapter 5 have revealed that the first-type error of the proposed test statistics is controlled, and they are good also in containing the second-type ones. The comparative study we performed shows that the proposed strategy is competitive with existing methods. It also shows that some

popular testing procedures like *DEseq*, *edgeR* and *DSS* actually do not control the first-type error.

The considered methods have been employed also for the analysis of a real dataset on prostate cancer cells, aiming to compare the expression level of thousands of genes in an androgens-treated group of a patients and a control one. The results have been shown in Chapter 6, and they pointed out a high accordance level between the proposed test statistics (especially for the "Difference" and the "Log Ratio" ones) and the already known methods.

In this thesis we have focused on two sample comparison, but the procedure can indeed be adapted to any contrast, in an obvious manner, especially when using the "Difference" statistic. In a similar way, because our approach can be cast in the generalized linear model framework, normalization or correction for some exogenous effects could also be considered. Indeed, for example, since the RNA-seq replicates are often observed during specific time intervals, it would be interesting to relate the gene expression levels to time, thus allowing to measure the evolution of the expression of the genes at several time points in correspondence to different conditions. Finally, an extension of the model could be examined considering the imposition of an additional Poisson component (i.e. a NB distribution with dispersion parameter α that goes to infinity) in the mixture model. Indeed, since an iterative algorithm is used for the estimation of the dispersion parameter, the imposition of an upper bound to limit the searching interval is needed. The presence of a mixture component with $\alpha \to \infty$ could be a good way for overcoming this forced limit.
An `R` package implementing the EM algorithm and the proposed test statistics will be available soon.

Appendix A

Appendix

A.1 Appendix - The λ estimator

From the Maximization step that is described in Section 4.1.2, for the λ_{ij} estimators we have that $\widehat{\lambda_{ij}} = \frac{\sum_r y_{ijr}}{\sum_k f(\mathbf{z}_i|\mathbf{y}_i) \sum_r E(u_{ijr}|\mathbf{y}_i,\mathbf{z}_i)}$. It can be proved that the denominator $\sum_k f(\mathbf{z}_i|\mathbf{y}_i) \sum_r E(u_{ijr}|\mathbf{y}_i, \mathbf{z}_i)$ is simply equal to n_j:

Proof. We set:

$$\delta_{ij} = \sum_{k=1}^{K} f(\mathbf{z}_i|\mathbf{y}_i) \sum_{r=1}^{n_j} E(u_{ijr}|\mathbf{y}_i, \mathbf{z}_i)$$

and for the (4.14)

$$\delta_{ij} = \sum_{k=1}^{K} f(\mathbf{z}_i|\mathbf{y}_i) \sum_{r=1}^{n_j} \frac{y_{ijr} + \alpha_k}{\lambda_{ij} + \alpha_k}$$

We note that we are considering one specific gene i in the condition j, and given that the mixture structure involves the gene level, y_{ijr} with $r = 1, ..., n_j$ can be considered as independently distributed according to a negative binomial, with dispersion parameter depending on the group membership of the gene i to the k-th component of the mixture; therefore $f(\mathbf{z}_i|\mathbf{y}_i)$ is equal to 1 in correspondence to the $k - th$ group at which the gene i belongs, and 0 otherwise:

$$\delta_{ij} = \sum_{r=1}^{n_j} \frac{y_{ijr} + \alpha_k}{\lambda_{ij} + \alpha_k}$$

therefore:

$$\sum_{r=1}^{n_j} (y_{ijr} + \alpha_k) = \delta_{ij}(\lambda_{ij} + \alpha_k)$$

$$\sum_{r=1}^{n_j} y_{ijr} + n_j\alpha_k = \delta_{ij}\lambda_{ij} + \delta_j\alpha_k$$

but since $\sum_{r=1}^{n_j} y_{ijr} = \delta_j\lambda_{ij}$, $n_j\alpha_k$ must be equal to $\delta_{ij}\alpha_k$ and $n_j = \delta_{ij}$. □

And finally

$$\widehat{\lambda_{ij}} = \frac{\sum_r y_{ijr}}{n_j}. \tag{A.1}$$

A.2 Appendix - Simulation B

A.2.1 First-type error

An appropriate measure for studying the behavior of the variability of the first-type errors as n_j varies could be based on the quartile coefficient of dispersion (QCD), [Johnson, 2014] that is defined as $\frac{Q_3-Q_1}{Q_3+Q_1}$. This indicator is useful in evaluating the dispersion of the first type errors without being affected by dimensions, and the QCDs that we get in correspondence of each simulation scheme for the different confidence levels are reported in Table A.1. These results indicate that actually the variability of the first type errors decreases as n_j increases, as expected. Since the first-type errors are skewed (and especially the ones for the smaller confidence levels), for some strategies we have that both the first and the third quartiles are equal to 0 and therefore we get QCD= “NaN”.

Table A.1: Simulation B: quartile coefficients of dispersion for the empirical first-type errors (Section 5.2).

Statistic	$n_j = 3$	$n_j = 5$	$n_j = 10$
Confidence level= 0.05			
Difference	0.4025	0.3023	0.2292
Ratio	0.3774	0.3005	0.2343
Log Ratio	0.4083	0.2962	0.2292
DESeq	0.5714	0.5455	0.4286
edgeR	0.5767	0.4470	0.3629
DSS	0.6444	0.6087	0.5814
Confidence level= 0.01			
Difference	0.6000	0.5294	0.4286
Ratio	0.5714	0.4545	0.3636
Log Ratio	0.7778	0.5294	0.4286
DESeq	1.0000	1.0000	1.0000
edgeR	0.7500	0.7778	0.6364
DSS	1.0000	1.0000	1.0000
Confidence level= 0.001			
Difference	1.0000	1.0000	1.0000
Ratio	1.0000	1.0000	1.0000
Log Ratio	1.0000	1.0000	1.0000
DESeq	NaN	NaN	NaN
edgeR	1.0000	1.0000	1.0000
DSS	1.0000	1.0000	NaN

Tables A.2 and A.3 report the medians and standard errors of the first- and second-type errors computed in the simulation study B described in the Section 5.2.

Table A.2: Simulation B: medians and standard errors for first-type errors, at different confidence levels.

Statistic	$n_j = 3$	$n_j = 5$	$n_j = 10$
Confidence level= 0.05			
Difference	0.0235 (0.0356)	0.0385 (0.0273)	0.0445 (0.0213)
Ratio	0.0260 (0.0351)	0.0395 (0.0267)	0.0460 (0.0211)
Log Ratio	0.0230 (0.0366)	0.0385 (0.0278)	0.0445 (0.0217)
DESeq	0.0060 (0.0242)	0.0092 (0.0206)	0.0132 (0.0187)
edgeR	0.0210 (0.0454)	0.0260 (0.0335)	0.0310 (0.0229)
DSS	0.0170 (0.0624)	0.0180 (0.0499)	0.0180 (0.0318)
Confidence level= 0.01			
Difference	0.0030 (0.0179)	0.0070 (0.0134)	0.0090 (0.0098)
Ratio	0.0060 (0.0197)	0.0090 (0.0142)	0.0100 (0.0104)
Log Ratio	0.0030 (0.0190)	0.0070 (0.0138)	0.0090 (0.0100)
DESeq	0.0000 (0.0111)	0.0010 (0.0072)	0.0011 (0.0061)
edgeR	0.0030 (0.0252)	0.0040 (0.0155)	0.0050 (0.0085)
DSS	0.0020 (0.0382)	0.0020 (0.0260)	0.0020 (0.0125)
Confidence level= 0.001			
Difference	0.0000 (0.0086)	0.0010 (0.0047)	0.0010 (0.0032)
Ratio	0.0010 (0.0105)	0.0010 (0.0063)	0.0010 (0.0039)
Log Ratio	0.0000 (0.0092)	0.0010 (0.0051)	0.0010 (0.0034)
DESeq	0.0000 (0.0053)	0.0000 (0.0023)	0.0000 (0.0012)
edgeR	0.0000 (0.0126)	0.0000 (0.0058)	0.0000 (0.0024)
DSS	0.0000 (0.0211)	0.0000 (0.0117)	0.0000 (0.0038)

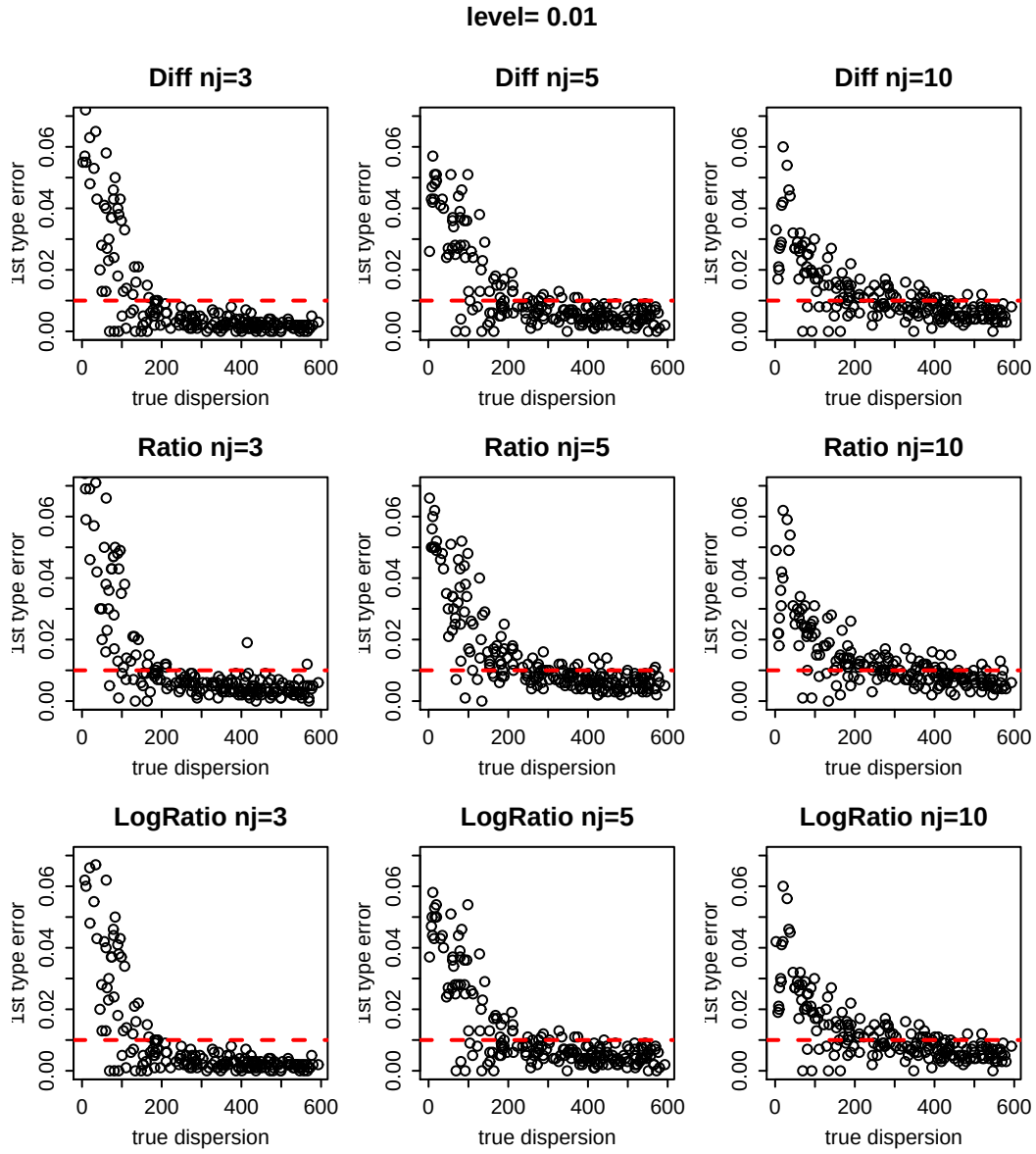

Figure A.1: Empirical first-type errors at a confidence level of 0.01 as a function of the true dispersion parameters α_i, computed on the 200 null genes for the proposed test statistics, as n_j varies. Red dashed lines indicate the nominal levels.

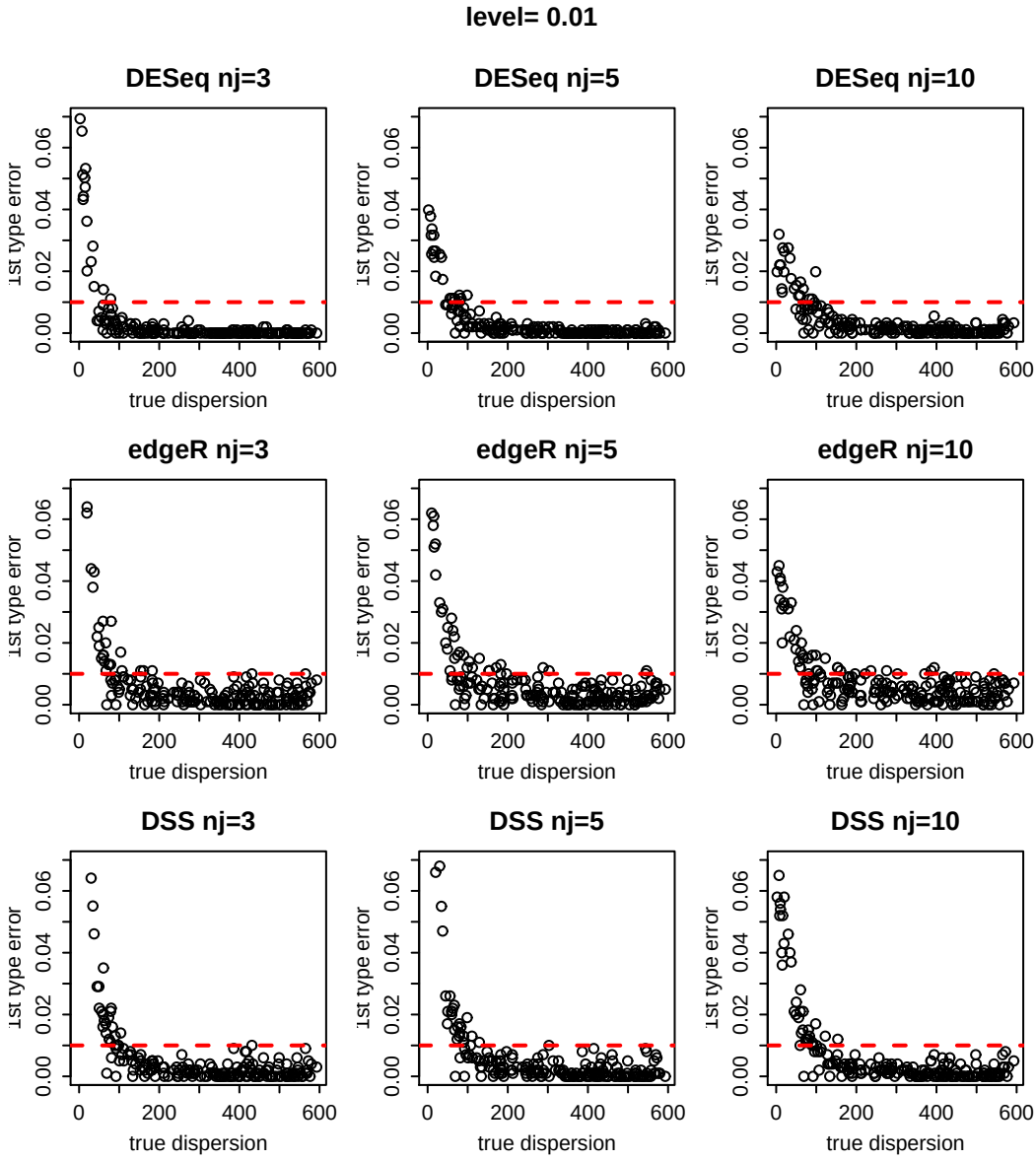

Figure A.2: Empirical first-type errors at a confidence level of 0.01 as a function of the true dispersion parameters α_i, computed on the 200 null genes for the already-known strategies, as n_j varies. Red dashed lines indicate the nominal levels.

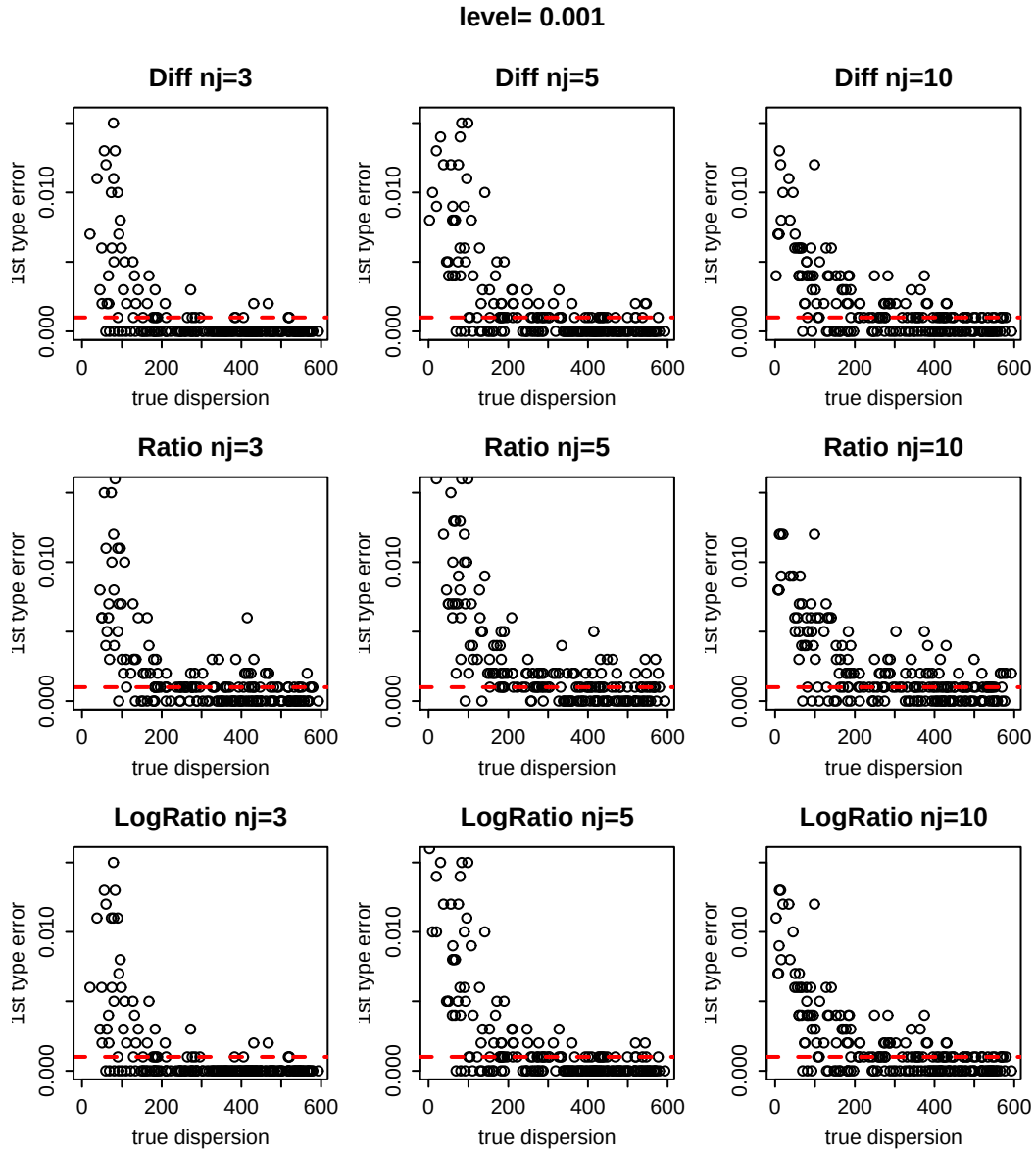

Figure A.3: Empirical first-type errors at a confidence level of 0.001 as a function of the true dispersion parameters α_i, computed on the 200 null genes for the proposed test statistics, as n_j varies. Red dashed lines indicate the nominal levels.

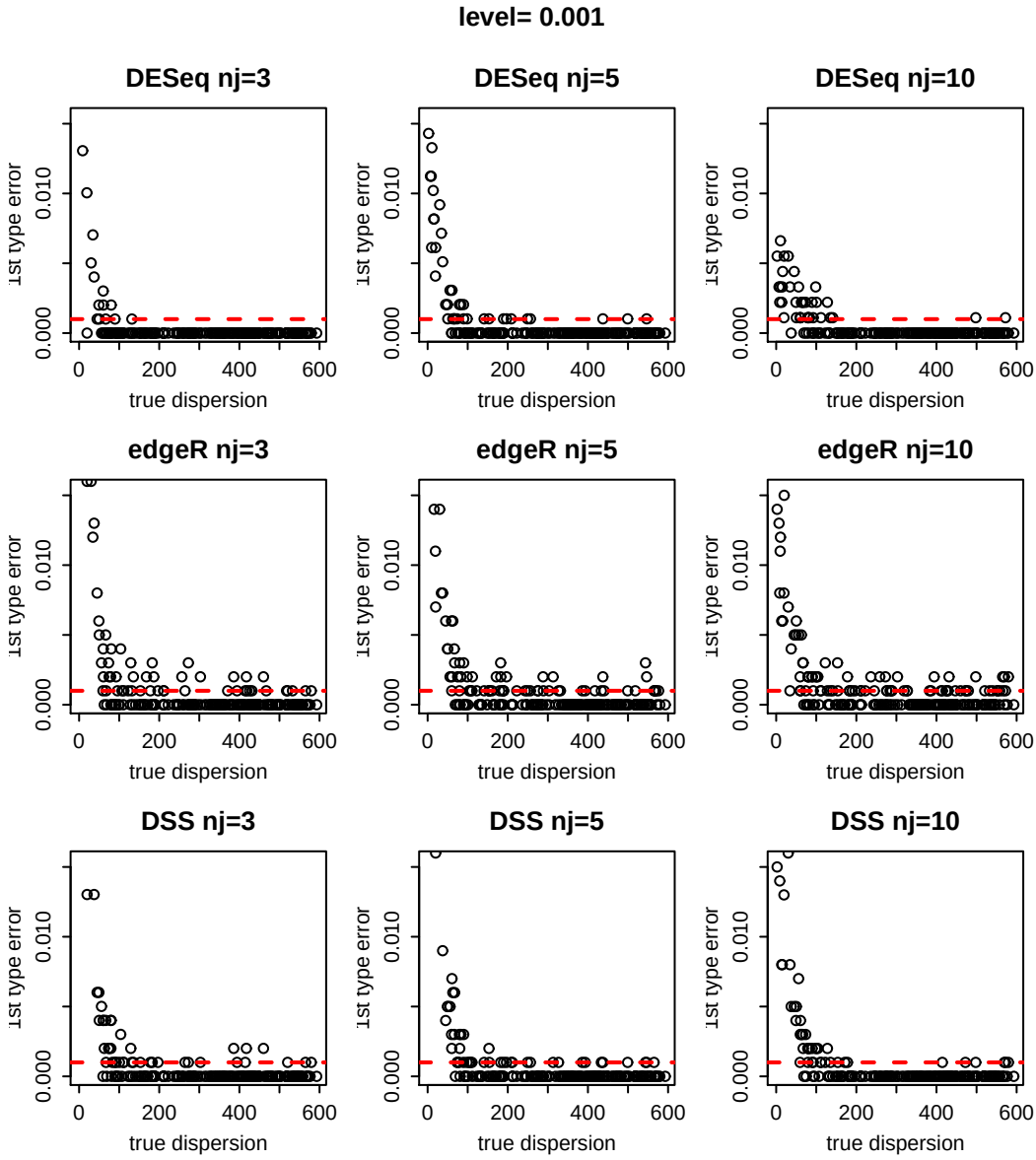

Figure A.4: Empirical first-type errors at a confidence level of 0.001 as a function of the true dispersion parameters α_i, computed on the 200 null genes for the already-known strategies, as n_j varies. Red dashed lines indicate the nominal levels.

A.2.2 Second-type error

Table A.3: Simulation B: medians and standard errors for second-type errors, at different confidence levels.

Statistic	$n_j = 3$	$n_j = 5$	$n_j = 10$
Confidence level= 0.05			
Difference	0.0140 (0.2738)	0.0000 (0.2267)	0.0000 (0.1455)
Ratio	0.0245 (0.3259)	0.0000 (0.2812)	0.0000 (0.2046)
Log Ratio	0.0140 (0.2726)	0.0000 (0.2246)	0.0000 (0.1443)
DESeq	0.0302 (0.3007)	0.0000 (0.2568)	0.0000 (0.1809)
edgeR	0.0160 (0.2526)	0.0000 (0.2197)	0.0000 (0.1533)
DSS	0.0120 (0.2449)	0.0000 (0.2109)	0.0000 (0.1526)
Confidence level= 0.01			
Difference	0.0530 (0.3289)	0.0020 (0.2867)	0.0000 (0.2199)
Ratio	0.1210 (0.3874)	0.0050 (0.3334)	0.0000 (0.2775)
Log Ratio	0.0530 (0.3278)	0.0010 (0.2845)	0.0000 (0.2167)
DESeq	0.1693 (0.3472)	0.0056 (0.3102)	0.0000 (0.2462)
edgeR	0.0865 (0.2997)	0.0030 (0.2740)	0.0000 (0.2170)
DSS	0.0581 (0.3014)	0.0020 (0.2710)	0.0000 (0.2181)
Confidence level= 0.001			
Difference	0.1460 (0.3703)	0.0135 (0.3345)	0.0000 (0.2834)
Ratio	0.4440 (0.3996)	0.0460 (0.3847)	0.0000 (0.3260)
Log Ratio	0.1445 (0.3693)	0.0120 (0.3333)	0.0000 (0.2799)
DESeq	0.4894 (0.3635)	0.0552 (0.3572)	0.0000 (0.3016)
edgeR	0.2810 (0.3359)	0.0315 (0.3193)	0.0000 (0.2753)
DSS	0.2580 (0.3471)	0.0225 (0.3230)	0.0000 (0.2758)

Figures A.5, A.6 and A.7 show the box-plots of the empirical second-type errors obtained by the “Difference”, “Ratio” and “Log Ratio” test statistic respectively and of the other considered approaches, in relation to the number of replicates ($n_j = 3, 5, 10$) and for the different levels of the test (0.05, 0.01 and 0.001).

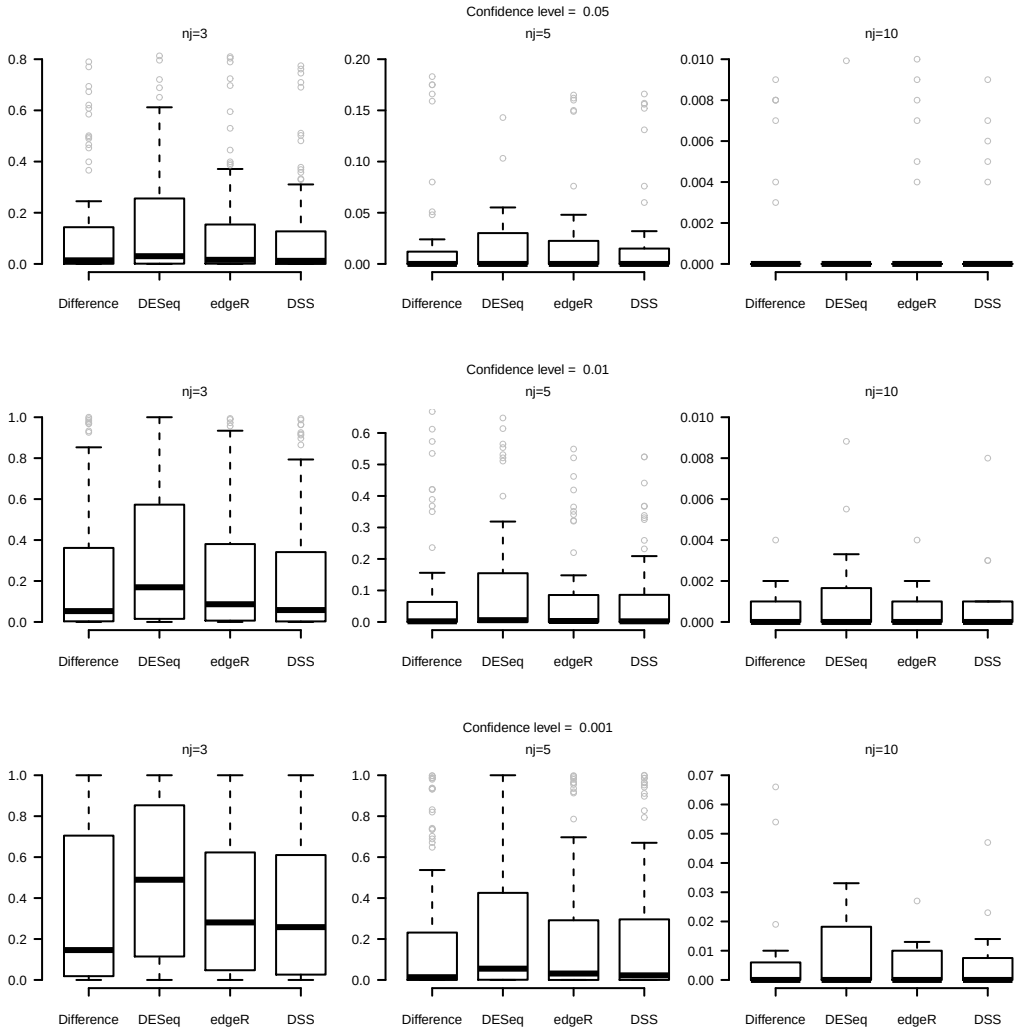

Figure A.5: Box-plots of the distribution of the second-type errors computed on the DE genes. Comparison between the performances of the proposed "Difference" test statistic, DESeq, edgeR and DSS as n_j varies.

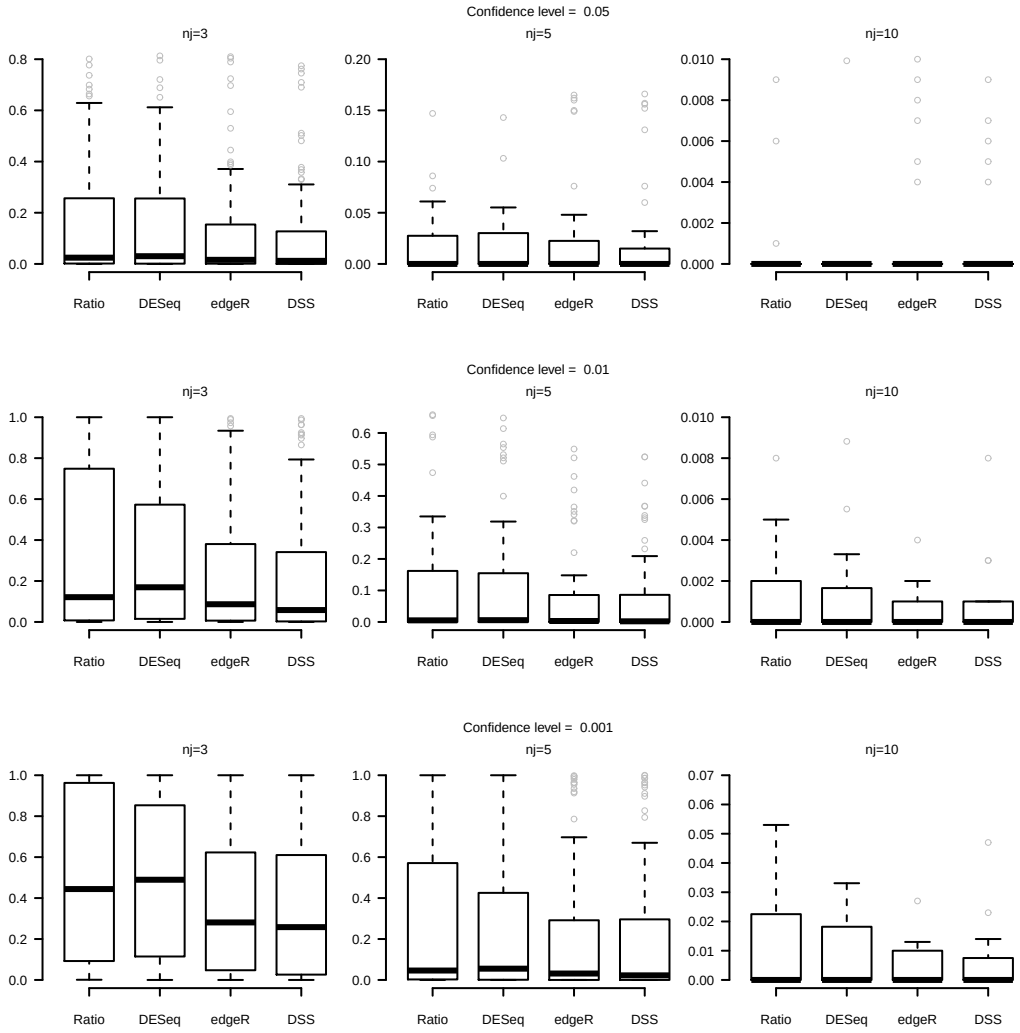

Figure A.6: Box-plots of the distribution of the second-type errors computed on the DE genes. Comparison between the performances of the proposed "Ratio" test statistic, DESeq, edgeR and DSS as n_j varies.

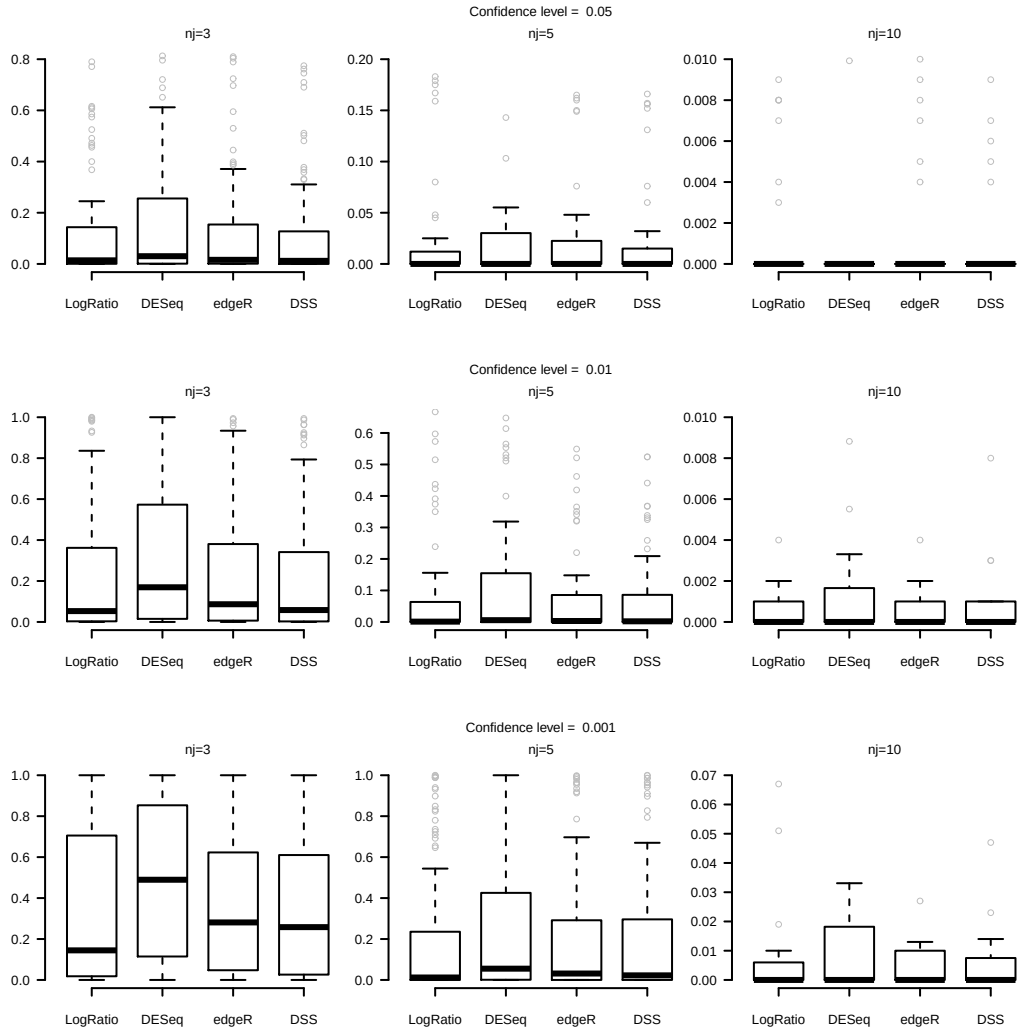

Figure A.7: Box-plots of the distribution of the second-type errors computed on the DE genes. Comparison between the performances of the proposed "Log Ratio" test statistic, DESeq, edgeR and DSS as n_j varies.

Bibliography

H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, pages 3099–3132, 2009.

S. Anders. Analysing RNA-Seq data with the DESeq package. *Molecular biology*, pages 1–17, 2010.

S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome biol*, 11(10):R106, 2010.

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

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(7):719–725, 2000.

J. S. Bloom, Z. Khan, L. Kruglyak, M. Singh, and A. A. Caudy. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC genomics*, 10(1):221, 2009.

D. Böhning. The EM algorithm with gradient function update for discrete mixtures with known (fixed) number of components. *Statistics and Computing*, 13(3): 257–265, 2003.

R.P. Brent. *Algorithms for Minimization without Derivatives, Chapter 4*. Prentice-Hall. Englewood Cliffs, NJ, 1973. ISBN 0-13-022335-2.

J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics*, 11(1):94, 2010.

R. H. Byrd, J. Nocedal, and R. B. Schnabel. Representations of quasi-newton matrices and their use in limited memory methods. *Mathematical Programming*, 63(1-3):129–156, 1994.

R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

A.C. Cameron and P.K. Trivedi. *Regression Analysis of Count Data*. Econometric Society Monographs. Cambridge University Press, 1998. ISBN 9780521635677.

Y. Chen, A. T.L. Lun, and G. K Smyth. Differential expression analysis of complex RNA-seq experiments using edgeR. 2014.

C. Cox. Fieller's theorem, the likelihood and the delta method. *Biometrics*, 46(3): pp. 709–718, 1990. ISSN 0006341X.

P. Delmar, S. Robin, T.L. Roux, J.J. Daudin, et al. Mixture model on the variance for the differential analysis of gene expression data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):31–50, 2005.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

M.A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, et al. A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics*, 14(6):671–683, 2013.

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

A. C. Frazee, S. Sabunciyan, K. D. Hansen, R. A. Irizarry, and J. T. Leek. Differential expression analysis of RNA-seq data at single-base resolution. *Biostatistics*, page kxt053, 2014.

M. Guindani, N. Sepúlveda, C. D. Paulino, and P. Müller. A bayesian semiparametric approach for the differential analysis of sequence counts data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(3):385–404, 2014.

T. Hardcastle and K. Kelly. BaySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11(422): 1–15, 2010.

J.M. Hilbe. *Negative Binomial Regression*. Cambridge University Press, 2011. ISBN 9781139500067.

N. Ismail and H. Zamani. Estimation of claim count data using negative binomial, generalized poisson, zero-inflated negative binomial and zero-inflated generalized poisson regression models. In *Casualty Actuarial Society E-Forum, Spring 2013*, 2013.

D. E. Johnson. 14 descriptive statistics. *Research Methods in Linguistics*, page 288, 2014.

V. M. Kvam, P. Liu, and Y. Si. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *American journal of botany*, 99(2):248–256, 2012.

H. Li, M. T. Lovci, Y. S. Kwon, M. G. Rosenfeld, X. D. Fu, and G. W. Yeo. Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. *Proceedings of the National Academy of Sciences*, 105(51):20179–20184, 2008.

J. Li, D. M. Witten, I. M. Johnstone, and R. Tibshirani. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, page kxr031, 2011.

J.C. Marioni, C.E. Mason, S.M. Mane, M. Stephens, and Y. Gilad. RNA-seq: An assessment of techincal reproducibility and comparison with gene expression arrays. *Genome Research*, 18:1509–1517, 2008.

G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Wiley, 1996. ISBN 9780471123583.

G. McLachlan and D. Peel. *Finite Mixture Models, Willey Series in Probability and Statistics*. John Wiley & Sons, New York, 2000.

D. Nityasuddhi and D. Böhning. Asymptotic properties of the EM algorithm estimate for normal mixture models with component specific variances. *Computational statistics & data analysis*, 41(3):591–601, 2003.

A. Oshlack, M. D. Robinson, M. D. Young, et al. From rna-seq reads to differential expression results. *Genome biol*, 11(12):220, 2010.

K. Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

D. Risso, K. Schwartz, G. Sherlock, and S. Dudoit. Gc-content normalization for rna-seq data. *BMC Bioinformatics*, 12(1):480, 2011.

M. D. Robinson and G. K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.

M. D. Robinson and G. K. Smyth. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2):321–332, 2008.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

S. Russell. The EM algorithm. 1998.

T. Sapatinas. Identifiability of mixtures of power-series distributions and related characterizations. *Annals of the Institute of Statistical Mathematics*, 47(3):447–459, 1995.

G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6 (2):461–464, 1978.

Y. Si and P. Liu. An optimal test with maximum average power while controlling FDR with application to RNA-seq data. *Biometrics*, 69(3):594–605, 2013.

G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1), 2004.

W.W. Soon, M. Hariharan, and M.P. Snyde. High-throughput sequencing for biology and medicine. *Molecular Systems Biology*, 9(640):1–14, 2013.

S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa. Differential expression in RNA-seq: a matter of depth. *Genome research*, 21(12):2213–2223, 2011.

A.W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000. ISBN 9780521784504.

A. Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54(3):pp. 426–482, 1943. ISSN 00029947.

L. Wang, Z. Feng, X. Wang, and X. Zhang. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, 26:136–138, 2010.

X. Wang. Approximating bayesian inference by weighted likelihood. *Canadian Journal of Statistics*, 34(2):279–298, 2006.

Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

L. Whitaker. On the poisson law of small numbers. *Biometrika*, 10:36–71, 1914.

D. M. Witten et al. Classification and clustering of sequencing data using a poisson model. *The Annals of Applied Statistics*, 5(4):2493–2518, 2011.

H. Wu, C. Wang, and Z. Wu. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, 14(2):232–243, 2013.

S. J. Yakowitz and J. D. Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, pages 209–214, 1968.

D. Yu, W. Huber, and O. Vitek. Shrinkage estimation of dispersion in negative binomial models for RNA-seq experiments with small sample size. *Bioinformatics*, 29(10):1275–1282, 2013.

Y. Zhou and F. A. Wright. BBSeq: A method to handle RNA-seq count data. 2011.