Mixtures of Negative Binomial distributions for modelling overdispersion in RNA-Seq data

Cinzia Viroli[1]
joint with E. Bonafede[1], S. Robin[2] & F. Picard[3]

[1] Department of Statistical Sciences, University of Bologna, Italy
[2] UMR 518 MIA, INRA/AgroParisTech, France,
[3] LBBE, University C. Bernard Lyon, France.

April, 3rd 2015

Statlearn 2015 - Grenoble

NGS technologies

The recent **Next Generation Sequencing (NGS)** technologies are becoming the mostly used tools to study gene expression.
RNA-Seq experiments: quantification of the transcriptome.

NGS technologies

The recent **Next Generation Sequencing (NGS)** technologies are becoming the mostly used tools to study gene expression.
RNA-Seq experiments: quantification of the transcriptome.

Before their advent, the expression level of a target genome was measured through *microarray technologies*;
NGS experiments: wider range of expression levels, cheaper and faster experiments.

NGS technologies

The recent **Next Generation Sequencing (NGS)** technologies are becoming the mostly used tools to study gene expression.
RNA-Seq experiments: quantification of the transcriptome.

Before their advent, the expression level of a target genome was measured through *microarray technologies*;
NGS experiments: wider range of expression levels, cheaper and faster experiments.

Microarray technologies: data are measured as fluorescence intensity $\rightarrow$ *continuous real data*;

NGS experiments: read counts assigned to a target genome $\rightarrow$ *discrete* measurements

Differential analysis

- **RNA-Seq**: a target gene or exon
- two (or more) biological conditions: disease states, treatments etc.
- comparison of the read counts of a genome region between the conditions

Data structure and notation

Y_{ijr} is the random variable that expresses the read counts mapped to:

- gene i (i=$1, \ldots, p$),
- in condition j ($j = 1, \ldots, d$; here $d = 2$ w.l.o.g),
- in sample r (r= $1, ..., n_j$),

Data structure and notation

Y_{ijr} is the random variable that expresses the read counts mapped to:

- gene i (i=$1, \ldots, p$),
- in condition j ($j = 1, \ldots, d$; here $d = 2$ w.l.o.g),
- in sample r (r= $1, ..., n_j$),
- p is large but n_j is small,

Data structure and notation

Y_{ijr} is the random variable that expresses the read counts mapped to:

- gene i (i=$1, \ldots, p$),
- in condition j ($j = 1, \ldots, d$; here $d = 2$ w.l.o.g),
- in sample r (r= $1, ..., n_j$),
- p is large but n_j is small,
- sometimes: excess of zeros,

Data structure and notation

Y_{ijr} is the random variable that expresses the read counts mapped to:

- gene i (i=$1, \dots, p$),
- in condition j ($j = 1, \dots, d$; here $d = 2$ w.l.o.g),
- in sample r (r= $1, ..., n_j$),
- p is large but n_j is small,
- sometimes: excess of zeros,
- 'overdispersion', i.e. the variance usually exceeds the mean.

Data structure and notation

Y_{ijr} is the random variable that expresses the read counts mapped to:

- gene i (i=$1, \ldots, p$),
- in condition j ($j = 1, \ldots, d$; here $d = 2$ w.l.o.g),
- in sample r (r= $1, ..., n_j$),
- p is large but n_j is small,
- sometimes: excess of zeros,
- 'overdispersion', i.e. the variance usually exceeds the mean.

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

1. first-level units: the replicates
2. second level: the conditions
3. third level: the 'genes'

Dealing with count data

Modeling nonnegative count data:

- *Poisson distribution*: the benchmark for count data, simple but imposes equidispersion (not adequate);

Dealing with count data

Modeling nonnegative count data:

Poisson distribution: the benchmark for count data, simple but imposes equidispersion (not adequate);

Zero-inflated Poisson (ZIP): often used, it allows to model over-dispersion due to excess of zeros, but it does not have an explicit parameter for variance;

Dealing with count data

Modeling nonnegative count data:

Poisson distribution: the benchmark for count data, simple but imposes equidispersion (not adequate);

Zero-inflated Poisson (ZIP): often used, it allows to model over-dispersion due to excess of zeros, but it does not have an explicit parameter for variance;

Negative binomial distribution (NB): two parameters, a mean and a dispersion parameter ($\rightarrow$ flexibility, overdispersion);

Dealing with count data

Modeling nonnegative count data:

Poisson distribution: the benchmark for count data, simple but imposes equidispersion (not adequate);

Zero-inflated Poisson (ZIP): often used, it allows to model over-dispersion due to excess of zeros, but it does not have an explicit parameter for variance;

Negative binomial distribution (NB): two parameters, a mean and a dispersion parameter ($\rightarrow$ flexibility, overdispersion);

Zero-inflated Negative binomial distribution (ZINB): empirical results proved that the difference in fit between ZINB and NB is usually trivial (*"Do we really need zero-inflated models?"* by P. Allison);

The NB distribution

$Y \sim \mathit{NegBin}(\lambda, \alpha)$

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

with: E(Y) = λ Var(Y) = $\lambda\left(1+\frac{1}{\alpha}\lambda\right)$

The NB distribution

$Y \sim NegBin(\lambda, \alpha)$

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

with: E(Y) = λ Var(Y) = $\lambda\left(1+\frac{1}{\alpha}\lambda\right)$

Two opposite strategies:

- a common dispersion parameter → not realistic

$$Y_{ijr} \sim NegBin(\lambda_{ij}, \alpha)$$

- p gene-specific dispersion parameters → estimation difficulties because of the limited number of replicates (p large, n_j small)

$$Y_{ijr} \sim NegBin(\lambda_{ij}, \alpha_i)$$

Estimating the dispersion parameters

Some solutions in the statistical literature that assume the NB probability model:

- **Robinson and Smyth (2007) - edgeR**: maximizes a weighted combination of the conditional log-likelihoods with per-gene dispersion and of the conditional log-likelihood with common dispersion;

Estimating the dispersion parameters

Some solutions in the statistical literature that assume the NB probability model:

- **Robinson and Smyth (2007) - edgeR**: maximizes a weighted combination of the conditional log-likelihoods with per-gene dispersion and of the conditional log-likelihood with common dispersion;
- **Anders and Huber (2010) - DESeq**: allows specifications of separate variances for genes and conditions and models the variances as smooth functions of the expected values through local regression;

Estimating the dispersion parameters

Some solutions in the statistical literature that assume the NB probability model:

- **Robinson and Smyth (2007) - edgeR**: maximizes a weighted combination of the conditional log-likelihoods with per-gene dispersion and of the conditional log-likelihood with common dispersion;
- **Anders and Huber (2010) - DESeq**: allows specifications of separate variances for genes and conditions and models the variances as smooth functions of the expected values through local regression;
- **Hardcastle and Kelly (2010) - baySeq**: same model as edgeR but it considers non-parametric priors on sets of parameters and it maximizes per-gene integrated quasi-likelihood (computational intensive)

Estimating the dispersion parameters

Some solutions in the statistical literature that assume the NB probability model:

- **Robinson and Smyth (2007) - edgeR**: maximizes a weighted combination of the conditional log-likelihoods with per-gene dispersion and of the conditional log-likelihood with common dispersion;
- **Anders and Huber (2010) - DESeq**: allows specifications of separate variances for genes and conditions and models the variances as smooth functions of the expected values through local regression;
- **Hardcastle and Kelly (2010) - baySeq**: same model as edgeR but it considers non-parametric priors on sets of parameters and it maximizes per-gene integrated quasi-likelihood (computational intensive)
- **Wu et al (2013) - DSS**: a shrinkage estimator imposing a log-normal prior on the dispersion parameters (Bayesian hierarchical model).

Estimating the dispersion parameters

Some solutions in the statistical literature that assume the NB probability model:

- **Robinson and Smyth (2007) - edgeR**: maximizes a weighted combination of the conditional log-likelihoods with per-gene dispersion and of the conditional log-likelihood with common dispersion;
- **Anders and Huber (2010) - DESeq**: allows specifications of separate variances for genes and conditions and models the variances as smooth functions of the expected values through local regression;
- **Hardcastle and Kelly (2010) - baySeq**: same model as edgeR but it considers non-parametric priors on sets of parameters and it maximizes per-gene integrated quasi-likelihood (computational intensive)
- **Wu et al (2013) - DSS**: a shrinkage estimator imposing a log-normal prior on the dispersion parameters (Bayesian hierarchical model).
- **Klambauer et al (2013) - DEXUS**: it assumes a mixture of *d* NBs for all the genes where the parameters are condition-specific, where each component is an (unkown) condition

Our proposal and outline

- Instead of fitting p NB models, we assume a mixture model with component-specific dispersion (and gene-specific means):
  - sharing information among genes that exhibit similar dispersion
  - an intermediate solution between the trade-off common vs gene-specific dispersion

Our proposal and outline

- Instead of fitting p NB models, we assume a mixture model with component-specific dispersion (and gene-specific means):
  - sharing information among genes that exhibit similar dispersion
  - an intermediate solution between the trade-off common vs gene-specific dispersion
- Theory for a statistical testing procedure is then developed within the model based clustering framework

Our proposal and outline

- Instead of fitting *p* NB models, we assume a mixture model with component-specific dispersion (and gene-specific means):
  - sharing information among genes that exhibit similar dispersion
  - an intermediate solution between the trade-off common vs gene-specific dispersion
- Theory for a statistical testing procedure is then developed within the model based clustering framework
- Through a wide simulation study we will show that the proposed approach is the best one in reaching the nominal value for the first-type error, while keeping elevate power

Our proposal

The NB parametrization can be derived from a Poisson-Gamma mixed model:

$$\begin{tikzcd}
U \sim \textit{Gamma}(\alpha, \alpha) \arrow[d] \\
Y|U = u \sim \textit{Pois}(\lambda u)
\end{tikzcd}$$

Our proposal

The NB parametrization can be derived from a Poisson-Gamma mixed model:

$$\begin{tikzcd} U \sim \mathit{Gamma}(\alpha, \alpha) \arrow[d] \\ Y|U = u \sim \mathit{Pois}(\lambda u) \end{tikzcd}$$

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

$$Y \sim \mathit{NegBin}(\lambda, \alpha).$$

The proposal

We assume that:

$$f(\mathbf{u}_i) = \sum_{k=1}^{K} w_k f_k(\mathbf{u}_i) = \sum_{k=1}^{K} w_k \prod_{j=1}^{d} \prod_{r=1}^{n_j} Gamma(u_{ijr}; \alpha_k, \alpha_k),$$

Mixtures of NB

Therefore the hierarchical structure becomes:

$$\begin{tikzcd}
Z_i \sim \mathit{Multinom}(1, \mathbf{w}) \text{ where } \mathbf{w} = (w_1, ..., w_K) \arrow[d] \\
U_{ijr} \,|Z_{ik} = 1 \sim \mathit{Gamma}(\alpha_k, \alpha_k) \arrow[d] \\
Y_{ijr}|U_{ijr} = u_{ijr} \sim \mathit{Pois}(\lambda_{ij} u_{ijr})
\end{tikzcd}$$

Mixtures of NB

Therefore the hierarchical structure becomes:

$$Z_i \sim \mathit{Multinom}(1, \mathbf{w}) \text{ where } \mathbf{w} = (w_1, ..., w_K)$$

$$\Big\downarrow$$

$$U_{ijr} \,|Z_{ik} = 1 \sim \mathit{Gamma}(\alpha_k, \alpha_k)$$

$$\Big\downarrow$$

$$Y_{ijr}|U_{ijr} = u_{ijr} \sim \mathit{Pois}(\lambda_{ij} u_{ijr})$$

Marginalizing with respect to U and Z:

$$\mathbf{Y}_i \sim \sum_k w_k \prod_{j=1}^{d} \prod_{r=1}^{n_j} \mathit{NegBin}\left(y_{ijr}; \lambda_{ij}, \alpha_k\right)$$

Estimation

Let $\boldsymbol{\theta} = \{\lambda_{ij}, w_k, \alpha_k\}_{i=1,\ldots p; j=1,\ldots,d; k=1,\ldots,K}$ be the whole set of model parameters.
The log-likelihood of the model is given by

$$\ln L(\boldsymbol{\theta}) = \ln \prod_{i=1}^{p} \sum_{k=1}^{K} w_k \prod_{j=1}^{d} \prod_{r=1}^{n_j} NegBin(y_{ijr}; \lambda_{ij}, \alpha_k)$$

Estimation

Let $\boldsymbol{\theta} = \{\lambda_{ij}, w_k, \alpha_k\}_{i=1,\dots p; j=1,\dots,d; k=1,\dots,K}$ be the whole set of model parameters.
The log-likelihood of the model is given by

$$\ln L(\boldsymbol{\theta}) = \ln \prod_{i=1}^{p} \sum_{k=1}^{K} w_k \prod_{j=1}^{d} \prod_{r=1}^{n_j} NegBin(y_{ijr}; \lambda_{ij}, \alpha_k)$$

A direct maximization of $\ln L(\boldsymbol{\theta})$ is not analytically possible, but the maximum likelihood estimates can be derived by the EM algorithm:

$$\arg\max_{\boldsymbol{\theta}} E_{\mathbf{z},\mathbf{u}|\mathbf{y};\boldsymbol{\theta}'} [\ln L_c(\boldsymbol{\theta})] = \arg\max_{\boldsymbol{\theta}} E_{\mathbf{z},\mathbf{u}|\mathbf{y};\boldsymbol{\theta}'} [\ln f(\mathbf{y}, \mathbf{u}, \mathbf{z}|\boldsymbol{\theta})]$$

which leads to iterating the E and M steps until convergence.

EM algorithm

By evaluating the score function of $E_{\mathbf{z},\mathbf{u}|\mathbf{y};\theta'}$ at zero, with respect to each parameter of the model we get:

EM algorithm

By evaluating the score function of $E_{\mathbf{z},\mathbf{u}|\mathbf{y};\theta'}$ at zero, with respect to each parameter of the model we get:

$$\widehat{\lambda_{ij}} = \frac{\sum_r y_{ijr}}{n_j}$$

EM algorithm

By evaluating the score function of $E_{\mathbf{z},\mathbf{u}|\mathbf{y};\theta'}$ at zero, with respect to each parameter of the model we get:

$$\widehat{\lambda_{ij}} = \frac{\sum_r y_{ijr}}{n_j}$$

the estimates for α_k are not in closed-form therefore we will use quasi-Newton algorithms to find the root of the score equation:

$$\frac{\partial}{\partial \alpha_k} \int_0^{+\infty} \sum_{k=1}^{K} \sum_{i=1}^{p} \sum_{j=1}^{d} \sum_{r=1}^{n_j} \ln f(u_{ijr}|\mathbf{z}_i) f(u_{ijr}, \mathbf{z}_i|\mathbf{y}_i) du_{ijr} = 0$$

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

Three test statistics for differential analysis

Differential analysis: statistical testing to decide whether, for a given gene, an observed difference in read counts between two biological conditions is significant or if it is just due to natural random variability.

Three test statistics for differential analysis

Differential analysis: statistical testing to decide whether, for a given gene, an observed difference in read counts between two biological conditions is significant or if it is just due to natural random variability.

Different ways to accomplish this aim:

- $H_0 : \lambda_{i1} - \lambda_{i2} = 0$

Three test statistics for differential analysis

Differential analysis: statistical testing to decide whether, for a given gene, an observed difference in read counts between two biological conditions is significant or if it is just due to natural random variability.

Different ways to accomplish this aim:

- $H_0 : \lambda_{i1} - \lambda_{i2} = 0$
- $H_0 : \dfrac{\lambda_{i1}}{\lambda_{i2}} = 1$

Three test statistics for differential analysis

Differential analysis: statistical testing to decide whether, for a given gene, an observed difference in read counts between two biological conditions is significant or if it is just due to natural random variability.

Different ways to accomplish this aim:

- $H_0 : \lambda_{i1} - \lambda_{i2} = 0$
- $H_0 : \dfrac{\lambda_{i1}}{\lambda_{i2}} = 1$
- $H_0 : \ln \dfrac{\lambda_{i1}}{\lambda_{i2}} = \ln(\lambda_{i1}) - \ln(\lambda_{i2}) = 0$

Difference test statistic

$H_0 : \lambda_{i1} - \lambda_{i2} = 0$

$$\frac{\widehat{\lambda}_{i1} - \widehat{\lambda}_{i2}}{\sqrt{Var(\widehat{\lambda}_{i1} - \widehat{\lambda}_{i2})}} | H_0 \rightsquigarrow N(0, 1)$$

Difference test statistic

$H_0 : \lambda_{i1} - \lambda_{i2} = 0$

$$\frac{\widehat{\lambda}_{i1} - \widehat{\lambda}_{i2}}{\sqrt{Var(\widehat{\lambda}_{i1} - \widehat{\lambda}_{i2})}} | H_0 \rightsquigarrow N(0,1)$$

$$Var(\widehat{\lambda}_{i1} - \widehat{\lambda}_{i2}) = Var(\widehat{\lambda}_{i1}) + Var(\widehat{\lambda}_{i2})$$
$$Var(\widehat{\lambda}_{ij}) = Var\left(\frac{\sum_{r=1}^{n_j} y_{ijr}}{n_j}\right) = \frac{1}{n_j^2} n_j Var(y_{ijr})$$

Difference test statistic

$H_0 : \lambda_{i1} - \lambda_{i2} = 0$

$$\frac{\widehat{\lambda}_{i1} - \widehat{\lambda}_{i2}}{\sqrt{Var(\widehat{\lambda}_{i1} - \widehat{\lambda}_{i2})}} | H_0 \rightsquigarrow N(0,1)$$

$$Var(\widehat{\lambda}_{i1} - \widehat{\lambda}_{i2}) = Var(\widehat{\lambda}_{i1}) + Var(\widehat{\lambda}_{i2})$$
$$Var(\widehat{\lambda}_{ij}) = Var\left(\frac{\sum_{r=1}^{n_j} y_{ijr}}{n_j}\right) = \frac{1}{n_j^2} n_j Var(y_{ijr})$$
$$Var(y_{ijr}) = E[Var(y_{ijr}|z_{ik} = 1)] + Var[E(y_{ijr}|z_{ik} = 1)]$$

Difference test statistic

$H_0 : \lambda_{i1} - \lambda_{i2} = 0$

$$\frac{\widehat{\lambda}_{i1} - \widehat{\lambda}_{i2}}{\sqrt{Var(\widehat{\lambda}_{i1} - \widehat{\lambda}_{i2})}} | H_0 \rightsquigarrow N(0,1)$$

$$Var(\widehat{\lambda}_{i1} - \widehat{\lambda}_{i2}) = Var(\widehat{\lambda}_{i1}) + Var(\widehat{\lambda}_{i2})$$
$$Var(\widehat{\lambda}_{ij}) = Var\left(\frac{\sum_{r=1}^{n_j} y_{ijr}}{n_j}\right) = \frac{1}{n_j^2} n_j Var(y_{ijr})$$
$$Var(y_{ijr}) = E[Var(y_{ijr}|z_{ik} = 1)] + Var[E(y_{ijr}|z_{ik} = 1)]$$

and for $E[Var(y_{ijr}|z_{ik} = 1))]$ we consider the conditional expectation given the observed data

$$Var(y_{ijr}) \quad = \quad E_{\mathbf{z}_i|\mathbf{y}_i}[Var(y_{ijr}|z_{ik} = 1)] = \widehat{\lambda}_{ij}\left(1 + \sum_k \frac{f(z_{ik}|\mathbf{y}_i)}{\widehat{\alpha_k}}\widehat{\lambda}_{ij}\right)$$

Ratio test statistic

$H_0 : \frac{\lambda_{i1}}{\lambda_{i2}} = 1$

$$\frac{\frac{\widehat{\lambda}_{i1}}{\widehat{\lambda}_{i2}} - 1}{\sqrt{Var\left(\frac{\widehat{\lambda}_{i1}}{\widehat{\lambda}_{i2}}\right)}} | H_0 \rightsquigarrow N(0,1)$$

Ratio test statistic

$H_0 : \frac{\lambda_{i1}}{\lambda_{i2}} = 1$

$$\frac{\frac{\widehat{\lambda}_{i1}}{\widehat{\lambda}_{i2}} - 1}{\sqrt{Var\left(\frac{\widehat{\lambda}_{i1}}{\widehat{\lambda}_{i2}}\right)}} | H_0 \rightsquigarrow N(0,1)$$

using Delta method:

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

Log Ratio test statistic

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

$$\frac{\ln \widehat{\lambda}_{i1} - \ln \widehat{\lambda}_{i2}}{\sqrt{Var(\ln \widehat{\lambda}_{i1} - \ln \widehat{\lambda}_{i2})}} | H_0 \rightsquigarrow N(0, 1)$$

Log Ratio test statistic

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

$$\frac{\ln \widehat{\lambda}_{i1} - \ln \widehat{\lambda}_{i2}}{\sqrt{Var(\ln \widehat{\lambda}_{i1} - \ln \widehat{\lambda}_{i2})}} | H_0 \rightsquigarrow N(0,1)$$

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

Log Ratio test statistic

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

$$\frac{\ln \widehat{\lambda}_{i1} - \ln \widehat{\lambda}_{i2}}{\sqrt{Var(\ln \widehat{\lambda}_{i1} - \ln \widehat{\lambda}_{i2})}} | H_0 \rightsquigarrow N(0,1)$$

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

$$Var(\ln \widehat{\lambda}_{ij}) = Var\left(\ln\left(\frac{\sum_r y_{ijr}}{n_j}\right)\right) = Var(\ln(\sum_r y_{ijr}))$$

Log Ratio test statistic

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

$$\frac{\ln \widehat{\lambda}_{i1} - \ln \widehat{\lambda}_{i2}}{\sqrt{Var(\ln \widehat{\lambda}_{i1} - \ln \widehat{\lambda}_{i2})}} | H_0 \rightsquigarrow N(0,1)$$

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

$$Var(\ln \widehat{\lambda}_{ij}) = Var\left(\ln\left(\frac{\sum_r y_{ijr}}{n_j}\right)\right) = Var(\ln(\sum_r y_{ijr}))$$

through the Delta method: $Var(\ln(\sum_r y_{ijr})) = \frac{1}{(\sum_r y_{ijr})^2} n_j Var(y_{ijr})$

Simulation A

Evaluating the capability of the proposed mixture model to estimate the variances of the genes as K increases.

Simulation A

Evaluating the capability of the proposed mixture model to estimate the variances of the genes as K increases.

A set of $H = 100$ datasets with:

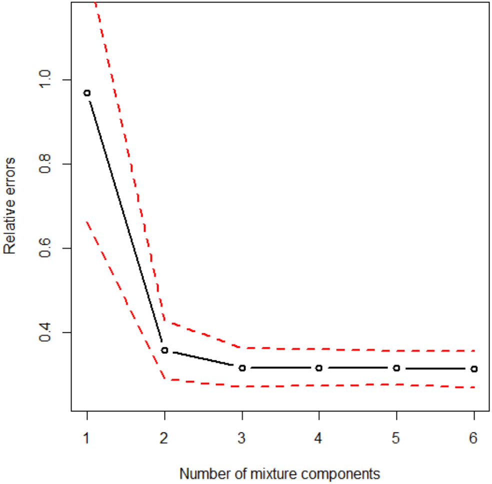
- $d = 2$ conditions
- $n_1 = n_2 = 5$ replicates
- $p = 300$ genes:
  - $\frac{1}{3}$ genes DE ($\lambda_{i1} \neq \lambda_{i2}$)
    $\lambda_{i1} \sim Unif(0, 250)$, $\lambda_{i2} = \frac{\lambda_{i1}}{e^{\phi_i}}$ where $\phi_i \sim N(\mu = 0.5, \sigma = 0.125)$
  - $\frac{2}{3}$ genes not DE ($\lambda_{i1} = \lambda_{i2}$)
    $\lambda_{i1} = \lambda_{i2} \sim Unif(0, 250)$
- $\alpha_i \sim Unif(0.5, 600)$ ($i = 1, \dots, p$)

Simulation A

Average of the relative errors in absolute values across the 100 datasets between the estimated variances and the true ones as K varies.

Simulation A

Average of the relative errors in absolute values across the 100 datasets between the estimated variances and the true ones as *K* varies.



Simulation A

Comparison with the others

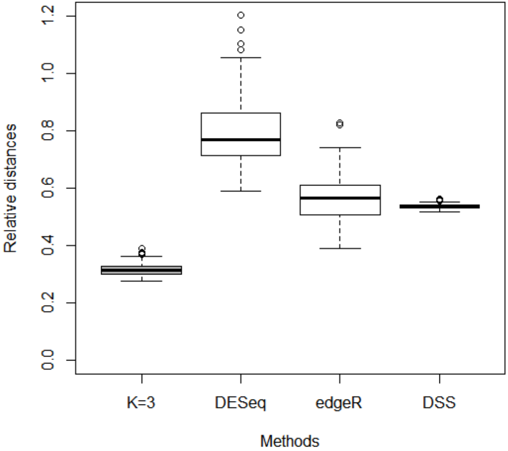
Comparison with Robinson et al 2010 (*edgeR* package), Anders and Huber 2010 (*DESeq* package), Wu et al 2013 (*DSS* package)

Simulation A

Comparison with the others

Comparison with Robinson et al 2010 (*edgeR* package), Anders and Huber 2010 (*DESeq* package), Wu et al 2013 (*DSS* package)

Relative distances between the estimated variances and the true ones (across the 100 datasets).



Simulation B

Evaluation of the adequateness of the statistical procedure: by observing the approximation of the empirical first-type error towards the nominal significance level under the null hypothesis as the number of replicates increases.

Simulation B

Evaluation of the adequateness of the statistical procedure: by observing the approximation of the empirical first-type error towards the nominal significance level under the null hypothesis as the number of replicates increases.

The same simulation design presented before: $d = 2$ conditions, 100 genes DE ($\lambda_{i1} \neq \lambda_{i2}$), 200 genes not DE ($\lambda_{i1} = \lambda_{i2}$), $\alpha_i \sim Unif(0.5, 600)$

Simulation B

Evaluation of the adequateness of the statistical procedure: by observing the approximation of the empirical first-type error towards the nominal significance level under the null hypothesis as the number of replicates increases.

The same simulation design presented before: $d = 2$ conditions, 100 genes DE ($\lambda_{i1} \neq \lambda_{i2}$), 200 genes not DE ($\lambda_{i1} = \lambda_{i2}$), $\alpha_i \sim Unif(0.5, 600)$

with:

- $H = 1000$ datasets;
- a varying number of replicates $n_j = 3, 5, 10$;
- $K = 3$ components

Simulation B

Evaluation of the adequateness of the statistical procedure: by observing the approximation of the empirical first-type error towards the nominal significance level under the null hypothesis as the number of replicates increases.

The same simulation design presented before: $d = 2$ conditions, 100 genes DE ($\lambda_{i1} \neq \lambda_{i2}$), 200 genes not DE ($\lambda_{i1} = \lambda_{i2}$), $\alpha_i \sim \mathit{Unif}(0.5, 600)$

with:

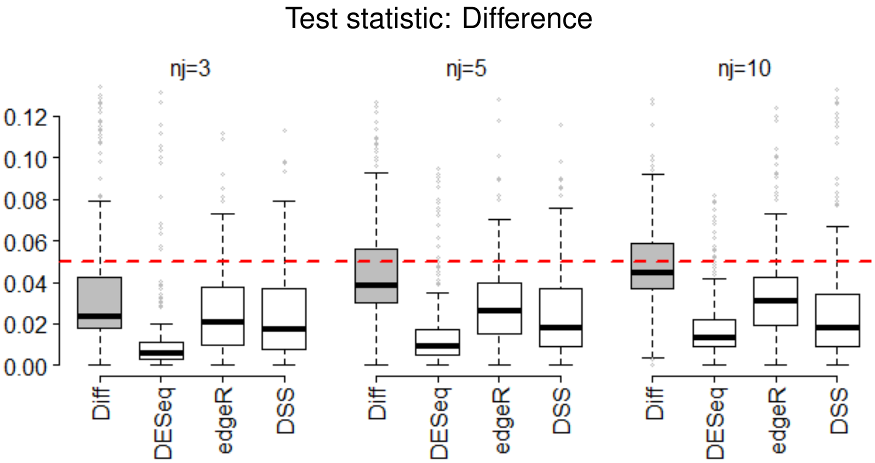
- $H = 1000$ datasets;
- a varying number of replicates $n_j = 3, 5, 10$;
- $K = 3$ components

Comparison with Robinson et al 2010 (*edgeR* package), Anders and Huber 2010 (*DESeq* package), Wu et al 2013 (*DSS* package)

Simulation B

First-type errors

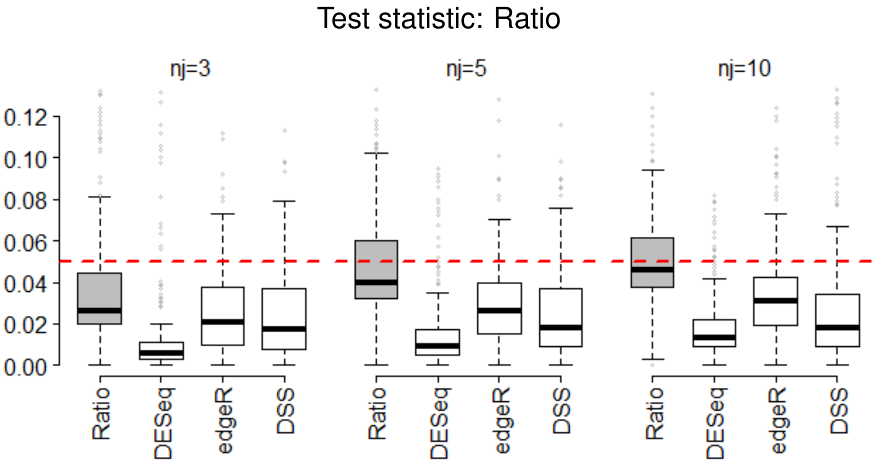
Confidence level= 0.05



Simulation B

First-type errors

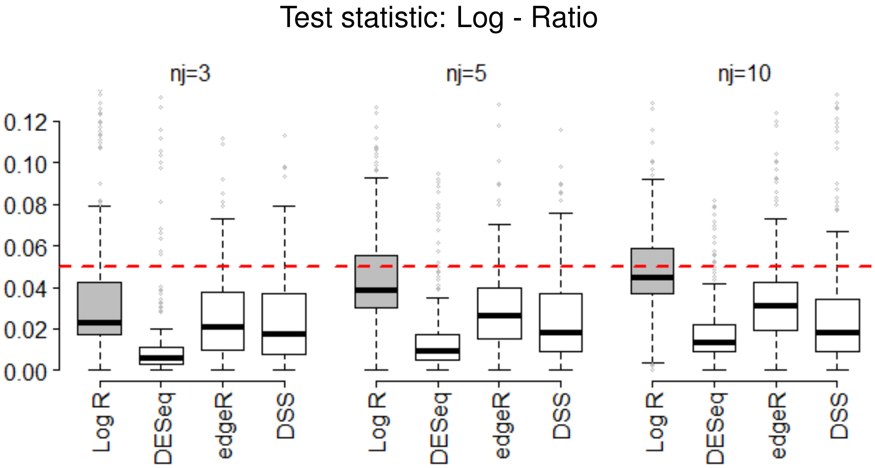
Confidence level= 0.05



Simulation B

First-type errors

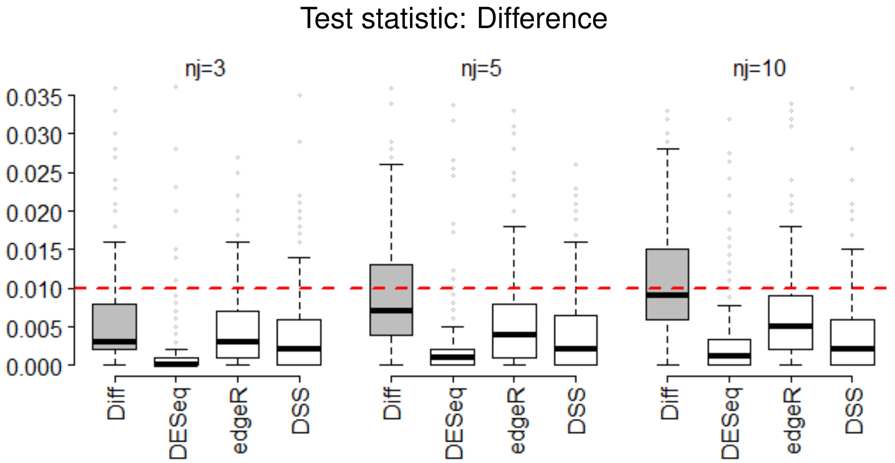
Confidence level= 0.05



Simulation B

First-type errors

Confidence level= 0.01

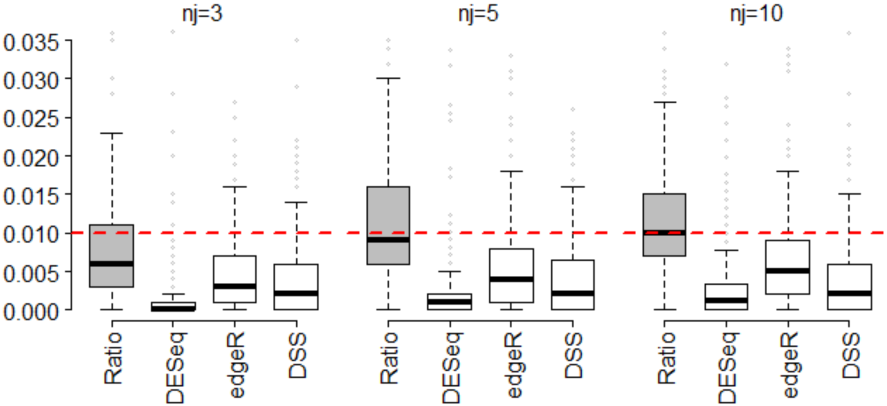


Simulation B

First-type errors

Confidence level= 0.01

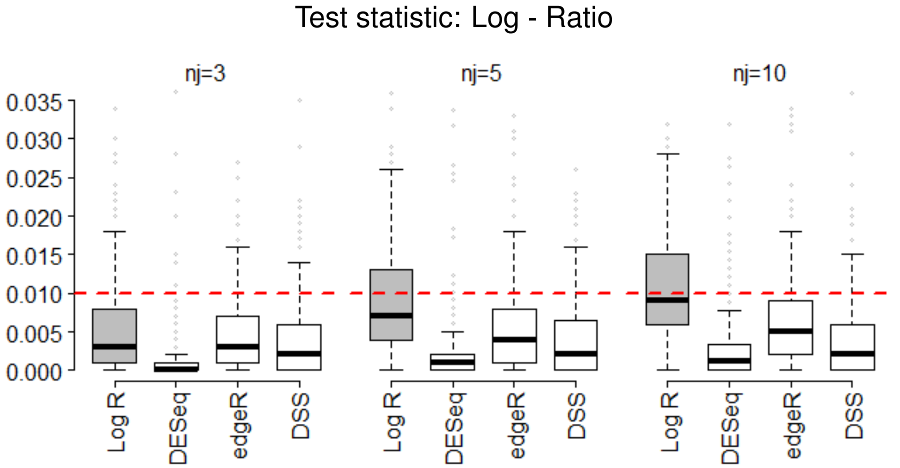
Test statistic: Ratio



Simulation B

First-type errors

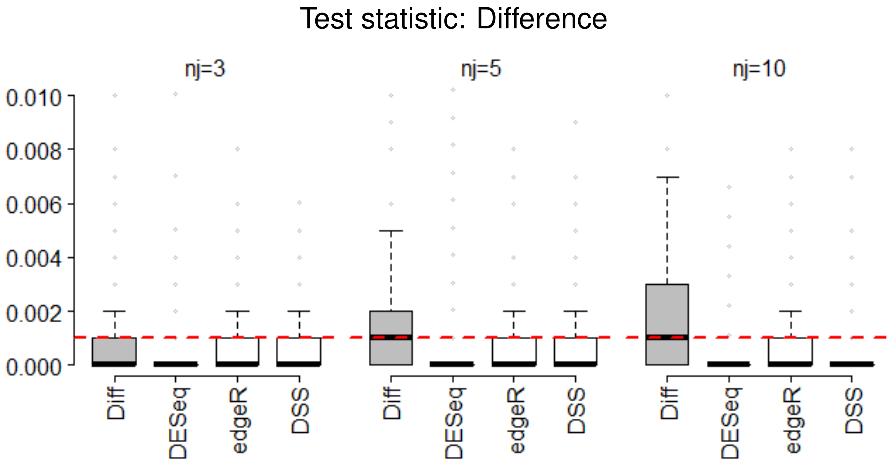
Confidence level= 0.01



Simulation B

First-type errors

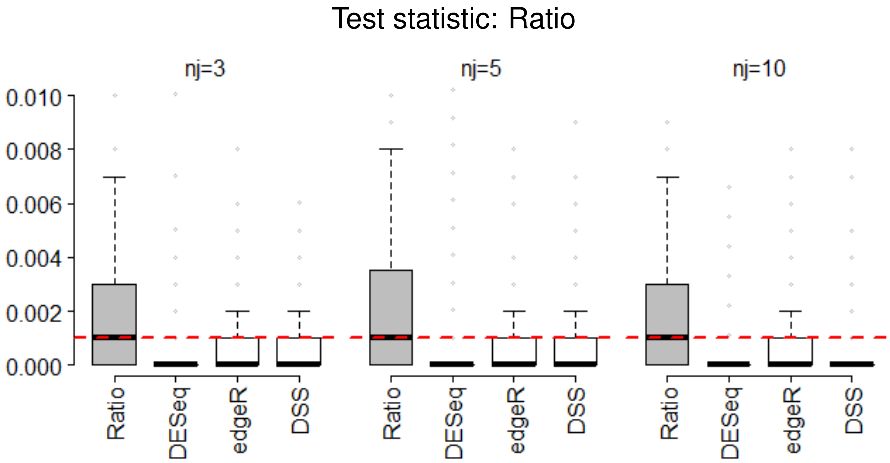
Confidence level= 0.001



Simulation B

First-type errors

Confidence level= 0.001

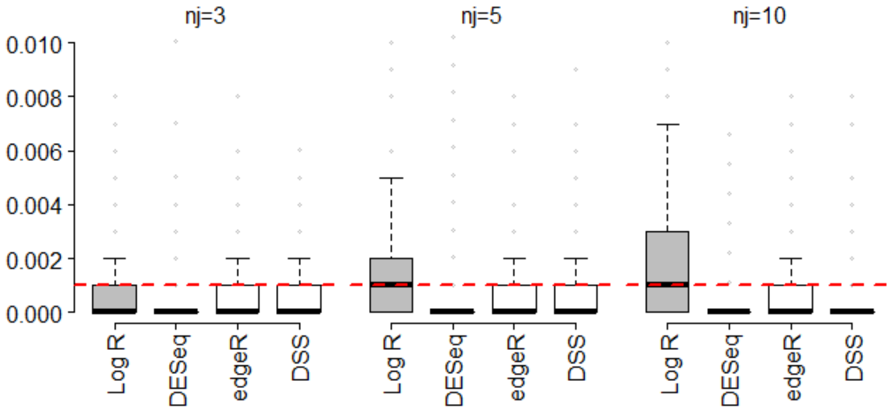


Simulation B

First-type errors

Confidence level= 0.001

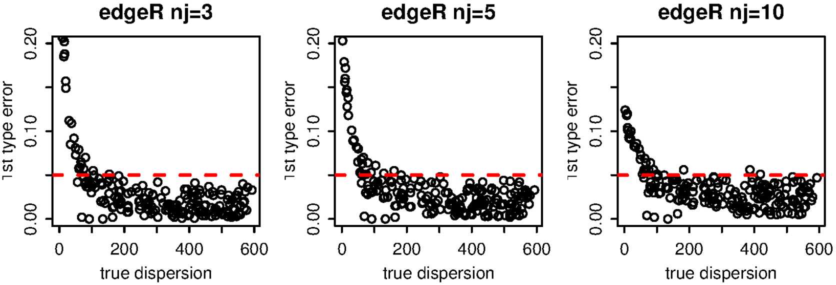
Test statistic: Log - Ratio



Simulation B: Empirical first-type errors as a function of the real dispersion parameters α_i.

1^{st} type errors and real α_i - edgeR

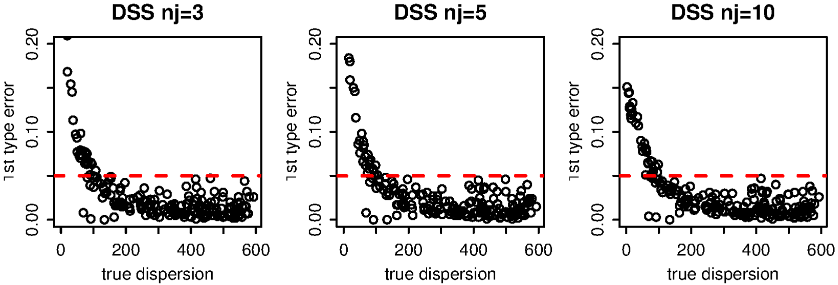
Confidence level= 0.05



Simulation B

1st type errors and real α_i - DSS

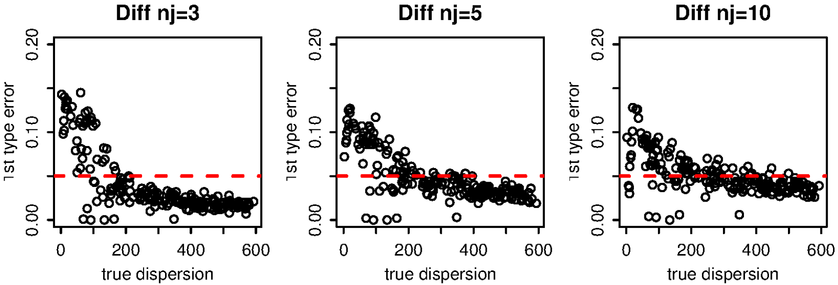
Confidence level= 0.05



Simulation B

1st type errors and real α_i - Difference

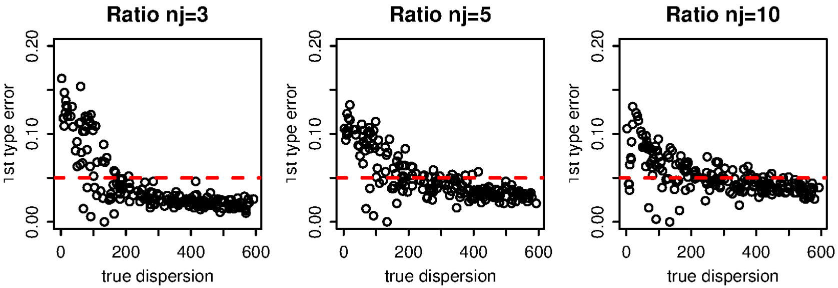
Confidence level= 0.05



Simulation B

1^{st} type errors and real α_i - Ratio

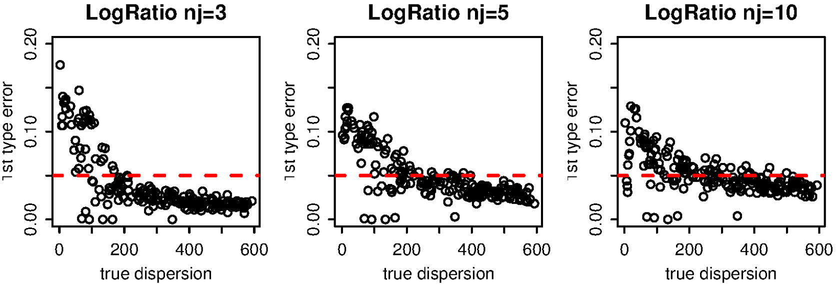
Confidence level= 0.05



Simulation B

1^{st} type errors and real α_i - Log Ratio

Confidence level= 0.05



Simulation B

ECDF of the null p-values

The capability of controlling the first-type error can be checked also by looking at the empirical cumulative density function (ECDF) of the null p-values;

Simulation B

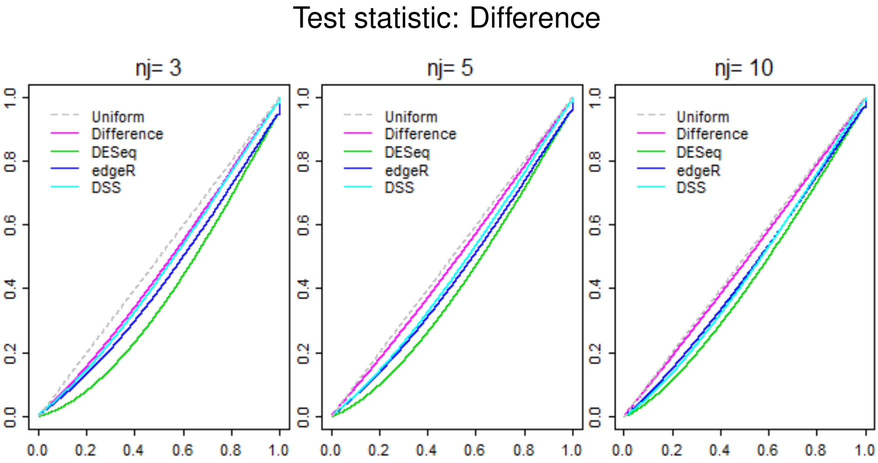
ECDF of the null p-values

The capability of controlling the first-type error can be checked also by looking at the empirical cumulative density function (ECDF) of the null p-values;

the closer their distribution is to the diagonal, the better is the approximation to the uniform distribution, as requested by the *probability integral transform theorem*.

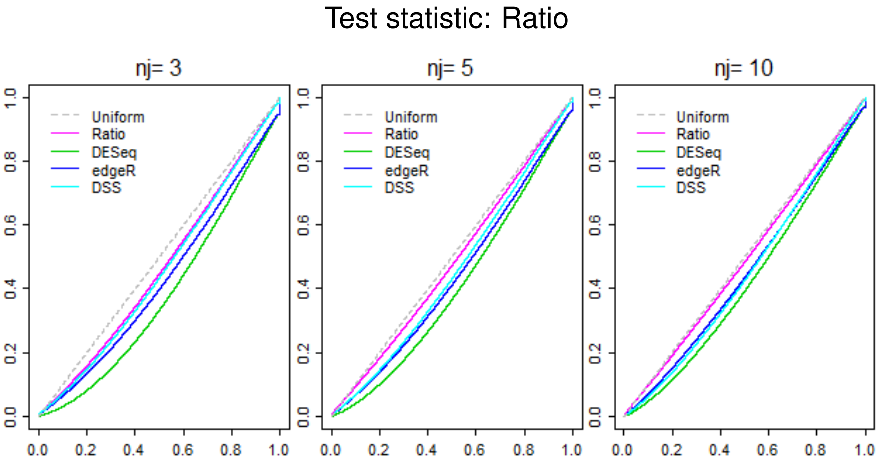
Simulation B

ECDF of the null p-values



Simulation B

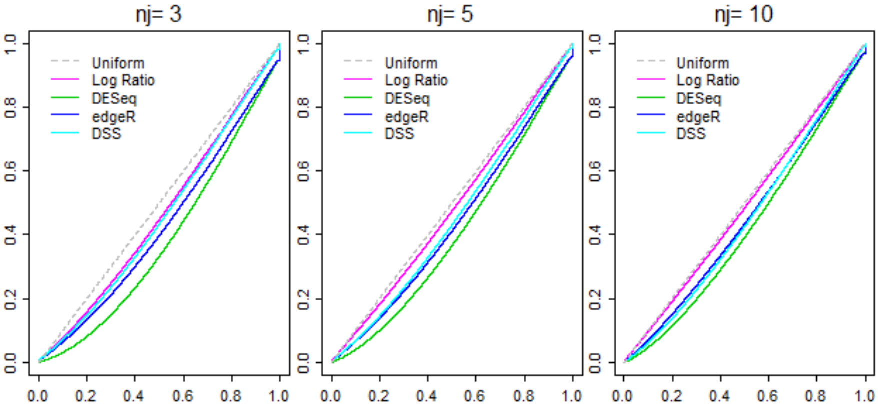
ECDF of the null p-values



Simulation B

ECDF of the null p-values

Test statistic: Log - Ratio



nj= 3

nj= 5

nj= 10

Application to prostate cancer data

The dataset

RNA-seq data on prostate cancer cells, two conditions:

1. treated with androgens ($n_j = 3$ patients)
2. control (inactive compound) ($n_j = 4$ patients)

37435 genes were sequenced; for the analysis we have considered the $p = 16424$ genes with mean count greater than 1.

Application to prostate cancer data

The dataset

RNA-seq data on prostate cancer cells, two conditions:

1. treated with androgens ($n_j = 3$ patients)
2. control (inactive compound) ($n_j = 4$ patients)

37435 genes were sequenced; for the analysis we have considered the $p = 16424$ genes with mean count greater than 1.

Androgen hormones: stimulate some genes
have a positive effect in curing prostate cancer cells

$\Rightarrow$ Differential analysis: investigation of the connection between these stimulated genes and survival of these cells

Application to prostate cancer data

The dataset

Preliminaries: the data have been normalized in order to account for possible technical biases and for the gene lengths.

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
⋮							

Application to prostate cancer data

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
$\Rightarrow$ Information criteria (AIC, BIC): $K = 3$.

Application to prostate cancer data

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
$\Rightarrow$ Information criteria (AIC, BIC): $K = 3$.

Differential expression analysis has been conducted by computing the three proposed test statistics and also using the *DESeq*, *edgeR* and *DSS* methods implemented in R using the default settings.

The proposed NB mixture model has been fitted on the data with a number of components K ranging from 1 to 6
$\Rightarrow$ Information criteria (AIC, BIC): $K = 3$.

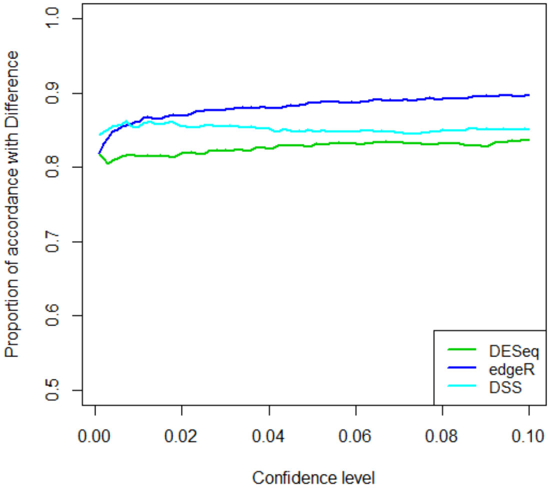
Differential expression analysis has been conducted by computing the three proposed test statistics and also using the *DESeq*, *edgeR* and *DSS* methods implemented in R using the default settings.

$$\text{Acc. level} = \frac{\text{num. of genes jointly declared DE}}{\text{average (num. of genes marginally declared DE)}}$$

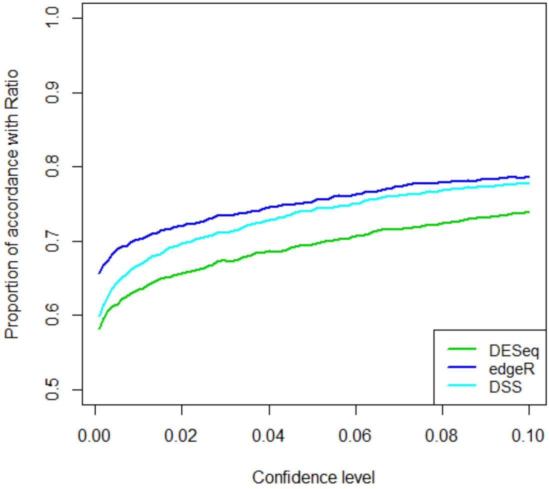
Application to prostate cancer data

Analysis and results

Difference test statistic



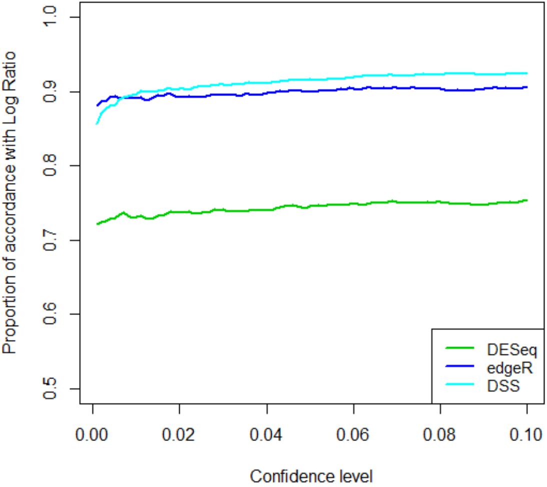
Ratio test statistic



Application to prostate cancer data

Analysis and results

Log Ratio test statistic



Conclusions

- The proposed mixture of Negative Binomials is a new way for sharing information among genes about their dispersion levels, and to gain a more accurate estimation of the variances;

Conclusions

- The proposed mixture of Negative Binomials is a new way for sharing information among genes about their dispersion levels, and to gain a more accurate estimation of the variances;
- Three different statistical tests have been proposed, compared and investigated in a wide simulation study;

Conclusions

- The proposed mixture of Negative Binomials is a new way for sharing information among genes about their dispersion levels, and to gain a more accurate estimation of the variances;
- Three different statistical tests have been proposed, compared and investigated in a wide simulation study;
- The simulation study results show that the proposed test statistics are the only ones that actually reach the nominal values for the first-type errors (and they are good also in restraining the second-type ones).

Some References

Anders, Simon and Huber, Wolfgang. (2010). Differential expression analysis for sequence count data. Genome Biology 11(10), 112.

Cox, Christopher. (1990). Fiellers theorem, the likelihood and the delta method. Biometrics 46(3), pp. 709718.

Dempster, M., Laird, N. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). Journal of the Royal Statistical Society B 39, 138.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis and density estimation. Journal of the American Statistical Association 97, 611631.

Hilbe, J.M. (2011). Negative Binomial Regression. Cambridge University Press.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. and Gilad, Y. (2008). RNAseq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Research 18, 15091517.

Some References

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

Robinson, Mark D, McCarthy, Davis J and Smyth, Gordon K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1), 139140.

Robinson, Mark D. and Smyth, Gordon K. (2007). Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 23(21), 28812887.

Wang, Z., Gerstein, M. and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10(1), 5763.

Wu H., Wang C., Wu Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. Biostatistics. 2013 Apr;14(2):232-43.

First type errors (mean and SD)
Confidence level= 0.05

Statistic	$n_j = 3$	$n_j = 5$	$n_j = 10$
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)

First type errors (mean and SD)
Confidence level= 0.01

Statistic	$n_j = 3$	$n_j = 5$	$n_j = 10$
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)

First type errors (mean and SD)
Confidence level= 0.001

Statistic	$n_j = 3$	$n_j = 5$	$n_j = 10$
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)

Second type errors (mean and SD)
Confidence level= 0.05

Statistic	$n_j = 3$	$n_j = 5$	$n_j = 10$
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)

Second type errors (mean and SD)
Confidence level= 0.01

Statistic	$n_j = 3$	$n_j = 5$	$n_j = 10$
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)

Second type errors (mean and SD)
Confidence level= 0.001

Statistic	$n_j = 3$	$n_j = 5$	$n_j = 10$
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)

AUC (adjusted p-values; average on the H= 1000 datasets)

	$n_j = 3$	$n_j = 5$	$n_j = 10$
Difference	0.950	0.968	0.986
Ratio	0.936	0.959	0.981
Log Ratio	0.951	0.968	0.986
DESeq	0.952	0.970	0.986
edgeR	0.956	0.972	0.987
DSS	0.958	0.974	0.988