How to use the LBE package

C. Dalmasso

April 25, 2023

1 Introduction

In the context of genome-wide studies for which a large number of statistical tests are simultaneously performed, the False Discovery Rate (FDR) which is defined as the expected proportion of false discoveries [1] is one of the most used criterion for taking into account the multiple testing problem.

In the framework of estimating procedures based on the marginal distribution of the p-values without assumption on the conditional distribution related to the alternative hypothesis, estimators of the FDR rely on the formula introduced by Storey [4]:

$$FDR(\Gamma) = \frac{\pi_0 \Pr(P \in G | H = 0)}{\Pr(P \in G)} \tag{1}$$

where H is the variable such that $H = 0$ if the null hypothesis H_0 is true, $H = 1$ if the alternative hypothesis H_1 is true, $\pi_0 = \Pr(H = 0)$ is the probability of not being modified and P is the random variable corresponding to the p-values.

This document provides a tutorial for using the `LBE` package that contains functions for estimating the proportion of true null hypotheses π_0 and the FDR (or the q-values that are defined for each p-value by q-value$(p_i) = FDR([0, p_i])$). We describe here the implemented functions and illustrate their use with the real dataset from Golub et al. [3]. In the last section, the LBE function is compared with the `qvalue` function from the `qvalue` package.

2 The leukemia data set from Golub et al. (1999)

The aim of the study of Golub et al. [3] was to identify differentially expressed genes between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). The samples were assayed using Affymetrix Hgu6800 chips and the data on the expression of 7129 genes are available in the Bioconductor package `golubEsets`.

The dataset `golub.pval` provided with the LBE package contains the p-values obtained from a two-sample t-test analysis. The variance-stabilizing method included in the `vsn` package was applied for normalizing the data.

```
> library(LBE)
> data(golub.pval)
```

3 Implemented functions

3.1 LBE

Under the null hypothesis, the p-values are supposed to be uniformly distributed on [0,1] so that $\Pr(P \in [0,\gamma]|H=0) = \gamma$. The FDR estimation is then obtained from the relation (1) by the separate estimation of $\Pr(P \in [0,\gamma])$ and π_0, the proportion of true null hypotheses. While $\Pr(P \in [0,\gamma])$ can easily be estimated by the empirical cumulative distribution, if no distributional assumption is made for the marginal distribution of the p-values, only an upper bound estimate can be obtained for π_0.

Let m be the total number of p-values. From a classical two-components mixture model for the distribution of the p-values, we have introduced [2] a conservatively biased estimator of π_0:

$$\widehat{\pi}_0 = 2\frac{1}{m}\sum_{i=1}^{m} P_i \tag{2}$$

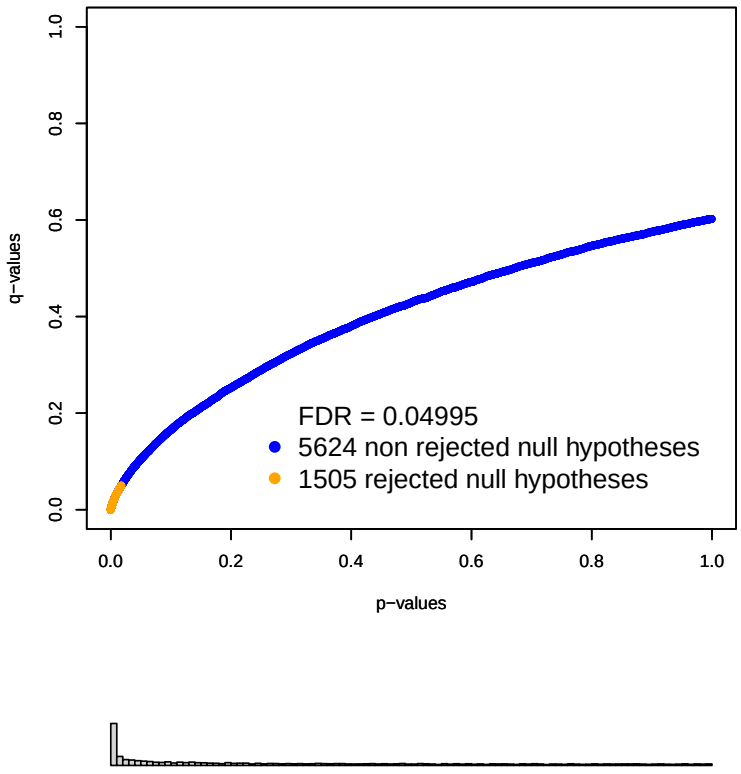
From this estimator of π_0, we have demonstrated that under suitable conditions for a function φ, transformed p-values lead to a less biased estimator of π_0 (see the theorem 1 in [2]). In this context, we have considered the functions $\varphi_a(P) = -\ln(1-x)^a,\ a \in [1,+\inf[$, and we have demonstrated that these functions lead to a family of estimators for which the bias for π_0 is decreasing with a. The obtained results can easily be extended to real values of a leading to the following family of estimators:

$$\hat{\pi}_{0(a)} = \frac{\frac{1}{m}\sum_{i=1}^{m}[-\ln(1-p_i)]^a}{\Gamma(a+1)},\ a \in [1,+\inf[. \tag{3}$$

For this family of estimators,an upper bound of the asymptotic variance can be obtained for independent p-values: $\frac{1}{m} \times \left(\frac{\Gamma(2a+1)}{\Gamma(a+1)^2} - 1\right)$ leading to a confidence interval for π_0. As there is a balance between bias (decreasing as a increase) and variance (increasing as a increase), for a specified number m of tested hypotheses, we have proposed to choose a as the greatest value such that the upper bound of the standard deviation is less than a threshold l for the variance's upper bound. Other rules may obviously be considered and the `LBE` function allows to set a independently from the variance upper bound.

The `LBE` function only requires a vector of p-values as input. We first create the object `LBE.res` by applying the `LBE` function with default arguments. A plot of the q-values versus the p-values is displayed together with the histogram of the p-values and informative numerical values in the legend such as the FDR and the number of rejected null hypotheses. Among the saved results `LBE.res`, we display the estimate of π_0, its confidence interval and the (default) level for the confidence interval. Then, we apply the `LBE` function once again by changing the level for the confidence interval and we display the new confidence interval for π_0. The argument plot.type is set to "none" so that the plot is not displayed.

```
> LBE.res <- LBE(golub.pval)
```




```
> #LBE.res <- LBE(golub.pval)
> names(LBE.res)

 [1] "call"          "FDR"           "pi0"           "pi0.ci"
 [5] "ci.level"      "a"             "l"             "qvalues"
 [9] "pvalues"       "significant"   "n.significant"

> LBE.res$pi0; LBE.res$pi0.ci; LBE.res$ci.level

[1] 0.6022541

[1] 0.0000000 0.6844972

[1] 0.95

> LBE.res2 <- LBE(golub.pval, ci.level=0.8,plot.type="none")
> LBE.res2$pi0.ci; LBE.res2$ci.level
```

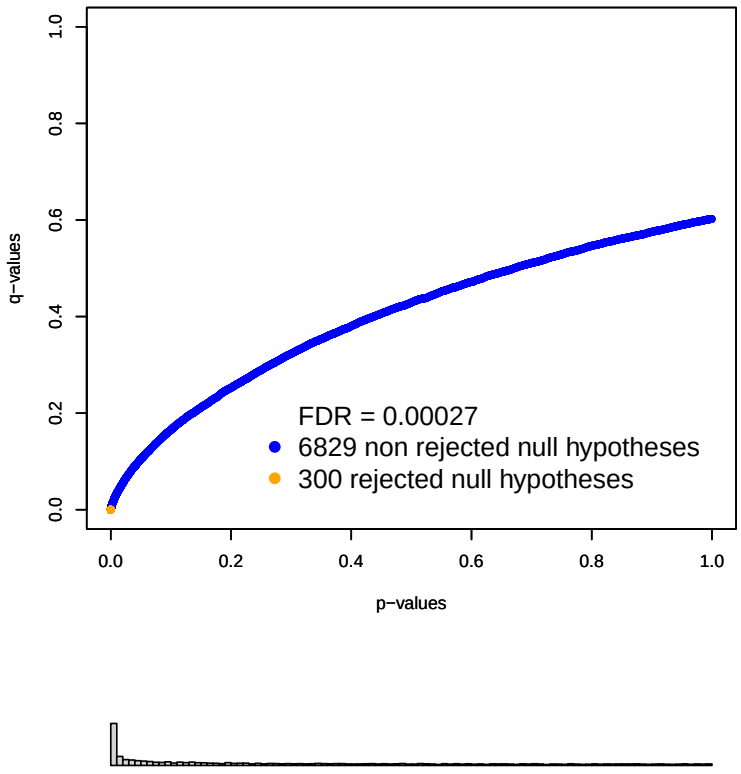
```
[1] 0.0000000 0.6443354
```

```
[1] 0.8
```

If the argument `qvalues` is set to `FALSE`, only π_0 is estimated. Otherwise (default value), the q-values (which are defined for each gene by q-value$(p_i) = FDR([0, p_i])$ are estimated and the `LBE` function returns either the number of significant genes when controlling the FDR at a specific level (the default value for the argument `FDR.level` is 0.05), either the estimated FDR for a specific number of significant genes.

First, we apply the `LBE` function by changing the level at which control the FDR, then we estimate the FDR when 300 genes are declared significant. It is worth noting that the estimated q-values remain unchanged.

```
> LBE.res3 <- LBE(golub.pval, FDR.level=0.1,plot.type="none")
> LBE.res3$qvalues[1:10]

 [1] 0.1725651 0.5072883 0.4221644 0.1578850 0.4381401 0.5353073 0.2620085
 [8] 0.1581277 0.2720757 0.4948859

> LBE.res3$n.significant

[1] 2008

> LBE.res4 <- LBE(golub.pval,FDR.level=NA,n.significant=300)
```




```
> LBE.res4$qvalues[1:10]

 [1] 0.1725651 0.5072883 0.4221644 0.1578850 0.4381401 0.5353073 0.2620085
 [8] 0.1581277 0.2720757 0.4948859
```

Using the proposed rule for choosing a particular estimator in the family (3), the parameter `a` is set according to a threshold `l` for the asymptotic standard deviation. The default value for `l` is 0.05 that is considered to be small enough, but other values can be chosen : the function `LBEa` described below illustrates the relation between `a` and `l` for a fixed number of tested hypotheses.

A particular value for `a` can also be directly set (without using the parameter `l`). Choosing values of `a` less that 1 leads to use the identity function for transforming the p-values, that is to say the estimator (2).

First we apply the function `LBE` by changing the upper bound for the asymptotic standard deviation, then, we apply `LBE` by arbitrarily setting $a = 2$ and finally, we do not transform the p-values (by setting $a = -1$).

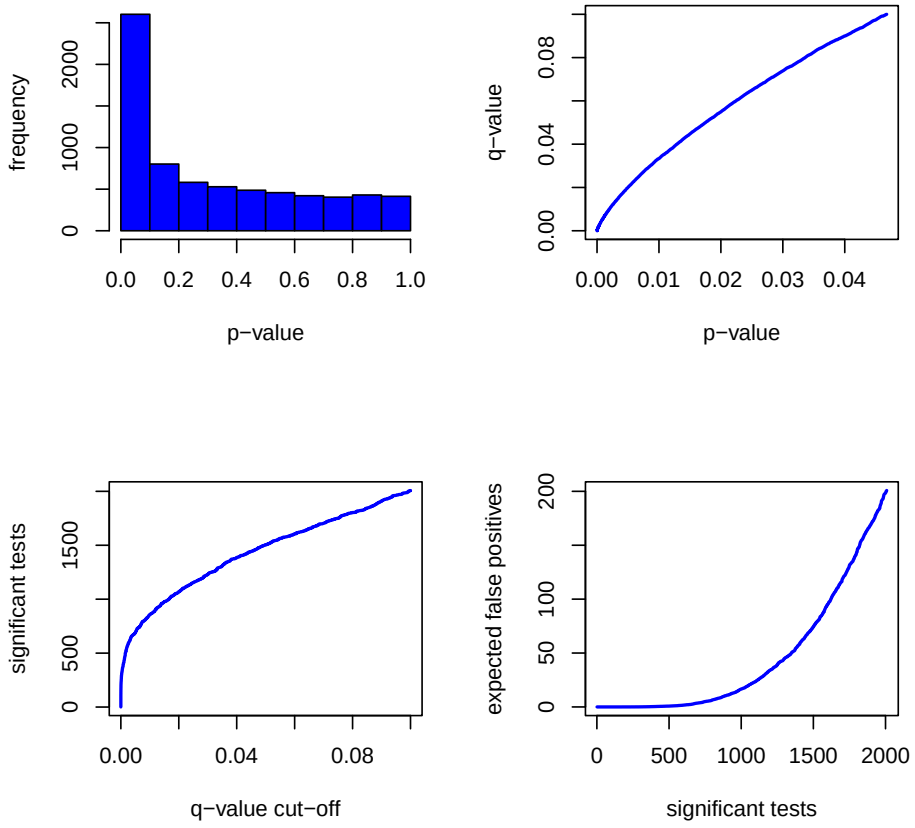
```
> LBE.res5 <- LBE(golub.pval,a=2,l=NA,plot.type="none")
> LBE.res5$a; LBE.res5$l; LBE.res5$pi0; LBE.res5$pi0.ci; LBE.res5$n.significant

[1] 2

[1] 0.02648321

[1] 0.5996953

[1] 0.0000000 0.6432563

[1] 1505

> LBE.res6 <- LBE(golub.pval,a=NA,l=0.1,plot.type="none")
> LBE.res6$a; LBE.res6$l; LBE.res6$pi0; LBE.res6$pi0.ci; LBE.res6$n.significant

[1] 4.025365

[1] 0.1000004

[1] 0.6015066

[1] 0.0000000 0.7659925

[1] 1505

> LBE.res7 <- LBE(golub.pval,a=-1,l=NA,plot.type="none")
> LBE.res7$a; LBE.res7$l; LBE.res7$pi0; LBE.res7$pi0.ci; LBE.res7$n.significant

[1] NA

[1] 0.006837937

[1] 0.6501887

[1] 0.0000000 0.6614361

[1] 1457
```

3.2 LBEplot

If `plot.type="main"`, the function `LBEplot`, that is called by the main function `LBE`, displays the plot of the q-values versus the p-values together with the histogram of the p-values. The FDR and the numbers of significant and non significant genes are displayed in the legend.

If `plot.type="multiple"` (default value), the function LBEplot produces four plots: an histogram of the p-values, the plot of the q-values versus the p-values, the number of significant genes versus the q-values and the number of expected false positives by the number of significant genes.

```
> LBEplot(LBE.res,plot.type="multiple")
```

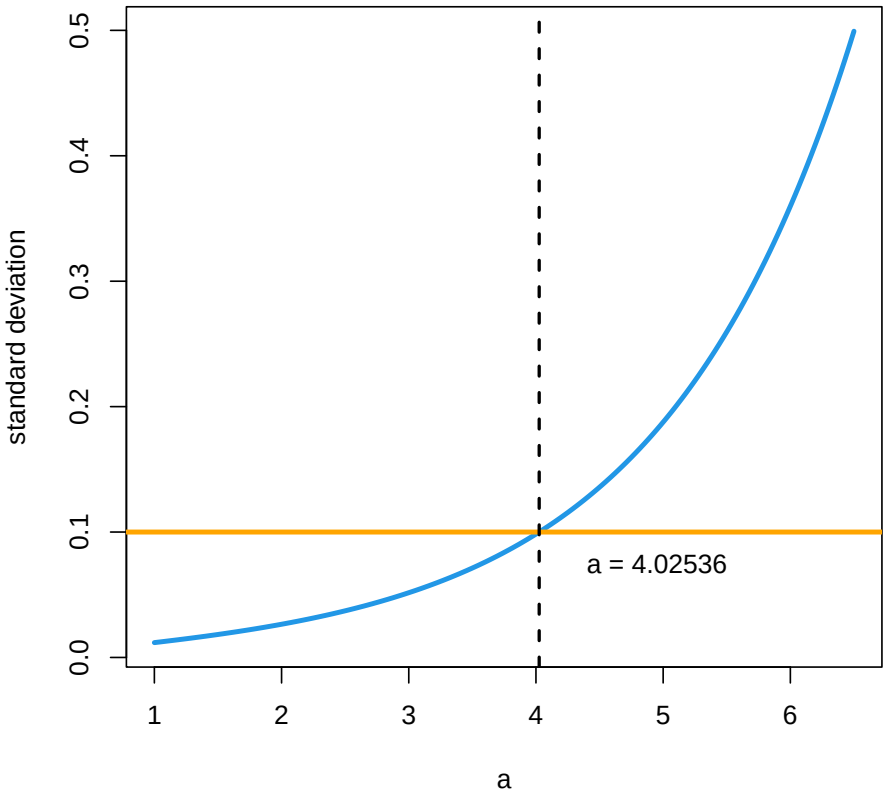


3.3 LBEa

The `LBEa` function is called by the main function `LBE` for choosing the greatest value of `a` such that the upper bound of the asymptotic standard deviation is less than a threshold `l`. A plot illustrating the relation between `a` and `l` for a fixed number of tested hypotheses is displayed.

```
> LBEa(length(golub.pval),l=0.1)
```

```
[1] 4.025365
```



3.4 LBEsummary

The function `LBEsummary` is analogous to the function `summary` from the `qvalue` package. It reports an estimate and a confidence interval for the proportion of true null hypotheses and presents a table comparing p-values to q-values.

```
> LBEsummary(LBE.res)

Call:
LBE(pval = golub.pval)

pi0:         0.6022541
Confidence Interval (level=0.95): [0,0.6844972]

Cumulative number of significant calls:

         <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1    <1
```

```
p-value    415    731  1287   1654  2053 2602 7129
q-value    236    415   858   1154  1505 2008 7129
```

3.5 `LBEwrite`

The `LBEwrite` function is analogous to the `qwrite` function from the `qvalue` package. It writes the output of the function `LBE` to a file.

```
> LBEwrite(LBE.res)
```

4 Comparison with the `qvalue` package

The `qvalue` package contains functions for the estimation and presentation of the q-values following the method introduced by Storey and Tibshirani [5]. While the default method for estimating π_0 (in the `qvalue` function) relies on a smoothing method for estimating the marginal density evaluated at one, LBE is based on the expectation of a transformation of the p-values.

As regards to `qvalue`, the results of a simulation study [2] indicates good performances for `LBE`. Moreover, the thoretical results we have obtained allow the calculation of a confidence interval for π_0.

We compare here the two methods with the dataset from Golub *et al. [3].*

```
> library(qvalue)
> qvalue.res <- qvalue(golub.pval)
> summary(qvalue.res)

Call:
qvalue(p = golub.pval)

pi0:        0.5841632

Cumulative number of significant calls:

          <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1   <1
p-value      415    731  1287   1654  2053 2602 7129
q-value      236    415   863   1163  1519 2037 7129
local FDR    160    291   571    734   921 1188 5758

> LBEsummary(LBE.res)

Call:
LBE(pval = golub.pval)

pi0:        0.6022541
Confidence Interval (level=0.95): [0,0.6844972]

Cumulative number of significant calls:
```

```
        <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1   <1
p-value    415    731  1287   1654  2053 2602 7129
q-value    236    415   858   1154  1505 2008 7129
```

References

[1] Benjamini Y, Hochberg Y. (1995) Controlling the false discovery rate : a practical and powerful approach to multiple testing. J R Stat Soc Ser B, 57, 289-300.

[2] Dalmasso, C; Broet, P.; Moreau, T. (2005) A simple procedure for estimating the false discovery rate. Bioinformatics. Bioinformatics, 21: 660 - 668.

[3] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD and Lander ES (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science, 531-537.

[4] Storey JD. (2001) A direct approach to false discovery rates. J R Stat Soc Ser B; 64, 479-498.

[5] Storey JD, Tibshirani R. (2003b) Statistical significance for genome-wide studies. Proc Natl Acad Sci, 100, 9440-9445.