XI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 24–25–26 de outubro de 2013

A new R Package for Multiple Comparisons Problems

Irene Castro-Conde[1], Jacobo de Uña-Álvarez[1,2]

[1]Grupo SiDOR, Universidad de Vigo.
[2]Departamento de Estadística e I.O., Universidad de Vigo

ABSTRACT

In this work we introduce a new R package called **sgof** which is devoted to the multiple hypotheses testing problems. The principal aim of this package is to implement, for the first time, the SGoF (Carvajal et al., 2009) and BB-SGoF (de Uña-Álvarez, 2012) multitesting procedures which have been proved to be more powerful than the classical FDR- and FWER-based methods in many situations, specially when the number of tests is large. BH (Benjamini and Hochberg, 1995) false discovery rate controlling procedure is also implemented. Number of rejections, FDR and adjusted p-values are implemented, as well as, some plots of interest. Finally, we give an example of the performance of this package.

Keywords: R, SGoF, BB-SGoF, correlated tests, FDR, multiple testing.

1. INTRODUCTION

The interest in the multiple testing problems has grown in the last years, since the advent of the Omic's Sciences like genomics and protenomics. In these areas, often thousands of null hypotheses are tested simultaneously, producing as a result a number of significant p-values or effects. Moreover, these hypotheses may have complex and unknown dependence structure among themselves. One of the main problems in multiple hypotheses testing is that, if one does not take the multiplicity of tests into account, then the probability that some of the true null hypotheses are rejected may be overly large. So, in the multitesting setting, a specific procedure for deciding which null hypotheses should be rejected is needed.

In this context, the family-wise error rate (FWER) and the false discovery rate (FDR), among other measures, have been proposed as suitable significance criteria to perform the multiple testing but their power may be rapidly decreased as the number of tests grows, being unable to detect even one effect in particular situations. See Benjamini and Hochberg (1995) or Dudoit and Van der Laan (2008) for basic definitions and reviews of existing literature.

The purpose of the **sgof** package is to provide the scientific community with an alternative tool of multiple comparisons by implementing some new methods, namely SGoF and BB-SGoF (see Carvajal et al., 2009, de Uña-Álvarez, 2011 and de Uña-Álvarez, 2012 for more information), which have been proved to be more powerful in many of the situations and specially when the number of tests is large (Castro-Conde and de Uña-Álvarez, 2013). For completeness, the well-known Benjamini-Hochberg (BH) procedure is also implemented. The **sgof** package is available from the Comprehensive R Archive Network (CRAN) at http://cran.r-project.org/packages/sgof and the reader can install the package directly from the R prompt via: install.packages("sgof"). **sgof** depends on R ($\geq$ 2.15.2) and the package **stats** R.

We end the introduction by listing some of the existing R packages for multiple comparisons.

1. The **stats** package includes the function *p.adjust* which, given a set of p-values, returns the adjusted p-values using one of several methods (Holm 1979, Hochberg 1988, Hommel 1988, Benjamini and Hochberg 1995).

2. The **multtest** package performs non-parametric bootstrap and permutation resampling-based multiple testing procedures for controlling the FWER, the generalized family-wise error rate (gFWER), the proportion os false positives (TPPFP), and the FDR. Results are reported in terms of adjusted p-values, confidence regions and test statistic cutoffs. The procedures are directly applicable to identifying differentially expressed genes in DNA microarray experiments.

3. The **siggenes** package performs multiple testing using Significance Analysis of Microarrays (SAM) and Efron's empirical Bayes approaches. Identification of differentially expressed genes and estimation of the FDR using both kind of approaches are implemented.

The rest of the paper is organized as follows. In Section 2, we show how SGoF, BBSGoF and BH methods are implemented in the package **sgof**. An example of the use of this package is reported in Section 3 and, finally, Section 4 contains the main conclusions of this work.

2. IMPLEMENTATION

The **sgof** package consists of 3 main functions: SGoF, BBSGoF and BH. Each one of these functions implements the corresponding method. All of them estimate the false discovery rate by the simple method proposed by: Dalmasso, Bort and Moreau (2005) by taking a=1 in their formula:

$$\pi_0(n) = \{\frac{1}{n}\sum_{i=1}^{n}[-\ln(1-p_i)]^a/\Gamma(a+1)\}, \quad (1)$$

where n is the length of the vector of p-values $\{p_i\}_{i=1}^n$ and the quantity $\pi_0(n)$ is the estimated proportion of true null hypotheses. The structure and performance of these 3 functions are detailed below. You can see the **sgof** manual for more information.

The **SGoF** function performs the Sequential-Goodness-of-Fit method for multiple hypotheses testing. It has the following 3 arguments:

```
SGoF(u, alpha = 0.05, gamma = 0.05)
```

- u: A (non-empty) numeric vector of p-values.
- alpha: The significance level of the metatest.
- gamma: The p-value threshold, so SGoF looks for significance in the amount of p-values below gamma.

The **BBSGoF** function performs the automatic version (automatic choice of the number of blocks) of the Beta-Binomial SGoF method for multiple hypotheses testing. This method works similarly to SGoF, but assuming the existence of a number of k independent blocks of tests. The within-block correlation comes from the randomness of the probability of that a p-value falls below *gamma* which is assumed to follow a Beta distribution. The function **BBSGoF** has the following 8 arguments:

```
BBSGoF(u, alpha = 0.05, gamma = 0.05, kmin = 2, kmax = min(length(u)%/%10, 100),
tol = 10, adjusted.pvalues = FALSE, blocks = NA)
```

- u: A (non-empty) numeric vector of p-values.
- alpha: The significance level of the metatest.
- gamma: The p-value threshold, so BBSGoF looks for significance in the amount of p-values below gamma.

- kmin: The smallest allowed number of blocks of correlated tests.
- kmax: The largest allowed number of blocks of correlated tests.
- tol:The tolerance in model fitting.
- adjusted.pvalues: Default is FALSE. A variable indicating whether to compute the adjusted p-values.
- blocks: The number of existing blocks.

The **BBSGoF** function consists of 2 functions: **BBSGoF.ap** which computes the adjusted p-values if the user requires it (*adjusted.pvalues=TRUE*), for the number of blocks specified, and **bbsgof** which computes the remaining results. The argument *tol* allows for a stronger (small *tol*) or weaker (large *tol*) criterion when removing poor fits of the beta-binomial model. When the variance of the estimated beta-binomial parameters for a given k is larger than *tol* times the median variance along *k=kmin,...,kmax*, the particular value of k is discarded. This function usually returns a warning message indicating which blocks k are removed because they provided negative or atypical variances.

The **BH** function performs the Benjamini-Hochberg FDR-controlling method for multiple hypotheses testing. It has the following 2 arguments:

```
BH(u, alpha = 0.05)
```

- u: A (non-empty) numeric vector of p-values.
- alpha: The significance level of the test.

For each of these three classes of functions there are 3 methods. The **print.method** function which prints the results in a nice way, the **summary.method** function which prints a summary of the main results reported by the function, and the **plot.method** function which provides some graphical representations.

3.AN ILLUSTRATIVE EXAMPLE

We consider the micro array study of hereditary breast cancer of Hedenfalk et al. (2001). One of the goals of this study was to find genes differentially expressed between BRCA1- and BRCA2-mutation positive tumors. Thus, for each of the 3,226 genes of interest, a p-value was assigned based on a suitable statistical test for the comparison. Following previous analysis of these data, 56 genes were eliminated. This left n =3,170 genes. The independence assumption among the tests was checked through the runs test, giving a two-sided p-value of 0.002654 indicating a significant positive dependence among the tests. The set of these p-values is included in the **sgof** package, and it is named *Hedenfalk*.

The first step to analyze the *Hedenfalk* data is to load the package and the data set. For this we use the next sentences:

```
R>library("sgof")
R>data("Hedenfalk")
```

As an illustrative example we will run the **BBSGoF** function. We start by the default implementation, which does not compute the adjusted p-values.

```
R>  m2<-BBSGoF(u=Hedenfalk$x)
R>  m2

Call:
BBSGoF(u = Hedenfalk$x)
```

```
Parameters:
alpha= 0.05
gamma= 0.05
kmin= 2
kmax= 100

Warning:
Blocks 2 3 4 5 6 7 8 9 11 15 18 19 have been removed because they provided
negative or atypical variances.

Rejections:
[1] 393.506

R> summary(m2)

Call:
BBSGoF(u=Hedenfalk$x)

Parameters:
alpha= 0.05
gamma= 0.05
kmin= 2
kmax=100

$Rejections
[1] 393.506

$FDR
[1] 0.13

$Tarone.pvalue.auto
[1] 0.0005118884

$beta.parameters
[1]  35.04046 148.41387

$betabinomial.parameters
[1] 0.191003717 0.005421396

$sd.betabinomial.parameters
[1] 0.010625190 0.003784251

$automatic.blocks
[1] 13
```

The **BBSGoF** function reports 393 effects with a FDR of 13%. The p-value of the Tarone's test shows that we can reject the null hypothesis of uncorrelated tests. The summary shows other results provided by this function. Figure 1 shows some plots of interest which are obtained as follows:

```
R> plot(m2)
```

In the upper left plot Tarone's test p-values are represented. We can see that there are many p-values falling below 0.05 so this suggest that there is a trend of positive correlation. In the upper right plot we see the estimated within-block correlation for each number of blocks which is positive and increases when the number of blocks k increases. In the lower left plot we represent the fitted beta density where the dashed line corresponds to the mean probability that a p-value falls below

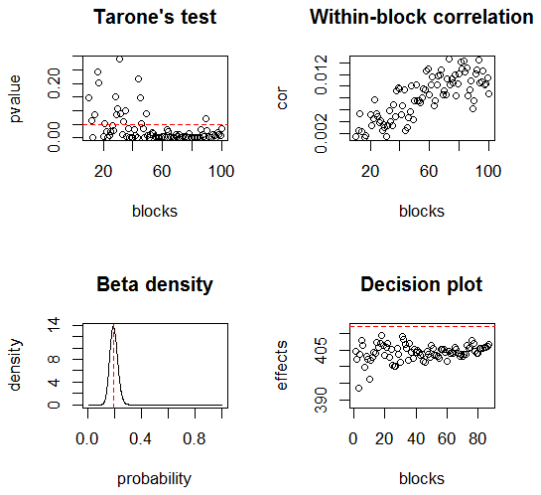

Figure 1: Plots of interest of the BB-SGoF procedure.

gamma (0.05). The mean value of the beta model (`m2$betabinomial.parameters[1]`) is above the expected value (*alpha*=0.05) under the intersection null of no effects. This indicates that some of the p-values could correspond to non-true nulls. On the other hand, the dispersion of the beta model is responsible of the within-block correlation among the indicators $I(p_i \leq gamma)$. Finally, the remaining figure shows the number of effects for each choice of the number of blocks.

Now, we choose the option *adjusted.pvalues=TRUE* and *blocks=13* to compute the adjusted p-values in the case of the existence of 13 independent blocks. Adjusted p-values for SGoF or BB-SGoF methods are defined as the minimum values of *alpha=gamma* at which the corresponding null hypothesis is rejected. These are the results obtained:

```
R>  m22<-BBSGoF(u=Hedenfalk$x,adjusted.pvalues=T,blocks=13)
R>  m22

Call:
BBSGoF(u = Hedenfalk$x, adjusted.pvalues = T, blocks = 13)

Parameters:
alpha= 0.05
gamma= 0.05
kmin= 2
kmax= 100

Warning:
Blocks 2 3 4 5 6 7 8 9 11 15 18 19 have been removed because
they provided negative or atypical variances.

Rejections:
[1] 393.506

R> summary( m22)

Call:
BBSGoF(u = Hedenfalk$x, adjusted.pvalues = T, blocks = 13)

Parameters:
alpha= 0.05
```

```
gamma= 0.05
kmin= 2
kmaobject=

 $Rejections
[1] 393.506

 $FDR
[1] 0.13

 $Adjusted.pvalues
 >gamma <=gamma
   2777     393

 $Tarone.pvalue.auto
[1] 0.0005118885

 $beta.parameters
[1]  35.04043 148.41374

 $betabinomial.parameters
[1] 0.191003721 0.005421401

 $sd.betabinomial.parameters
[1] 0.010625193 0.003784255

 $automatic.blocks
[1] 13
```

The summary shows that there are 393 adjusted p-values falling below *gamma* which matches with the number of rejections reported by the **BBSGoF** function, as it should be in theory. Figure 2 shows the plot of the adjusted p-values versus the original ones. It is seen, as expected, that the adjusted p-values are greater than the original ones. Besides, all the p-values above `sort(u)[sum(m22$Adjusted.pvalues!1)]`= 0.05393691 report adjusted p-values equal to 1; this means that the corresponding null hypotheses are accepted regardless the particular values of *alpha=gamma* at which BB-SGoF method is performed.

```
R> plot(m22)
```

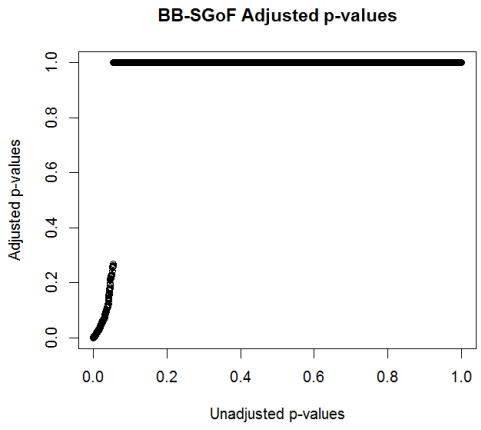

Figure 2: Adjusted p-values reported by BBSGoF method versus the original ones.

4. CONCLUSIONS

In this paper we have introduce the *sgof* package to implement three different methods for solving multitesting problems: BH, SGoF and BB-SGoF. We have explained how they are implemented. Moreover, we have shown an example of application to illustrate how *sgof* works, using the Hedenfalk data. Finally, we hope that the *sgof* package will be useful to the community, by providing a simple and powerful tool for solving multiple hypotheses problems.

ACKNOWLEDGEMENTS

Financial support from the Grant MTM2011-23204 (FEDER support included) of the Spanish Ministry of Science and Innovation is acknowledged.

REFERENCES

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A Practical and Powerful approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological), 57, 1, 289–300.

Carvajal-Rodríguez, A. and De Uña-Álvarez, J. and Rolán-Álvarez, E. (2009). A new multitest correction (SGoF) that increases its statistical power when increasing the number of tests. BMC Bioinformatics, 10, 209.

Castro-Conde, I. and de Uña-Álvarez, J. (2013). Performance of Beta-Binomial SGoF multitesting method for dependent gene expression levels: a simulation study. Proceedings of BIOINFORMATICS 2013 International Conference on Bioinformatics Models, Methods and Algorithms, 97–93.

Dalmasso, C., Bort, P. and Moreau, T. (2005). A simple procedure for estimating the false discovery rate. Bioinformatics, 21, 5, 660–668.

De Uña-Álvarez, J. (2012). The Beta-Binomial SGoF method for multiple dependent tests. Statistical Applications in Genetics and Molecular Biology, 11, 14.

Dudoit, S. and Van der Laan, M.J. (2008). Multiple Testing Procedures with Applications to Genomics. New York: Springer.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M. et al. (2001). Gene-Expression Profiles in Hereditary Breast Cancer. New England Journal of Medicine 344, 539–548.

Hochberg, Y. (1988). A Sharper Bonferroni Procedure for Multiple Tests of Significance. Biometrika, 75, 4, 800–802.

Hochberg, Y. and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. Statistics in Medicine, 9, 811–818.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics, 6, 2, 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika, 75, 2, 383.