Samroc example

Per Broberg

April 4, 2013

Analysis of the data from Golub *et al.*

Consider the microarray experiment in Golub et al. (1999) where ALL and AML subtypes of leukemia are compared. The data are available within package *multtest*.

We can analyse those data in *SAGx* with the function *samrocNboot*. The ideas behind it are presented in Broberg (2003). Briefly, the method relies on a penalised t-test statistica $d = (\bar{x}_1 - \bar{x}_2)/(S + a)$ with fudge factor a Efron et al. (2001). In this case the effect estimated consists of a difference in group means. In general the method can estimate and test one such effect in the presence of explanatory variables such as AGE or GENDER using a linear model. In such a case the function `samrocN` provides a solution. Example code now follows.

```
> library("SAGx")
> library("multtest")
> data(golub)
> set.seed(849867)
> samroc.res <- samrocN(data = golub, formula = ~as.factor(golub.cl))
> show(samroc.res)

Samroc result:
Data: 38 samples with 3051 genes.
Model: ~ as.factor(golub.cl)
Using 100 permutations
Fudge factor: 0 . Estimated proportion unchanged genes: 0.42 .
Annotation: Thu Apr  4 00:05:00 2013
Call: samrocN golub ~as.factor(golub.cl)
```

The function `samrocN` is used to perform a penalised t-test. Its value is an object of class *samroc.result*. The functions `show` and `plot` are defined for such objects. In Figure 1 the densities of the test statistic and its permutation null distribution are displayed. The graph was produced by invoking the `plot` function

```
> plot(samroc.res)
```

```
> par(bg = "cornsilk")
> plot(samroc.res)
```

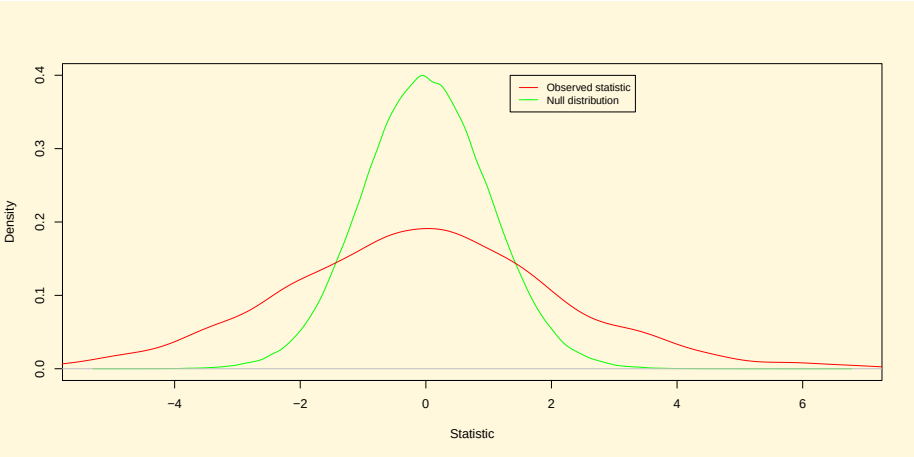

Figure 1: Densities of the test statistic and of its permutation null distribution

One can also perform a simple Gene Set Enrichment Analysis based on the output from `samrocNboot` by invoking `GSEA.mean.t`, cf. Tian et al. (2005) which describes a similar idea. The package *hu6800.db* maps KEGG pathways Kanehisa and Goto (2000) onto probeset identifiers. The following code analyses one KEGG pathway (00970 Aminoacyl-tRNA biosynthesis) and outputs a p-value based on the average over the pathway of the absolute value of the test statistic d. The algorithm includes restandardization following Efron and Tibshirani (2006).

```
> library("hu6800.db")
> kegg <- as.list(hu6800PATH2PROBE)
> probeset <- golub.gnames[,3]
> GSEA.mean.t(samroc = samroc.res, probeset = probeset, pway = kegg[1],
+ type = "original", two.side = FALSE)

      normal p-value mean statistic Wilcoxon p-value median statistic
04610     0.03276032      0.7982671        0.2237629        0.9306652

>
```

The estimated proportion unchanged genes equals 0.42. The distribution of p-values is shown in Figure 2, which confirms that many genes are changed. Furthermore, using the function *pava.fdr* we obtain estimates of the FDR and of the local FDR, see Figure 3. This function is presented in Broberg (2005) and combines the local FDR estimator of Aubert et al. (2004) with Poisson regression (see Efron (2004)) and isotonic regression.

```
> par(bg = "cornsilk")
> hist(samroc.res@pvalues, xlab = "p-value", main ="", col = 'orange', freq  = F)
> print(abline(samroc.res@p0,0, col = 'red'))
```

```
NULL
```

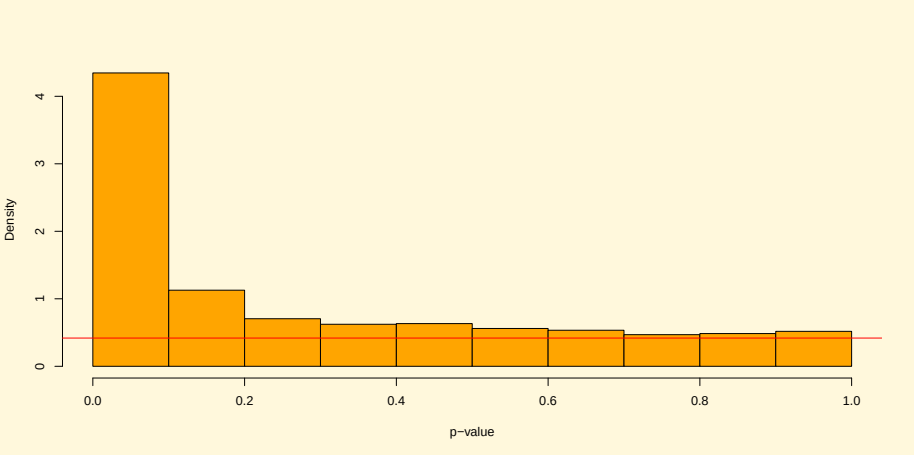

Figure 2: Histogram of the p-values generated by function *samrocNboot*

```
> par(bg = "cornsilk")
> fdrs <- pava.fdr(ps = samroc.res@pvalues)
> plot(samroc.res@pvalues, fdrs$pava.local.fdr, type = 'n', xlab = "p-value", ylab = "False
> lines(lowess(samroc.res@pvalues, fdrs$pava.local.fdr), col = 'red')
> lines(lowess(samroc.res@pvalues, fdrs$pava.fdr), col = 'blue')
> legend(0.1,0.9,pch=NULL,col=c("red","blue"),c("pava local FDR","pava FDR"),lty = 1)
```

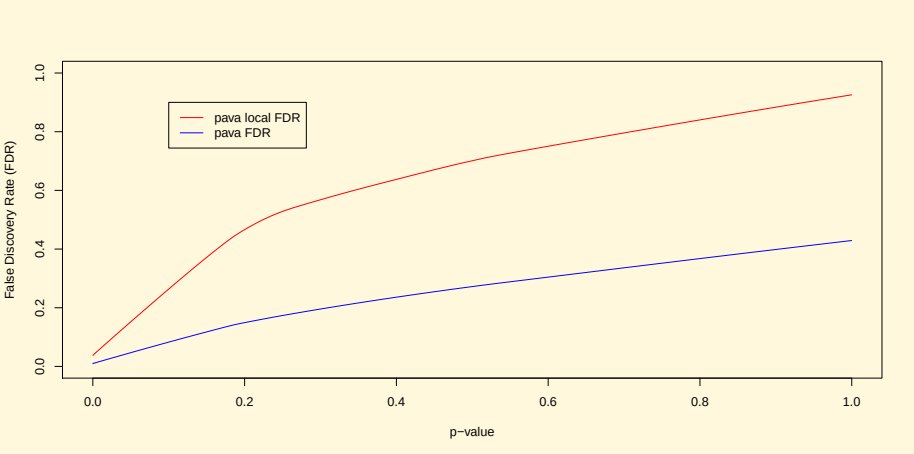

Figure 3: Scatter plot of the local false discovery rate and the false discovery rate as estimated by function *pava.fdr*

1 On the calculation of p-values

Following Tusher et al. (2001), Broberg (2003) defines a permutation p-value for gene i out of a total N as

$$p_i = \frac{\#\{d^{*k}(j)| : |d^{*k}(j)| > |d(i)|\}}{N \times B} \tag{1}$$

, denoting by $d(i)$ the test statistic corresponding to gene i, and by $d^{*k}(i)$ the permutation null statistic in the k^{th} iteration out of a total B.

This has the unfortunate side effect of occasionally returning p-values equal to zero. To solve this the definition from Davison and Hinkley (1997) is employed. Denote by F_n the empirical distribution function of all $-|d^{*k}|$. The estimate then becomes:

$$p_i = \frac{B \times N \times F_n(-|d(i)|) + 1}{B \times N + 1} \tag{2}$$

This follows from $\{t^* \geq t\} \Leftrightarrow \{-t^* \leq -t\}$.

Various functions from SAGx were used in Pierrou et al. (2007).

References

J Aubert, A Bar-Hen, JJ Daudin, and S Robin. Determination of the differentially expressed genes in microarray experiments using local fdr. *BMC Bioinformatics*, 5:125, 2004. doi: http://dx.doi.org/10.1186/1471-2105-5-125. 3

P Broberg. Statistical methods for ranking differentially expressed genes. *Genome Biology*, 4:R41, 2003. doi: http://dx.doi.org/10.1186/gb-2003-4-6-r41. URL http://genomebiology.com/2003/4/6/R41. 1, 5

Per Broberg. A comparative review of estimates of the proportion unchanged genes and the false discovery rate. *BMC Bioinformatics*, 6(1):199, 2005. ISSN 1471-2105. doi: http://dx.doi.org/10.1186/1471-2105-6-199. URL http://www.biomedcentral.com/1471-2105/6/199. 3

A.C. Davison and D.V Hinkley. *Bootstrap Methods and Their Application*. Cambridge University Press, 1997. 5

Bradley Efron. Selection and estimation for large-scale simultaneous inference. preprint available at the `http://www-stat.stanford.edu/ brad/papers/Selection.pdf`, 2004. 3

Bradley Efron and Robert Tibshirani. On testing of the significance of sets of genes. Technical report, Stanford Statistics, 2006. 2

Bradley Efron, Robert Tibshirani, John Storey, and Victoria Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 2001. 1

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439):531–537, 1999. doi: 10.1126/science.286.5439.531. URL http://www.sciencemag.org/cgi/content/abstract/286/5439/531. 1

Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl. Acids Res.*, 28(1):27–30, 2000. doi: 10.1093/nar/28.1.27. URL http://nar.oxfordjournals.org/cgi/content/abstract/28/1/27. 2

Stefan Pierrou, Per Broberg, Rory A. O'Donnell, Krzysztof Pawlowski, Robert Virtala, Eva Lindqvist, Audrey Richter, Susan J. Wilson, Gilbert Angco, Sebastian Moller, Hakan Bergstrand, Witte Koopmann, Elisabet Wieslander, Per-Erik Stromstedt, Stephen T. Holgate, Donna E. Davies, Johan Lund, and Ratko Djukanovic. Expression of Genes Involved in Oxidative Stress Responses in Airway Epithelial Cells of Smokers with

Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Crit. Care Med.*, 175(6):577–586, 2007. doi: 10.1164/rccm.200607-931OC. URL http://ajrccm.atsjournals.org/cgi/content/abstract/175/6/577. 5

Lu Tian, Steven A. Greenberg, Sek Won Kong, Josiah Altschuler, Isaac S. Kohane, and Peter J. Park. Discovering statistically significant pathways in expression profiling studies. *PNAS*, 102(38):13544–13549, 2005. doi: 10.1073/pnas.0506577102. URL http://www.pnas.org/cgi/content/abstract/102/38/13544. 2

Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98(9):5116–5121, 2001. doi: 10.1073/pnas.091062498. URL http://www.pnas.org/cgi/content/abstract/98/9/5116. 5