Samroc example

Per Broberg

October 17, 2016

Analysis of the data from Golub *et al.*

Consider the microarray experiment in **?** 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 **?**. Briefly, the method relies on a penalised t-test statistica $d = (\bar{x}_1 - \bar{x}_2)/(S+a)$ with fudge factor a **?**. 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: Mon Oct 17 16:16:35 2016
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 **??** 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)
```

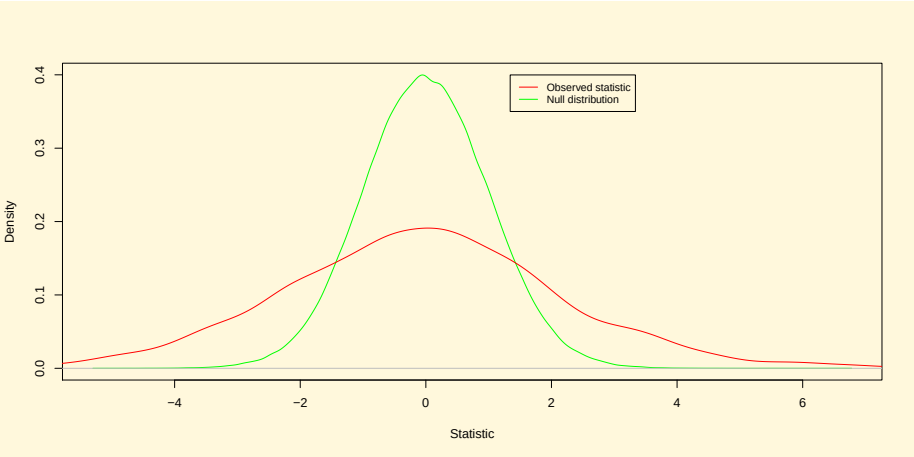

Figure 1: Densities of the test statistic and of its permutation null distribution

One can also perform a simple Gene Set Enrichment Analysis based on the output from `samrocNboot` by invoking `GSEA.mean.t`, cf. **?** which describes a similar idea. The package *hu6800.db* maps KEGG pathways **?** 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 **?**.

```
> 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 **??**, 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 **??**. This function is presented in **?** and combines the local FDR estimator of **?** with Poisson regression (see **?**) 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
```

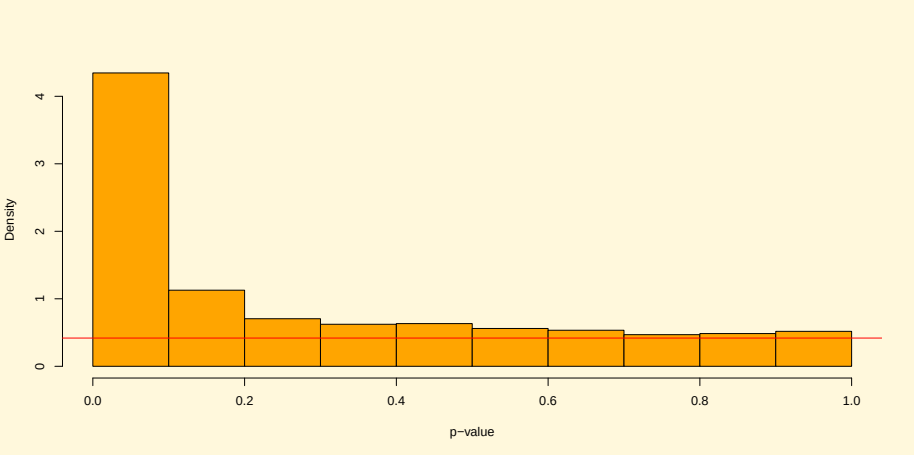

Figure 2: Histogram of the p-values generated by function *samrocNboot*

```
> 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)
```

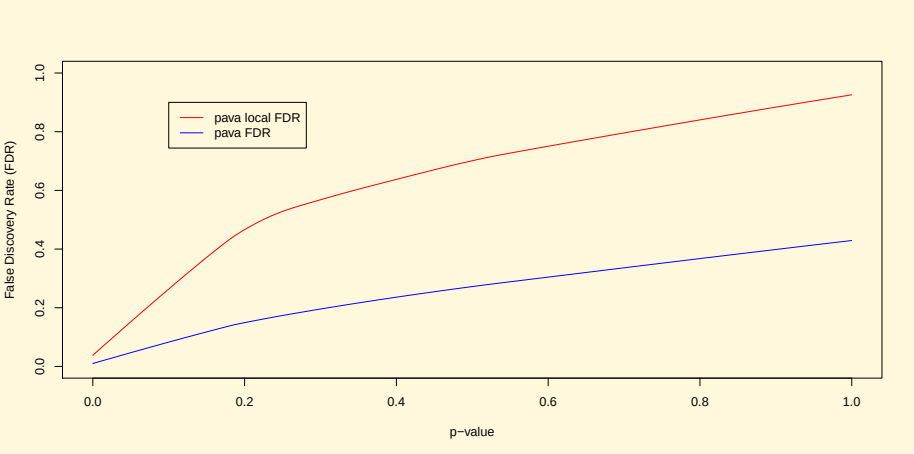

Figure 3: Scatter plot of the local false discovery rate and the false discovery rate as estimated by function *pava.fdr*

1 On the calculation of p-values

Following **?**, **?** 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 **?** 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 **?**.