SAM

Guillermo Ayala Gallego

2024-04-09

Table of contents

Método SAM

- Es la abreviatura de Significance Analysis of Microarrays.
- El conjunto de muestras es $\{1, \dots, n\}$.
- Empezamos con el caso en que pretendemos comparar dos grupos.
- Para un gen dado las expresiones del grupo 1 (o del grupo 2) corresponden con los valores donde $y_j = 1$ (respectivamente $y_j = 2$).
- Tendremos J_1 y J_2 donde J_g son las muestras del grupo g con g= 1,2.
- Tenemos

$$\bar{x}_{ig} = \frac{\sum_{j \in J_g} x_{ij}}{n_g}$$

siendo n_g el número de muestras en J_g.

- Para el i-ésimo gen consideramos la **diferencia relativa** dada por

$$d_i = \frac{\bar{x}_{i1} - \bar{x}_{i2}}{s_i + s_0}$$

donde

$$s_i = \sqrt{a \sum_{g=1}^{2} \sum_{j \in J_g} (x_{ij} - \bar{x}_{ig})^2}$$

y

$$a = \frac{1/n_1 + 1/n_2}{n_1 + n_2}$$

- El estadístico d_i es el t-estadístico al que le modificamos el error estándar.
- Si el valor s_i es muy pequeño entonces el valor d_i es muy grande.
- Los distintos valores d_i no son comparables para los distintos genes, no son **intercambiables**.
- Por ello se le suma un valor s_0 que **estabiliza** la varianza (y nos permite comparar los t-valores entre si).
- Este valor hay que **estimarlo** o calcularlo a partir de los propios datos.

El procedimiento

1. Calculamos los estadísticos d_i y los ordenamos:

$$d_{(1)} \leq ... \leq d_{(N)}.$$

2. Realizamos B permutaciones aleatorias del vector y que nos indica el grupo de la columna.
3. Para cada permutación, b, calculamos los nuevos valores d_i y los denotamos por $d_i(b)$.
4. Ordenamos

$$d_{(1)}(b) \leq ... \leq d_{(N)}(b).$$

5. Calculamos, para las B permutaciones, el valor medio de estos estadísticos ordenados.

$$\bar{d}_{(i)} = \sum_{b=1}^{B} \frac{d_{(i)}(b)}{B}.$$

6. Para un valor fijo positivo Δ determinamos:

   - Empezando en el origen y moviéndonos a la derecha determinamos el primer $i = i_1$ tal que

$$d_{(i)} - \bar{d}_{(i)} > \Delta.$$

     Todos los genes que verifican esta desigualdad son llamados **significativamente positivos**.

- Empezando en el origen y moviéndonos a la izquierda determinamos el primer $i = i_2$ tal que
$$\bar{d}_{(i)} - d_{(i)} > \Delta.$$
Todos los genes que verifican esta desigualdad los llamamos **significativamente negativos**.
- Para cada Δ definimos un punto de corte superior como
$$u(\Delta) = \min\{d_i : i \text{ es significativamente positivo}\},$$
y
$$l(\Delta) = \max\{d_i : i \text{ es significativamente negativo}\}.$$

7. Para distintos valores de Δ, calculamos:

- El número total de genes significativos.
- En la b-ésima permutación aleatoria hemos obtenido $d_i(b)$ con $i = 1, \dots, N$. Calculamos el número de genes **falsamente llamados** en la aleatorización b-ésima como
$$c_b = |\{i : d_i(b) > u(\Delta) \text{ o } d_i(b) < l(\Delta)\}$$
Tendremos los valores c_b con $b = 1, \dots, B$. Notemos que los valores $d_i(b)$ han sido calculados bajo la hipótesis de no asociación, es cierta la hipótesis nula para cada gen. En consecuencia debieran de estar en el intervalo $[l(\Delta), u(\Delta)]$. Calculamos la mediana y el percentil de orden 0.90 de los valores c_b con $b = 1, \dots, B$ y los denotamos como $q_{0.5}(c)$ y $q_{0.90}(c)$.

8. Estimamos π_0, la proporción de genes sin expresión diferencial.

- Calculamos los percentiles de orden 0.25 y 0.75, $q_{0.25}$ y $q_{0.75}$, de $\{d_i(b) : i = 1, \dots, N; b = 1, \dots, B\}$.
- Calculamos
$$\hat{\pi}_0 = \frac{|\{d_i : d_i \in (q_{0.25}, q_{0.75})\}|}{N/2}$$
donde $\{d_1, \dots, d_N\}$ son los d valores originales.
- Truncamos el valor de $\hat{\pi}_0$ en 1, es decir, consideramos
$$\hat{\pi}_0 = \min\{\hat{\pi}_0, 1\}.$$

9. Multiplicamos la mediana, $q_{0.5}(c)$, y el percentil 0.9, $q_{0.90}(c)$ por $\hat{\pi}_0$.
10. Elegimos un valor de Δ y determinamos el conjunto de genes significativos.
11. La tasa de falsos positivos FDR es estimada como el cociente entre la mediana, $q_{0.5}(c)$, (o el percentil 0.90, $q_{0.90}(c)$) del número de genes falsamente llamados y el número de genes declarados significativos.
12. El **q-valor**: + Es pFDR cuando la lista de genes significativos está compuesta por este gen y todos los genes que son más significativos que este. + Se calcula buscando cuál es el valor más pequeño de Δ para el cual este gen es declarado como significativo. + Si este valor es Δ entonces la pFDR asociada a este valor es el q-valor asociado al gen.

Cálculo de s_0

1. Consideramos $s = (s_1, \dots, s_N)$ las desviaciones estándar muestrales.
2. Sea $q_\alpha(s)$ el percentil de orden α de los valores s_i. Definimos

$$d_i^{(\alpha)} = \frac{\bar{x}_{i1} - \bar{x}_{i2}}{s_i + q_\alpha(s)}.$$

3. Consideramos los percentiles $q_{j/100}(s)$ con $j = 1, \dots, 100$.
4. Para cada $\alpha \in \{ 0, 0.1, \dots, 1\}$:
    1. Se calcula
    $$v_j = \frac{1}{0.64} MAD(\{d_i^{(\alpha)} : s_i \in [q_{j/100}(s), q_{(j+1)/100}(s)]\})$$
    para $j = 1, \dots, 100$ donde MAD es la mediana de las desviaciones absolutas respecto de la mediana
    2. Calculamos el valor $CV(\alpha)$, el coeficiente de variación de los valores v_j.
    3. Se elige como valor de α: $\hat{\alpha}$ que minimiza $CV(\alpha)$.
5. Tomamos para s_0: $\hat{s}_0 = q_{\hat{\alpha}}(s)$.

samr

Cargamos el paquete.

```
pacman::p_load(samr)
```

Cargamos los datos.

```
data(golub,package="multtest")
fac0 = golub.cl +1 ## samr espero valores 1 y 2 para los grupos
```

Aplicamos el método: tenemos dos grupos independientes.

```
samfit = SAM(golub,fac0,resp.type="Two class unpaired",nperms=1000)
```

¿Qué valor de Δ se ha utilizado?

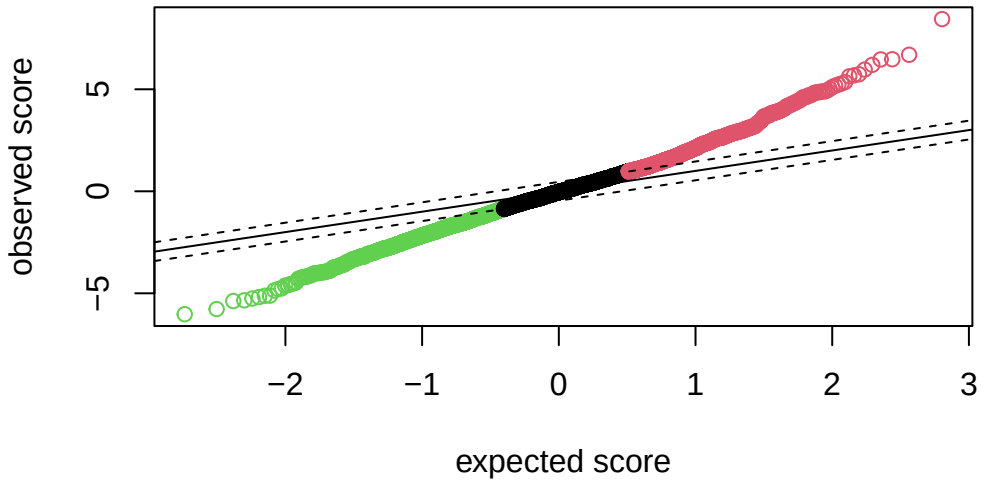
```
samfit$del
```

```
   delta
0.460144
```

```
sigtabla = samfit$siggenes.table
```

Veamos gráficamente el resultado.

```
plot(samfit)
```



Podemos ver también una exploración de posibles valores para Δ con

```
head(samfit$delta.table)
```

```
         delta # med false pos 90th perc false pos # called median FDR
[1,] 0.4601440       356.06686           501.40688     1689 0.21081519
[2,] 0.5282265       258.95772           383.04163     1506 0.17195068
[3,] 0.6010044       176.95444           288.14441     1339 0.13215418
[4,] 0.6784776       117.87971           210.40315     1189 0.09914189
[5,] 0.7606462        76.60833           147.44405     1072 0.07146299
[6,] 0.8475101        45.85710            98.72763      932 0.04920289
     90th perc FDR      cutlo     cuthi
```

```
[1,]     0.2968661 -0.8650472 0.9722594
[2,]     0.2543437 -0.9883173 1.1236393
[3,]     0.2151937 -1.1215032 1.2757362
[4,]     0.1769581 -1.2624441 1.4182065
[5,]     0.1375411 -1.4105379 1.5489360
[6,]     0.1059309 -1.5606679 1.7121587
```

Modificando FDR

Tenemos muchos genes significativos. Parece que podemos ser más exigentes con la tasa FDR.

```
samfit = SAM(golub,fac0,resp.type="Two class unpaired",nperms=1000,fdr.output=.001)
sigtabla = samfit$siggenes.table
```

La tabla resumen sería

```
sigtabla = samfit$siggenes.table
```

Otro tipo de covariables

Comparación de más de dos grupos

- Tenemos K grupos a comparar ($K > 2$) por lo que $y_j \in \{1, \dots, K\}$.
- J_k son los índices de las observaciones en grupo k.
- N_k es el cardinal de J_k ($\sum_{k=1}^{K} n_k = n$).
- Definimos d_i como

$$d_i = \frac{r_i}{s_i + s_0}$$

con

$$r_i = \sqrt{\frac{\sum_{k=1}^{K} n_k}{\prod_{k=1}^{K} n_k} \sum_{k=1}^{K} n_k (\bar{x}_{ik} - \bar{x}_{i.})^2}$$

$$s_i = \sqrt{\frac{1}{\sum_{k=1}^{K} (n_k - 1)} \left(\sum_{k=1}^{K} \frac{1}{n_k} \right) \sum_{k=1}^{K} \sum_{j \in J_k} (x_{ij} - \bar{x}_{ik})^2}.$$

Covariable continua

- Supongamos que los valores de y son valores numéricos.
- Definimos d_i como

$$d_i = \frac{r_i}{s_i + s_0}$$

- Tenemos:

$$r_i = \frac{\sum_{j=1}^{n} y_j (x_{ij} - \bar{x}_{i\cdot})}{\sum_{j=1}^{n} (y_j - \bar{y})^2}$$

y

$$s_i = \frac{\hat{\sigma}_i}{\sqrt{\sum_{j=1}^{n} (y_j - \bar{y})^2}}$$

donde

$$\hat{\sigma}_i = \sqrt{\frac{\sum_{j=1}^{n} (x_{ij} - \hat{x}_{ij})^2}{n-2}},$$

y

$$\hat{x}_{ij} = \hat{\beta}_{i0} + r_i y_j,$$

$$\hat{\beta}_{i0} = \bar{x}_{i\cdot} - r_i \bar{y}.$$