Componentes principales

Guillermo Ayala Gallego

Componentes principales

Guillermo Ayala Gallego

2024-03-25

Componentes principales

El problema

- En este tema nos ocupamos de problemas de **reducción de dimensión**.
- ¿Qué significa reducir la dimensión?
- Trabajando con expresión de genes tenemos tantas filas como genes y tantas columnas como muestras.
- En resumen miles de filas y decenas o centenares de columnas.
- En este tema tratamos el tema de cómo reducir el número de columnas o el número de filas.

Componentes principales con dimensión 2

- Expresión de las dos primeras muestras de los datos golub para aquellos genes que contienen la expresión **Cyclin**.

```
data(golub,package="multtest")
sel = grep("Cyclin",golub.gnames[,2])
golub.red = golub[sel,1:2]
plot(golub.red,xlab="Primera muestra",ylab="Segunda muestra")
```

- Centramos los datos: restamos a cada columna la media de la columna.
- La idea de las componentes principales es considerar una combinación lineal de los valores originales.
- Se pretende elegir un vector (de dimensión dos) \(a_{1} = (a_{11},a_{12})\) de modo que en lugar de utilizar \(x_i\) consideremos (el resumen) \(u_i = a_{11} x_{i1} + a_{12} x_{i2}\).

- ¿Qué \(a_1\) elegimos?
- La idea es lograr que los valores \(u_i\) tengan la mayor variabilidad que se pueda con objeto de no perder información.
- En concreto se elige \(a_1\) de modo que maximizamos \[\frac{1}{n}\sum_{i=1}^n (u_i - \bar{u}) ^2. \]

prcomp

- Obtengamos las **componentes principales**.

```
a.pca = prcomp(golub.red)
```

- Los vectores directores de las líneas sobre las que proyectamos aparecen en la siguiente figura.
- Estos vectores son

```
a.pca$rotation
```

```
            PC1        PC2
[1,] -0.7619878  0.6475914
[2,] -0.6475914 -0.7619878
```

- Mostrando las líneas sobre las que proyectamos (en azul) para obtener las dos componentes principales.
- Mostrando las líneas sobre las que proyectamos (en azul) para obtener las dos componentes principales y las proyecciones, esto es, la **primera componente** en verde.
- Y ahora consideremos la segunda componente principal.
- Los valores de estas proyecciones (o componentes principales) los obtenemos con

```
predict(a.pca)
```

```
              PC1           PC2
 [1,] -2.50309193 -1.541823e-01
 [2,]  0.01368602 -2.024163e-01
 [3,] -2.38702381  3.714339e-03
 [4,]  0.33489688 -6.847077e-05
 [5,]  0.76608286  2.806154e-01
 [6,]  0.27144878  2.899820e-02
 [7,]  0.31169639 -2.876394e-01
 [8,]  2.22052303 -8.232084e-02
 [9,] -0.93221244  1.836866e-01
[10,] -0.39946389 -7.239549e-03
[11,]  0.08293509  3.191733e-01
[12,]  2.22052303 -8.232084e-02
```

Componentes principales de los datos golub

```
golub.pca = prcomp(golub,scale=FALSE,center=TRUE)
```

- El argumento **center=TRUE** centra los datos restando la media de la columna de modo que las variables tengan medias nulas.
- El argumento **scale=TRUE** hace que las variables originales sean divididas por su desviación estándar de modo que la varianza (y la desviación estándar) de las nuevas variables sea la unidad.

Criterios de selección del número de componentes

- Uno puede ser la **proporción total explicada**. Fijar un nivel mínimo y quedarnos con el número de componentes necesario para superar este valor mínimo.
- El segundo puede ser que una componente no puede tener una desviación estándar menor que una de las variables originales. **Si hemos escalado** cada variable original dividiendo por su desviación estándar entonces la desviación estándar de cada componente ha de ser mayor que uno.
- Otro criterio puede ser ver en qué momento se produce un descenso de la desviación estándar muy notable. Quedarnos con las componentes previas.
- Un resumen de las componentes nos puede indicar con cuántas nos quedamos.

```
summary(golub.pca)
```

```
Importance of components:
                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     5.0436 1.44073 1.11734 1.03505 0.85821 0.74399 0.72104
Proportion of Variance 0.6694 0.05462 0.03285 0.02819 0.01938 0.01457 0.01368
Cumulative Proportion  0.6694 0.72405 0.75691 0.78510 0.80448 0.81905 0.83273
                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     0.69232 0.63819 0.63630 0.56700 0.55263 0.53868 0.52011
Proportion of Variance 0.01261 0.01072 0.01065 0.00846 0.00804 0.00764 0.00712
Cumulative Proportion  0.84534 0.85606 0.86672 0.87518 0.88321 0.89085 0.89797
                          PC15    PC16    PC17    PC18    PC19    PC20    PC21
Standard deviation     0.49568 0.48402 0.47719 0.47068 0.45421 0.43795 0.43410
Proportion of Variance 0.00647 0.00617 0.00599 0.00583 0.00543 0.00505 0.00496
Cumulative Proportion  0.90443 0.91060 0.91659 0.92242 0.92785 0.93290 0.93786
                          PC22    PC23    PC24    PC25   PC26    PC27    PC28
Standard deviation     0.42475 0.41582 0.40718 0.40066 0.3948 0.38731 0.38417
Proportion of Variance 0.00475 0.00455 0.00436 0.00422 0.0041 0.00395 0.00388
Cumulative Proportion  0.94260 0.94715 0.95152 0.95574 0.9598 0.96379 0.96767
                          PC29    PC30    PC31   PC32   PC33    PC34    PC35
Standard deviation     0.37882 0.37124 0.36957 0.3596 0.3593 0.35276 0.34218
Proportion of Variance 0.00378 0.00363 0.00359 0.0034 0.0034 0.00327 0.00308
```

```
Cumulative Proportion  0.97145 0.97508 0.97867 0.9821 0.9855 0.98875 0.99183
                          PC36    PC37    PC38
Standard deviation     0.33228 0.32572 0.30667
Proportion of Variance 0.00291 0.00279 0.00247
Cumulative Proportion  0.99473 0.99753 1.00000
```

- Si fijamos, por ejemplo un 80% de la variación total a explicar entonces debemos quedarnos con las cinco primeras componentes.

```
a = predict(golub.pca)
a = a[,1:5]
```

- Podemos representar, como es habitual, las dos primeras componentes.

```
plot(a[,1],a[,2],xlab="Primera componente",ylab="Segunda componente")
```

Interpretación

- Es interesante observar los valores del vector asociado a la primera componente.

```
golub.pca$rotation[,1]
```

```
 [1] 0.1715179 0.1690829 0.1650131 0.1726783 0.1659431 0.1668800 0.1686381
 [8] 0.1602445 0.1648769 0.1687936 0.1653992 0.1694389 0.1629073 0.1661268
[15] 0.1647691 0.1720833 0.1559293 0.1600159 0.1677201 0.1491867 0.1272725
[22] 0.1620961 0.1643597 0.1652554 0.1659262 0.1690494 0.1539691 0.1689052
[29] 0.1541333 0.1516988 0.1691436 0.1682306 0.1452419 0.1675335 0.1638384
[36] 0.1508645 0.1476137 0.1520465
```

- Podemos ver que son coeficientes muy parecidos y (casi) todos positivos.
- Básicamente tenemos la media muestral de todos los niveles de expresión en las 38 muestras.
- La primera componente es casi media sobre las 38 muestras.
- ¿Y la segunda componente?

```
golub.pca$rotation[,2]
```

```
 [1]  0.104190349 -0.036887376  0.069108679  0.100701406  0.170952497
 [6]  0.028349013  0.032390592  0.000505933  0.093593873  0.023532773
[11]  0.075375878 -0.089380731  0.233399832  0.077938472  0.237951078
[16]  0.184071755  0.078196661  0.041608386  0.114629249  0.247148154
[21]  0.201580365 -0.014147623  0.037858911  0.210585781 -0.044465104
[26]  0.122286768  0.021439090 -0.189278987 -0.174593342 -0.243775839
[31] -0.165316096 -0.150156242 -0.344034501 -0.157687744 -0.130649728
[36] -0.277921375 -0.344828851 -0.222765782
```

- Si observamos los coeficientes vemos que las primeros 27 valores son positivos y los 11 últimos son negativos.

- Además no hay una gran diferencia entre los 27 primeros y tampoco entre los 11 últimos.
- Básicamente (aunque no exactamente) estamos comparando, para cada gen, la media de los niveles de expresión sobre los datos ALL (leucemia linfoblástica aguda) con la media sobre los datos AML (leucemia mieloide aguda).

¿Componentes principales de las muestras o de los genes?

Genes (características)

- Podemos realizar un análisis de componentes principales de los genes.

```
tgolub.pca = prcomp(t(golub),scale=FALSE,center=TRUE)
summary(tgolub.pca)
```

```
Importance of components:
                           PC1      PC2     PC3    PC4     PC5     PC6     PC7
Standard deviation     13.0934 10.17462 9.40357 7.9010 6.82616 6.62780 6.30435
Proportion of Variance  0.1645  0.09934 0.08485 0.0599 0.04471 0.04215 0.03814
Cumulative Proportion   0.1645  0.26385 0.34870 0.4086 0.45332 0.49547 0.53361
                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     5.83194 5.79413 5.15726 5.01893 4.90719 4.72354 4.50857
Proportion of Variance 0.03264 0.03222 0.02552 0.02417 0.02311 0.02141 0.01951
Cumulative Proportion  0.56625 0.59846 0.62398 0.64816 0.67126 0.69267 0.71218
                          PC15    PC16    PC17    PC18    PC19    PC20    PC21
Standard deviation     4.40036 4.34750 4.27398 4.12411 3.98196 3.94862 3.85795
Proportion of Variance 0.01858 0.01814 0.01753 0.01632 0.01522 0.01496 0.01428
Cumulative Proportion  0.73076 0.74890 0.76643 0.78275 0.79796 0.81292 0.82721
                          PC22    PC23    PC24    PC25    PC26    PC27    PC28
Standard deviation     3.77535 3.69767 3.64804 3.58436 3.51711 3.49260 3.44527
Proportion of Variance 0.01368 0.01312 0.01277 0.01233 0.01187 0.01171 0.01139
Cumulative Proportion  0.84088 0.85400 0.86677 0.87910 0.89097 0.90268 0.91407
                          PC29    PC30    PC31    PC32    PC33    PC34    PC35
Standard deviation     3.37281 3.35604 3.26862 3.26192 3.21319 3.10703 3.01784
Proportion of Variance 0.01092 0.01081 0.01025 0.01021 0.00991 0.00926 0.00874
Cumulative Proportion  0.92498 0.93579 0.94604 0.95625 0.96616 0.97543 0.98416
                          PC36    PC37      PC38
Standard deviation     2.95738 2.78502 5.815e-15
Proportion of Variance 0.00839 0.00744 0.000e+00
Cumulative Proportion  0.99256 1.00000 1.000e+00
```

```
library(affycoretools)
plotPCA(t(golub),legend = FALSE)
```

Dos primeras componentes de las muestras

```
affycoretools::plotPCA(golub,legend = FALSE)
```

- Cuando realizamos un análisis de los genes sabemos que las observaciones están clasificadas (las muestras son de dos tipos de leucemia).
- La función **plotPCA** está preparada para mostrarnos si las componentes nos **reproducen** los grupos que sabemos que previamente tenemos.
- En la siguiente figura mostramos las dos primeras componentes pero diferenciando el tipo de muestra (según el tipo de leucemia).

```
tipo = factor(golub.cl+1,levels = 1:2,labels = c("ALL","AML"))
plotPCA(golub,groups = tipo, groupnames = tipo,legend=FALSE)
```

gse20986

```
pacman::p_load("Biobase")
data(gse20986,package="tamidata")
```

```
plotPCA(gse20986,groups = pData(gse20986)[,"tissue"],
        groupnames = levels(pData(gse20986)[,"tissue"]))
```

gse1397

Los datos

```
data(gse1397,package="tamidata")
pData(gse1397)[,c("tissue","type")]
```

```
       tissue    type
1    Cerebrum Euploid
2    Cerebrum Euploid
3    Cerebrum Euploid
4    Cerebrum Euploid
5  Cerebellum Euploid
6  Cerebellum Euploid
7  Cerebellum Euploid
8    Cerebrum    TS21
9    Cerebrum    TS21
10   Cerebrum    TS21
11   Cerebrum    TS21
12 Cerebellum    TS21
13 Cerebellum    TS21
14 Cerebellum    TS21
```

Tipo

```
plotPCA(gse1397, groups = pData(gse1397)[,'type'],
        groupnames = levels(pData(gse1397)[,'type']))
```

Tejido

```
plotPCA(gse1397, groups = pData(gse1397)[,'tissue'],
        groupnames = levels(pData(gse1397)[,'tissue']))
```