Técnicas de reducción de la dimensión con datos ómicos

Francesc Carmona y Alex Sánchez

5/7/2021

Introducción

En estadística, las técnicas para la reducción de la dimensión son los procesos para reducir el número de variables observadas a un conjunto de pocas (habitualmente dos o tres) variables "principales".

En el contexto de una matriz de datos de expresión para miles de genes en una muestra de pacientes, el objetivo es obtener unas pocas variables (dos si es posible) que cubran la variación en conjunto de los genes.

Estas técnicas son útiles para la visualización de los datos, el análisis de conglomerados o la modelización predictiva.

Análisis de componentes principales

El análisis de componentes principales (PCA) es una de las técnicas más populares.

La interpretación geométrica es la siguiente:

- El objetivo del PCA es hallar la dirección de máxima dispersión de los datos. Esa dirección o eje es una combinación lineal de las variables que llamamos primera componente principal.
- La segunda componente es otra combinación lineal de las variables ortogonal a la anterior y también en la dirección de máxima variabilidad.
- Y así sucesivamente

Solución del PCA

La solución se obtiene mediante la descomposición en valores propios de la matriz de covarianzas (o de correlaciones).

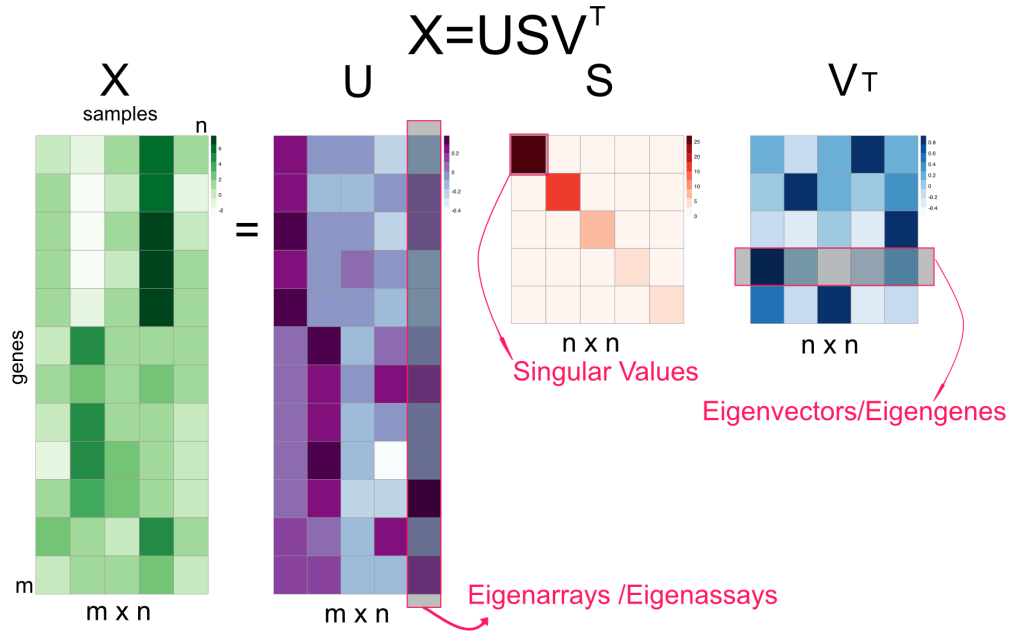
En R

```
prcomp(t(mat), center=TRUE, scale=TRUE)
eigen(cor(t(mat)))
```

Recordemos que los genes (variables) estan en las filas de la matriz.

Descomposición en valores singulares

Otra forma de calcular el PCA es con la llamada descomposición en valores singulares (SVD).



Relación entre valores propios y valores singulares

Los valores propios del PCA son las varianzas de cada eje principal.

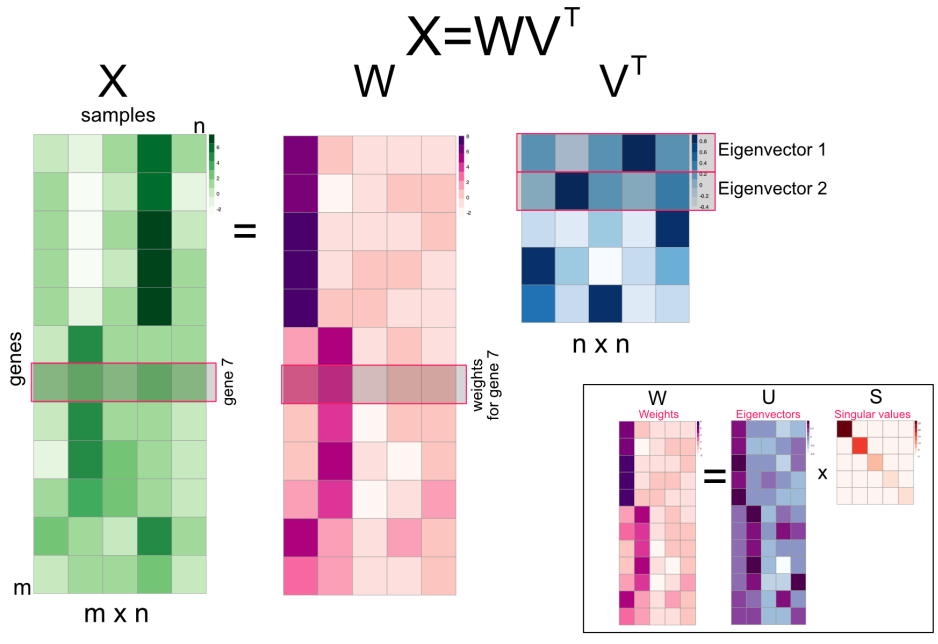
Se corresponden con el cuadrado de los valores singulares ya que para una matriz centrada y estandarizada:

$$\text{cov}(X^T) = XX^T = US^2U^T$$
$$X^TX = VS^2V^T$$

Luego hay dos tipos de vectores propios: los eigenvectors (eigengenes) y los eigenarrays (eigenassays). Los primeros para la representación de los genes y los segundos para la representación de las muestras.

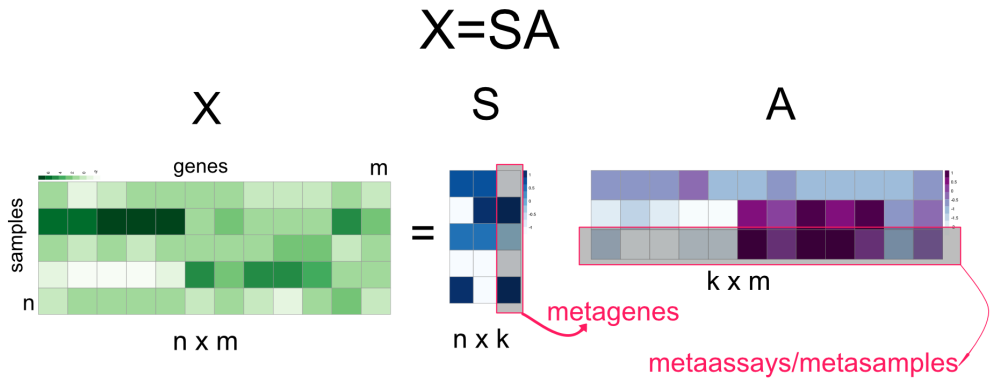
La proyección de las muestras es U^TX, es decir, SV^T.

Vectores propios como variables latentes



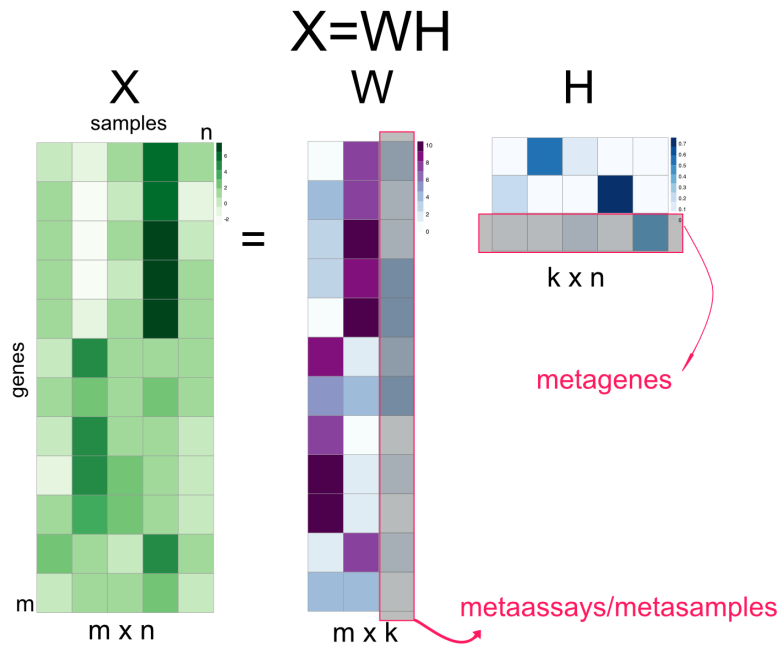
Análisis de componentes independientes (ICA)

El ICA es una extensión del PCA.



El ICA exige la indepencencia estadística entre las columnas de S. Hay distintos algoritmos.

Factorización con matrices no negativas



Las matrices W y H deben contener valores no negativos.

La solución pasa por minimizar una función de coste entre X y WH basada en una distancia.

Es mejor para la interpretación la presencia de muchos 0.

Escalado multidimensional (MDS)

Se trata de representar las distancias entre los objetos (o pacientes) obtenidas en un espacio de alta dimensión con un conjunto de puntos en un espacio de baja dimensión sin perder mucha información.

La solución clásica se conoce también con el nombre de Análisis de coordenadas principales (PCoA).

Extensiones recientes toman esta solución como punto de partida para optimizar otras funciones de coste. El resultado son los métodos MDS no métricos.

Incrustación estocástica de vecinos (t-SNE)

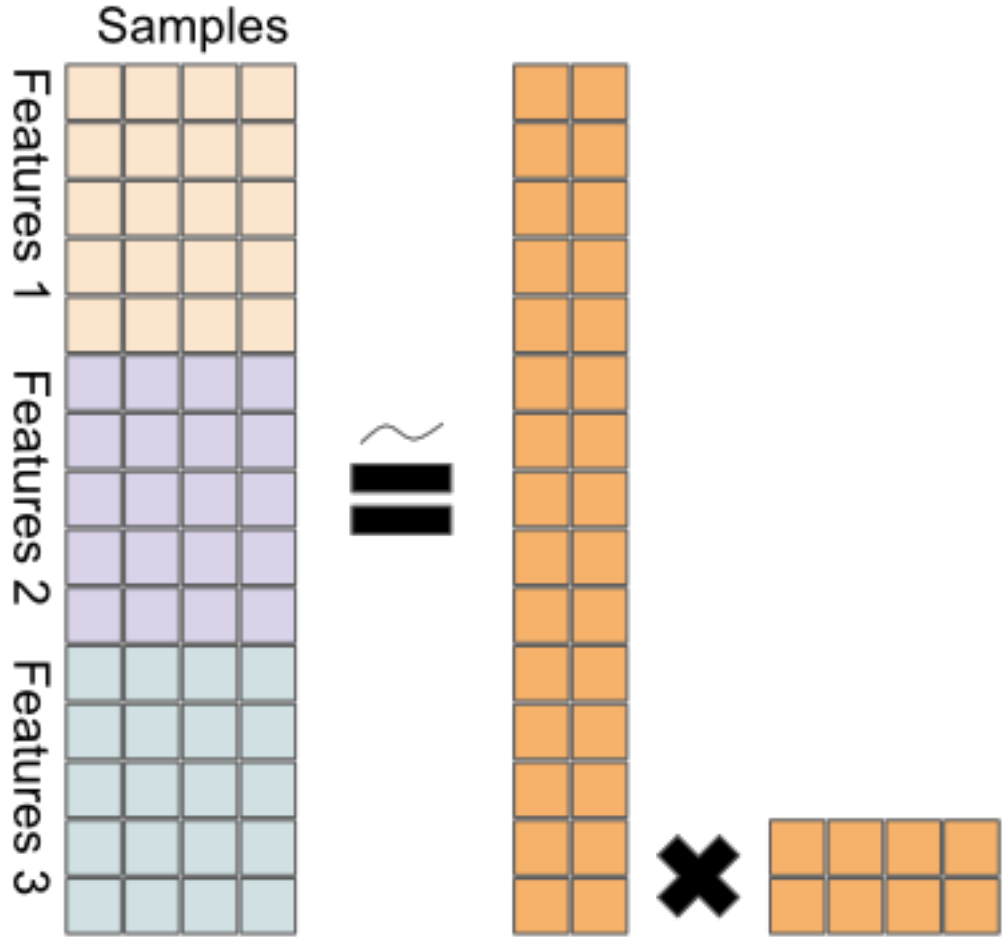
Es una técnica similar al MDS. Se trata de representar las distancias en el espacio de alta dimensión con puntos cuyas distancias reproducen las anteriores.

La diferencia con el MDS es que con el t-SNE se pretende preservar las estructuras locales y despreciar los elementos lejanos. Para ello se utiliza un parámetro de "perplejidad".

El método construye unas probabilidades (como similitudes) y trata de conseguir una estructura en dimensión reducida cuyas probabilidades se ajustan mediante la minimización de la divergencia de Kullback-Leibler por el descenso del gradiente.

Análisis de factores múltiple (MFA)

Es una extensión del PCA a varios conjuntos de datos.



Análisis de factores múltiple (MFA)

Como las medidas de los diferentes experimentos tienen diferentes escalas y, por tanto, diferentes varianzas, debemos normalizar cada tipo de datos:

$$X_n = \begin{bmatrix} X_1/\lambda_1^{(1)} \\ X_2/\lambda_1^{(2)} \\ \vdots \\ X_L/\lambda_1^{(L)} \end{bmatrix} = WH$$

donde $\lambda_1^{(i)}$ es el primer valor propio del PCA de X_i.

En R tenemos la función `MFA()` del paquete `FactoMineR`.

Factorización con matrices no negativas

La factorización adopta la forma familiar $X \approx WH$, con $X \geq 0$, $W \geq 0$, y $H \geq 0$.

Las restricciones no negativas hacen que una descomposición sin pérdidas (es decir, $X = WH$) sea generalmente imposible. Por lo tanto, el NMF intenta encontrar una solución que minimice la norma de Frobenius (norma euclídea matricial) de la reconstrucción:

$$\min \|X - WH\|_F \qquad W \geq 0, H \geq 0$$

Factorización conjunta con matrices no negativas

Como en el MFA, en el contexto de datos multi-ómicos, la idea es hallar una descomposición de la matriz conjunta.

También aquí hay que normalizar cada matiz de datos por separado

$$X = \begin{bmatrix} X_1^N/\alpha_1 \\ X_2^N/\alpha_2 \\ \vdots \\ X_L^N/\alpha_L \end{bmatrix}$$

donde X_i^N es la matriz de datos normalizada $X_i^N = \frac{x^{ij}}{\sum_j x^{ij}}$ y $\alpha_i = \|X_i^N\|_F$.

El NMF se aplica a la matriz conjunta X.