MÉTODOS DE REDUCCIÓN DE LA DIMENSIONALIDAD
ACP vs t-SNE

María Quílez Miguel

Trabajo de fin de grado de Matemáticas
Universidad de Zaragoza

Director del trabajo: José Tomás Alcalá Nalvaiz
Julio de 2022

Resumen

La visualización de datos y la reducción de la dimensionalidad son dos herramientas en pleno auge y desarrollo en la era del Big Data. Gracias a ellas se obtienen resultados más sencillos y manejables de los datos que facilitan su análisis y su interpretación. Debido a su gran importancia y utilidad han surgido varias técnicas dedicadas a las mencionadas herramientas. En esta memoria se analizarán la siguientes técnicas:

- El Análisis de Componentes Principales (ACP) se trata de una técnica lineal de análisis no supervisado que simplifica la complejidad de los datos originales al tiempo que conserva su variabilidad. Construye nuevas variables obtenidas mediante una proyección de los datos originales en dimensiones más bajas que sintetizan la información inicial [1].

- Una generalización de la anterior es el Escalado Multidimensional Métrico en el que los datos no corresponden a variables sino a similitudes o semejanzas entre elementos. Se trata de una técnica de análisis multivariante que, a partir de una matriz de distancias o de similitudes, obtiene una representación de los individuos en un sistema de referencia de forma que las distancias en ese sistema se aproximen a las iniciales [2].

- t-Distributed Stochastic Neighbor Embedding (t-SNE) es una técnica no lineal de reducción de la dimensión que proporciona una visualización dando a cada punto del espacio de alta dimensión una ubicación en el de baja dimensión. Esta técnica permite capturar gran parte de la estructura local de los datos en el espacio de alta dimensión mientras mantiene la estructura global de los mismos [3].

El objetivo del trabajo se centra en desarrollar detalladamente cada una de las técnicas mencionadas. En especial, se describirá la técnica t-SNE así como su antecesora SNE. Se realizarán una serie de experimentos en los que se podrán apreciar las diferencias entre las visualizaciones que proporcionan el ACP y t-SNE en dos conjuntos de datos. Se seleccionarán los parámetros que influyen en t-SNE en relación con la capacidad de agrupar los datos en grupos homogéneos ya existentes mediante algoritmos de análisis clúster.

Abstract

Visualization of high-dimensional data and dimensionality reduction are two tools on the rise and development in the era of Big Data. Thanks to them, simpler and more manageable results are obtained from the data that facilitate their analysis and interpretation. Due to its great importance and usefulness, several techniques dedicated to the aforementioned tasks have emerged. In this memory the following techniques will be analyzed:

- Principal Component Analysis (PCA) is a linear, unsupervised analysis technique that simplifies the complexity of the original data while preserving its variability. Construct new variables obtained by projecting the original data into a lower dimension linear subspace that synthesize the initial information [1].
- A generalization of the previous one is the Metric Multidimensional Scaling in which the data does not correspond to variables but to distances or similarities between elements. It is a multivariate analysis technique that, from a matrix of distances or similarities, obtains a representation of the individuals in a reference system so that the distances in that system are close to the initial ones [2].
- t-Distributed Stochastic Neighbor Embedding (t-SNE) is a nonlinear dimension reduction technique that provides a visualization methode by giving each point in high-dimensional space a location in low-dimensional space. This technique allows capturing much of the local structure of the data in high-dimensional space while maintaining the global structure of the data [3].

The principal components technique and metric multidimensional scaling are presented in the first chapter. In the second chapter, the SNE technique is developed, exposing its limitations corrected by its successor t-SNE. In the last chapter, a series of experiments will be carried out in which the differences between the visualization results provided by the PCA and t-SNE in two data sets can be appreciated. The parameters that influence t-SNE will be selected in such way that clustering algorithm recognize preexisting groups in data.

Índice general

Capítulo 1

Componentes principales

1.1. Planteamiento del problema

Dadas n observaciones de p variables distintas (linealmente independientes) el objetivo del análisis de componentes principales es analizar si es posible representar esta información con menos variables tomando combinaciones lineales de las primeras [4].

Se dispone de una matriz $\mathbf{X}^*$ de dimensión $n \times p$ donde las variables están en las columnas

$$\mathbf{X}^* = \left[\begin{array}{c|ccc|c} \mathbf{x}_1^* & | & \dots & | & \mathbf{x}_p^* \end{array} \right] \text{ con } \mathbf{x}_i^* \in \mathbb{R}^n \tag{1.1}$$

y las observaciones en las filas

$$\mathbf{X}^* = \begin{bmatrix} (\mathbf{x}_1^*)^{\mathrm{T}} \\ \vdots \\ (\mathbf{x}_n^*)^{\mathrm{T}} \end{bmatrix} \text{ con } \mathbf{x}_j^* \in \mathbb{R}^p. \tag{1.2}$$

Definición 1.1. Sea $\mathbf{x}^* \in \mathbb{R}^n$ una de las columnas en (1.1), entonces su valor medio se calcula como

$$\bar{x}^* = (\mathbf{x}^*)^{\mathrm{T}} \mathbb{1}/n$$

con $\mathbb{1}$ el vector de $\mathbb{R}^n$ con valor 1 en todas sus componentes.

A partir del vector de datos centrados $\mathbf{x} = (\mathbf{x}^* - \bar{x}^* \mathbb{1}) \in \mathbb{R}^n$ es inmediata la siguiente definición.

Definición 1.2. Dado un vector de observaciones $\mathbf{x} \in \mathbb{R}^n$ centrado se define su varianza (muestral) como

$$\mathrm{Var}(\mathbf{x}) = \frac{1}{n} \mathbf{x}^{\mathrm{T}} \mathbf{x}.$$

El centrado de un vector de datos se extiende a matrices de forma inmediata.

Definición 1.3. Sea $\mathbf{X}^*$ una matriz $n \times p$ se define la matriz de datos centrada como

$$\mathbf{X} = (\mathbf{I} - \frac{1}{n} \mathbb{1}\mathbb{1}^{\mathrm{T}}) \mathbf{X}^*$$

donde $\mathbf{I}$ es la matriz identidad de orden n.

Observación 1.4. Las variables (columnas) de la matriz $\mathbf{X}$ tienen media cero.

Definición 1.5. Se define la matriz de varianzas y coviaranzas de la matriz de datos centrada $\mathbf{X}$ como

$$\mathbf{S} = \frac{1}{n} \mathbf{X}^{\mathrm{T}} \mathbf{X}.$$

Observación 1.6. La matriz $\mathbf{S}$ tiene valores propios reales y no negativos ya que es simétrica y semidefinida positiva por ser sus variables linealmente independientes.

El objetivo es encontrar un subespacio de dimensión menor que p tal que al proyecctar sobre él los puntos, éstos mantengan su estructura con la menor distorsión posible. Se puede convertir esta idea intuitiva en un criterio matemático tratando de encontrar un subespacio de dimensión menor que p, por ejemplo de dimensión 1 (una recta), de forma que al proyectar los puntos sobre éste las distancias entre las proyecciones de los puntos y las distancias entre los puntos iniciales tengan la menor diferencia posible.

Definición 1.7. Dado un conjunto de observaciones $\{\mathbf{x}_1,\dots,\mathbf{x}_n\}$ donde $\mathbf{x}_i \in \mathbb{R}^p \ \forall i = 1,\dots,n$ se define el cuadrado de la distancia entre dos puntos del espacio de dimensión grande como $d_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j)$.

Definición 1.8. Dado un conjunto de puntos $\{\mathbf{y}_1,\dots,\mathbf{y}_n\}$ donde $\mathbf{y}_i \in \mathbb{R}^r \ \forall i = 1,\dots,n$ con $1 \leq r < p$ se define el cuadrado de la distancia entre dos puntos proyectados como $\hat{d}_{ij}^2 = (\mathbf{y}_i - \mathbf{y}_j)^{\mathrm{T}}(\mathbf{y}_i - \mathbf{y}_j)$.

Definición 1.9. A partir de las distancias definidas en 1.7 y 1.8 se define la diferencia de las distancias al cuadrado entre el conjunto de observaciones $\{\mathbf{x}_1,\dots,\mathbf{x}_n\}$ con $\mathbf{x}_i \in \mathbb{R}^p$ y los puntos $\{\mathbf{y}_1,\dots,\mathbf{y}_n\}$ con $\mathbf{y}_i \in \mathbb{R}^r \ \forall i = 1,\dots,n$ mediante la siguiente expresión

$$D = \sum_{i=1}^{n} \sum_{j=i+1}^{n} (d_{ij}^2 - \hat{d}_{ij}^2). \tag{1.3}$$

Notar que D se puede descomponer como $D = D_b - D_s$ donde D_b es la suma de los cuadrados de las distancias entre los puntos originales y D_s es la suma de los cuadrados de las distancias entre los puntos del espacio destino.

La técnica de componentes principales busca que la diferencia de las distancias al cuadrado (1.3) sea mínima.

Proposición 1.10. *Minimizar* (1.3) *es equivalente a maximizar la suma de las distancias entre los puntos proyectados* D_s.

Demostración. Se puede escribir la expresión (1.3) de la siguiente forma

$$D = D_b - D_s = \sum_{i=1}^{n} \sum_{j=i+1}^{n} d_{ij}^2 - \sum_{i=1}^{n} \sum_{j=i+1}^{n} \hat{d}_{ij}^2.$$

Como la suma de las distancias originales D_b es constante, para minimizar (1.3) basta maximizar D_s. □

Comenzamos maximizando la distancia entre los puntos proyectados $\mathbf{D}_s$ para el caso $r = 1$.

Teorema 1.11. *En el contexto anterior, maximizar las distancias entre los puntos proyectados es equivalente a maximizar la varianza de la variable definida por las proyecciones de los puntos.*

Demostración. Sea $\mathbf{y}_i = \mathbf{a}_1^{\mathrm{T}}\mathbf{x}_i \ \forall i = 1,\dots,n$ la proyección de una observación $\mathbf{x}_i$ sobre la dirección $\mathbf{a}_1$ con $\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 = 1$. Como la matriz de datos $\mathbf{X}$ tiene media cero entonces las proyecciones $\mathbf{Y} = (\mathbf{y}_1,\dots,\mathbf{y}_n)$ también son centradas ya que

$$\sum_{i=1}^{n} \mathbf{y}_i = \sum_{i=1}^{n} \mathbf{a}_1^{\mathrm{T}}\mathbf{x}_i = \mathbf{a}_1^{\mathrm{T}} \sum_{i=1}^{n} \mathbf{x}_i = 0.$$

Por un lado, se tiene que la suma de las distancias al cuadrado entre los puntos proyectados es

$$\mathbf{D}_s = \sum_{i=1}^{n} \sum_{h=i+1}^{n} (\mathbf{y}_i - \mathbf{y}_h)^2, \tag{1.4}$$

donde cada término $\mathbf{y}_i$ aparece $n-1$ veces ya que cada punto $\mathbf{x}_i$ se compara con los otros $n-1$ y habrá tantos dobles productos como parejas de puntos, es decir, $\binom{n}{2} = n(n-1)/2$. Así, se puede escribir (1.4) como

$$\mathbf{D}_s = (n-1)\sum_{i=1}^{n}\mathbf{y}_i^2 - 2\sum_{i=1}^{n}\sum_{h=i+1}^{n}\mathbf{y}_i\mathbf{y}_h = n\sum_{i=1}^{n}\mathbf{y}_i^2 - B, \tag{1.5}$$

con $B = \sum_{i=1}^{n}\mathbf{y}_i^2 - 2\sum_{i=1}^{n}\sum_{h=i+1}^{n}\mathbf{y}_i\mathbf{y}_h = \sum_{i=1}^{n}\mathbf{y}_i\sum_{i=1}^{n}\mathbf{y}_i = 0$.

Por tanto, maximizar (1.5) es equivalente a maximizar $n\sum_{i=1}^{n}\mathbf{y}_i^2$ que también es equivalente a maximizar la varianza de la variable $\mathbf{Y}$ (ver definición 1.2). □

Observación 1.12. Se puede generalizar la expresión de la suma de distancias al cuadrado entre los puntos proyectados D_s para un subespacio de dimensión $1 \leq r < p$ mediante la siguiente expresión

$$D_s = n\sum_{i=1}^{n}\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_i, \tag{1.6}$$

con $\mathbf{y}_i \in \mathbb{R}^r \ \forall i = 1,\ldots,n$ los vectores de puntos proyectados.

Teorema 1.13. *Maximizar la distancia de los puntos proyectados D_s es equivalente a encontrar una matriz $\boldsymbol{A} = \left[\ \boldsymbol{a}_1^T \mid \ \ldots \mid \ \boldsymbol{a}_r^T\ \right]^T$ ortonormal de dimensión $r \times n$ que maximice*

$$\sum_{i=1}^{r}\boldsymbol{a}_i^T\boldsymbol{S}\boldsymbol{a}_i. \tag{1.7}$$

Se considera dicha matriz como la matriz de proyección ya que proporciona el valor de las variables proyectadas en el subespacio de dimensión r según $\boldsymbol{Y} = \boldsymbol{XA}$.

Demostración. Se quiere aproximar la matriz $\mathbf{X}$ de rango p por otra matriz $\mathbf{Y}$ de rango $1 \leq r < p$. Para ello, a partir de la proyección $\mathbf{Y} = \mathbf{XA}$, se busca la matriz $\mathbf{A}$ de dimensión $r \times n$ y que cumple $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}$. La matriz de varianzas y covarianzas de las variables proyectadas viene dada por

$$\mathbf{S}_{\mathbf{Y}} = \frac{1}{n}\mathbf{Y}^{\mathrm{T}}\mathbf{Y} = \frac{1}{n}\mathbf{A}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{A} = \mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{A}. \tag{1.8}$$

Por otro lado, el valor de D_s según la expresión (1.6) puede reescribirse en términos $\mathbf{S}_{\mathbf{Y}}$ (1.8) como

$$D_s = n\sum_{i=1}^{n}\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_i = n\ \mathrm{traza}(\mathbf{Y}\mathbf{Y}^{\mathrm{T}}) = n\ \mathrm{traza}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y}) = n^2\ \mathrm{traza}(\mathbf{S}_{\mathbf{Y}}) = n^2\ \mathrm{traza}(\mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{A}), \tag{1.9}$$

donde $\mathbf{S}$ es la matriz de varianzas y covarianzas de las variables iniciales.

De acuerdo con la expresión (1.9), maximizar D_s es encontrar r vectores $\mathbf{A} = \left[\ \mathbf{a}_1^{\mathrm{T}} \mid \ \ldots \mid \ \mathbf{a}_r^{\mathrm{T}}\ \right]^{\mathrm{T}}$ que maximicen la suma de los elementos diagonales de $\mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{A}$, es decir, que maximicen $\sum_{i=1}^{r}\mathbf{a}_i^{\mathrm{T}}\mathbf{S}\mathbf{a}_i$.

□

Corolario 1.14. *Dada una matriz de proyección $\boldsymbol{A}$ y su matriz proyectada $\boldsymbol{Y} = \boldsymbol{XA}$, maximizar* (1.7) *es equivalente a maximizar la varianza total, la traza de su matriz de varianzas y covarianzas, de la matriz de proyecciones $\boldsymbol{Y}$.*

Demostración. Notar que (1.7) se puede desarollar como

$$\sum_{i=1}^{r}\mathbf{a}_i^{\mathrm{T}}\mathbf{S}\mathbf{a}_i = \ \mathrm{traza}(\mathbf{A}\mathbf{S}\mathbf{A}^{\mathrm{T}}) = \ \mathrm{traza}(\mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{A}) = \ \mathrm{traza}(\mathbf{S}_{\mathbf{Y}}).$$

□

1.2. Cálculo de la primera componente

Definición 1.15. La combinación lineal de las variables iniciales con varianza máxima forman la primera componente principal.

Teorema 1.16. *Sea **S** la matriz de varianzas y covarianzas de las observaciones. Los coeficientes de la primera componente principal vienen dados por el vector propio $\boldsymbol{a}_1$ asociado al valor propio λ, siendo éste el mayor de los valores propios de **S**.*

Demostración. Se representan los valores de la proyección de la primera componente principal por el vector $\mathbf{y}_1 = \mathbf{X}\mathbf{a}_1$. La media de $\mathbf{y}_1$ será cero ya que la de las variables originales lo es y su varianza será

$$\text{Var}(\mathbf{y}_1) = \frac{1}{n}\mathbf{y}_1^{\mathrm{T}}\mathbf{y}_1 = \frac{1}{n}\mathbf{a}_1^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{a}_1 = \mathbf{a}_1^{\mathrm{T}}\mathbf{S}\mathbf{a}_1. \tag{1.10}$$

Como el objetivo es maximizar (1.10) se introduce la restricción en el módulo del vector mediante el multiplicador de Lagrange con un vector de coeficientes normalizado obteniendo la siguiente función

$$M = \mathbf{a}_1^{\mathrm{T}}\mathbf{S}\mathbf{a}_1 - \lambda(\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 - 1). \tag{1.11}$$

Derivando en (1.11) con respecto $\mathbf{a}_1$ e igualando a cero se obtiene

$$\frac{\partial M}{\partial \mathbf{a}_1} = 2\mathbf{S}\mathbf{a}_1 - 2\lambda\mathbf{a}_1 = 0, \tag{1.12}$$

es decir, $\mathbf{S}\mathbf{a}_1 = \lambda\mathbf{a}_1$. Por tanto, $\mathbf{a}_1$ es un vector propio de $\mathbf{S}$ con valor propio λ, resultando que la varianza (1.10) de $\mathbf{y}_1$ es

$$\mathbf{a}_1^{\mathrm{T}}\mathbf{S}\mathbf{a}_1 = \mathbf{a}_1^{\mathrm{T}}\lambda\mathbf{a}_1 = \lambda. \tag{1.13}$$

Como se quiere maximizarla, se toma λ como el mayor valor propio de $\mathbf{S}$. Su vector propio asociado, $\mathbf{a}_1$, define los coeficientes de la primera componente principal. □

La primera componente principal es la mejor recta de proyección (mejor subespacio de dimensión 1) de la matriz de datos $\mathbf{X}$. La distancia D_s entre los puntos proyectados en una recta se maximiza al tomar $\mathbf{a}_1$, la dirección de proyección, como el vector propio asociado al mayor valor propio de $\mathbf{S}$.

1.3. Cálculo de la segunda componente

Definición 1.17. La combinación lineal de las variables originales, normalizada y ortogonal a $\mathbf{a}_1$, de varianza máxima forma la segunda componente principal.

Proposición 1.18. *Sea **S** la matriz de varianzas y covarianzas de las observaciones. Los coeficientes de la segunda componente principal vienen dados por el vector propio $\boldsymbol{a}_2$ asociado al valor propio λ_2, siendo éste el segundo mayor valor propio de **S**.*

Demostración. Procediendo como en la demostración del teorema 1.16 se representan los valores de la proyección de la segunda componente principal mediante el vector $\mathbf{y}_2 = \mathbf{X}\mathbf{a}_2$. La media de $\mathbf{y}_2$ será cero ya que la de las variables originales lo es y su varianza será

$$\text{Var}(\mathbf{y}_2) = \frac{1}{n}\mathbf{y}_2^{\mathrm{T}}\mathbf{y}_2 = \frac{1}{n}\mathbf{a}_2^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{a}_2 = \mathbf{a}_2^{\mathrm{T}}\mathbf{S}\mathbf{a}_2. \tag{1.14}$$

Como el objetivo es maximizar (1.14), introduciendo la restricción en el módulo del vector $\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 = 1$ mediante el multiplicador de Lagrange, se considera la siguiente función

$$M = \mathbf{a}_2^{\mathrm{T}}\mathbf{S}\mathbf{a}_2 - \lambda(\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 - 1). \tag{1.15}$$

Derivando en (1.15) e igualando a cero se obtiene $\mathbf{S}\mathbf{a}_2=\lambda\mathbf{a}_2$, es decir, que $\mathbf{a}_2$ es un vector propio de $\mathbf{S}$ con valor propio λ. Por tanto, la varianza (1.14) de $\mathbf{y}_2$ es

$$\mathbf{a}_2^{\mathrm{T}}\mathbf{S}\mathbf{a}_2 = \mathbf{a}_2^{\mathrm{T}}\lambda\mathbf{a}_2 = \lambda. \tag{1.16}$$

Como se quiere maximizarla, se toma ahora λ como el segundo mayor valor propio de $\mathbf{S}$. El vector propio correspondiente $\mathbf{a}_2$ y ortogonal a $\mathbf{a}_1$, define los coeficientes de cada variable en la segunda componente principal.

□

Teorema 1.19. *Sean $\boldsymbol{A} = \left[\ \boldsymbol{a}_1^T \mid \ \boldsymbol{a}_2^T\ \right]^T$ los vectores propios ortonormales asociados a los dos mayores valores propios de la matriz de varianzas y covarianzas $\boldsymbol{S}$. La matriz $\boldsymbol{Y} = (\boldsymbol{X}\boldsymbol{a}_1, \boldsymbol{X}\boldsymbol{a}_2)$ es el conjunto de puntos de $\mathbb{R}^2$ que maximiza D_s.*

Demostración. Según el teorema 1.13 buscamos los vectores $\mathbf{a}_1$ y $\mathbf{a}_2$ que maximizan

$$\sum_{j=1}^{2} \mathbf{a}_j^{\mathrm{T}}\mathbf{S}\mathbf{a}_j = \mathbf{a}_1^{\mathrm{T}}\mathbf{S}\mathbf{a}_1 + \mathbf{a}_2^{\mathrm{T}}\mathbf{S}\mathbf{a}_2. \tag{1.17}$$

Sin pérdida de generalidad podemos imporner las rectricciones de normalidad $\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 = 1$ y $\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 = 1$ obteniendo la siguiente función objetivo

$$M = \mathbf{a}_1^{\mathrm{T}}\mathbf{S}\mathbf{a}_1 + \mathbf{a}_2^{\mathrm{T}}\mathbf{S}\mathbf{a}_2 - \lambda_1(\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 - 1) - \lambda_2(\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 - 1). \tag{1.18}$$

Derivando (1.18) recpecto de los vectores $\mathbf{a}_1$ y $\mathbf{a}_2$ se obtiene

$$\frac{\partial M}{\partial \mathbf{a}_1} = 2\mathbf{S}\mathbf{a}_1 - 2\lambda_1\mathbf{a}_1, \qquad \frac{\partial M}{\partial \mathbf{a}_2} = 2\mathbf{S}\mathbf{a}_2 - 2\lambda_2\mathbf{a}_2 \tag{1.19}$$

e igualando a cero la solución que optimiza la función objetivo (1.18) es

$$\mathbf{S}\mathbf{a}_1 = \lambda_1\mathbf{a}_1, \qquad \mathbf{S}\mathbf{a}_2 = \lambda_2\mathbf{a}_2. \tag{1.20}$$

Según (1.20) los vectores $\mathbf{a}_1$ y $\mathbf{a}_2$ deben ser vectores propios de $\mathbf{S}$. Se toman los vectores propios de norma uno y se sustituyen en (1.18) de forma que, en el máximo, la función objetvivo toma el siguiente valor

$$M(\mathbf{a}_1, \mathbf{a}_2) = \lambda_1 + \lambda_2, \tag{1.21}$$

por lo que λ_1 y λ_2 tienen que ser los dos mayores valores propios de la matriz $\mathbf{S}$ con $\mathbf{a}_1$ y $\mathbf{a}_2$ sus vectores propios asociados. □

La pareja de las dos primeras componente principales es el mejor plano de proyección de la matriz de datos $\mathbf{X}$. La distancia D_s entre los puntos proyectados en un plano se maximiza definiendo los coeficientes de cada proyección como $\mathbf{Y} = (\mathbf{X}\mathbf{a}_1, \mathbf{X}\mathbf{a}_2)$ donde $\mathbf{a}_1$ y $\mathbf{a}_1$ son los vectores propios asociados a los mayores valores propios de $\mathbf{S}$.

1.4. Cálculo de la r-ésima componente principal

Definición 1.20. La combinación lineal normalizada y ortogonal a $\mathbf{a}_1, \ldots, \mathbf{a}_{r-1}$ de varianza máxima forma la $r-$ésima componente principal.

Proposición 1.21. *Sean $\boldsymbol{A} = \left[\ \boldsymbol{a}_1^T \mid \ \ldots \mid \ \boldsymbol{a}_r^T\ \right]^T$ los vectores propios ortonormales asociados a los r mayores valores propios de la matriz de varianzas y covarianzas $\boldsymbol{S}$. La matriz $\boldsymbol{Y} = (\boldsymbol{X}\boldsymbol{a}_1, \ldots, \boldsymbol{X}\boldsymbol{a}_r)$ es el conjunto de puntos de $\mathbb{R}^r$ que maximiza D_s.*

Demostración. Es una generalización del teorema 1.19 para un subespacio de dimensión r.

□

El espacio de dimensión r que mejor representa a los puntos originales viene definido por los vectores propios asociados a los r mayores valores propios de $\mathbf{S}$. Se puede abordar el problema tratando de encontrar una matriz $\mathbf{A}$ de rango $r < p$ que ofrezca la mejor aproximación de la matriz de covarianzas $\mathbf{S}$.

Teorema 1.22. *Dada una matriz $\boldsymbol{S}$ definida positiva de dimensión $p \times p$ y $1 \leq r \leq p$. Se considera la función*

$$\phi(\boldsymbol{B}) = traza((\boldsymbol{S}-\boldsymbol{B})^2),$$

siendo $\boldsymbol{B}$ una matriz semidefinida positiva de rango r. Entonces el valor mínimo de ϕ se alcanza cuando

$$\boldsymbol{B} = \sum_{i=1}^{r} \lambda_i \boldsymbol{a}_i \boldsymbol{a}_i^T,$$

con $\lambda_1, \cdots, \lambda_r$ los r mayores valores propios de $\boldsymbol{S}$ y $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_r$ sus correspondientes vectores propios ortonormales.

Demostración. Se puede escribir $\mathbf{B} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$ donde $\mathbf{A}$ es una matriz $p \times r$ de rango r. Se tiene

$$\phi(\mathbf{A}) = \text{traza}(\mathbf{S} - \mathbf{A}\mathbf{A}^{\mathrm{T}})^2. \tag{1.22}$$

Se quiere encontrar la matriz $\mathbf{A}$ que minimiza la función objetivo (1.22). De acuerdo con la expresión sobre la diferencial de funciones matriciales [5], la primera diferencial es

$$\frac{\partial \phi}{\partial \mathbf{A}} = -2 \text{ traza}(\mathbf{S} - \mathbf{A}\mathbf{A}^{\mathrm{T}}) \frac{\partial \mathbf{A}\mathbf{A}^{\mathrm{T}}}{\partial \mathbf{A}} = -4 \text{ traza}(\mathbf{A}^{\mathrm{T}}(\mathbf{S} - \mathbf{A}\mathbf{A}^{\mathrm{T}})) \frac{\partial \mathbf{A}}{\partial \mathbf{A}}. \tag{1.23}$$

Utilizando que la traza de una matriz coincide con la traza de su traspuesta, se tiene que la ecuación anterior (1.23) es

$$\frac{\partial \phi}{\partial A} = -4 \text{ traza}(\mathbf{A}^{\mathrm{T}}(\mathbf{S} - \mathbf{A}\mathbf{A}^{\mathrm{T}})) = -4 \text{ traza}(\mathbf{S}\mathbf{A} - \mathbf{A}\mathbf{A}^{\mathrm{T}}\mathbf{A})$$

y se anula cuando

$$\mathbf{S}\mathbf{A} = \mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A}). \tag{1.24}$$

Dado que $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ es simétrica se puede diagonzalizar, es decir, dados $\mu_1, \cdots, \mu_r$ los valores propios de $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ existe una matriz ortogonal Λ de dimensión $r \times r$ tal que

$$\Lambda^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}}\mathbf{A})\Lambda = \text{ diag } (\mu_1, \cdots, \mu_r) = \mathbf{M}. \tag{1.25}$$

Se define la matriz $\mathbf{P} = \mathbf{A}\Lambda\mathbf{M}$ de forma que

$$\mathbf{S}\mathbf{A} = \mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A}) \Leftrightarrow \mathbf{S}\mathbf{A}\Lambda = \mathbf{A}\Lambda\Lambda^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}}\mathbf{A})\Lambda \Leftrightarrow \mathbf{S}\mathbf{A}\Lambda = \mathbf{A}\Lambda\mathbf{M} \Leftrightarrow \mathbf{S}\mathbf{A}\Lambda\mathbf{M} = \mathbf{A}\Lambda\mathbf{M}\mathbf{M} \Leftrightarrow \mathbf{S}\mathbf{P} = \mathbf{P}\mathbf{M} \tag{1.26}$$

y así se puede reescribir (1.24) como

$$\mathbf{S}\mathbf{P} = \mathbf{P}\mathbf{M}, \qquad \mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I}_r. \tag{1.27}$$

Por tanto, todo valor propio de $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ es valor propio de $\mathbf{S}$ y $\mathbf{P}$ es su correspondiente matriz de vectores propios.

Desarrollando la expresión de la función objetivo (1.22), se puede escribir de la siguiente forma

$$\phi(\mathbf{B}) = \text{traza}(\mathbf{S}^2) - \text{traza}(\mathbf{M}^2), \tag{1.28}$$

pues $\text{traza}(\mathbf{A}\mathbf{A}^{\mathrm{T}}\mathbf{S}) = \text{traza}((\mathbf{A}\mathbf{A}^{\mathrm{T}})^2)$ por (1.24) y $\text{traza}(\mathbf{S}\mathbf{A}\mathbf{A}^{\mathrm{T}}) = \text{traza}(\mathbf{M}^2)$ por la definición de $\mathbf{M}$ (1.25).

El mínimo se alcanza tomando $\mu_1, \cdots, \mu_r$ igual a $\lambda_1, \cdots, \lambda_r$ los r mayores valores propios de $\mathbf{S}$. De esta forma, se tiene

$$\mathbf{B} = \mathbf{A}\mathbf{A}^{\mathrm{T}} = \mathbf{P}\mathbf{M}^{1/2}\Lambda^{\mathrm{T}}\Lambda\mathbf{M}^{1/2}\mathbf{P}^{\mathrm{T}} = \mathbf{P}\mathbf{M}\mathbf{P}^{\mathrm{T}} = \sum_{i=1}^{r} \lambda_i a_i a_i^{\mathrm{T}}. \tag{1.29}$$

□

1.5. Escalado multidimensional

El escalado multidimensional es una generalización de la idea de componentes principales. Se considera una matriz $\mathbf{D}$ cuadrada $n \times n$ de distancias o disimilaridades entre los n datos de un conjunto. El objetivo de este método es obtener una matriz $\mathbf{X}$ de dimensiones $n \times p$ a partir de la matriz $\mathbf{D}$ que pueda interpretarse como la matriz de p variables en los n datos donde la distancia euclídea entre los datos presente la menor distorsión posible respecto a la matriz $\mathbf{D}$.

1.5.1. Construcción de variables a partir de las distancias

A partir de la matriz $\mathbf{X}$ de nuestros datos centrados de dimensión $n \times p$ se pueden construir dos matrices cuadradas y semidefinidas positivas: la matriz de coviaranzas de la definición 1.5 y la matriz de productos cruzados.

Definición 1.23. Dada la matriz $\mathbf{X}$ de datos centrados de dimensiones $n \times p$ se define la matriz de productos cruzados de dimensión $n \times n$ como $\mathbf{Q}$=$\mathbf{X}\mathbf{X}^{\mathrm{T}}$. Sus términos contienen el producto escalar por pares de elementos q_{ij} =$\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j$, siendo $\mathbf{x}_i^{\mathrm{T}}$ la fila i de la matriz $\mathbf{X}$. Se puede considerar $\mathbf{Q}$ como una matriz de similitud siendo el producto escalar la medida que define la similitud entre elementos.

Proposición 1.24. *Dados dos puntos $\boldsymbol{x}_i$ y $\boldsymbol{x}_j$ se puede expresar su distancia euclídea al cuadrado d_{ij}^2 en función de los elementos de $\boldsymbol{Q}$ mediante la siguiente expresión*

$$d_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij}$$

Demostración. La distancia euclídea al cuadrado de dos puntos $\mathbf{x}_i$ y $\mathbf{x}_j$ es

$$d_{ij}^2 = \sum_{s=1}^{p}(x_{is} - x_{js})^2 = \sum_{s=1}^{p} x_{is}^2 + \sum_{s=1}^{p} x_{js}^2 - 2\sum_{s=1}^{p} x_{is}x_{js} = q_{ii} + q_{jj} - 2q_{ij} \tag{1.30}$$

□

Definición 1.25. A partir de la proposición 1.24 se define la matriz de distancias al cuadrado entre puntos como

$$\mathbf{D} = diag(\mathbf{Q})\mathbb{1}^{\mathrm{T}} + \mathbb{1}diag(\mathbf{Q})^{\mathrm{T}} - 2\mathbf{Q}$$

siendo $diag(\mathbf{Q})$ el vector formado por los términos de la diagonal de $\mathbf{Q}$.

El problema que se plantea en el escalado multidimensional es encontrar la matriz $\mathbf{X}$ de observaciones a partir de una matriz de distancias al cuadrado $\mathbf{D}$, obteniendo primero la matriz de similitud $\mathbf{Q}$.

Definición 1.26. Sea la matriz de distancias al cuadrado $\mathbf{D}$ se define la media de los elementos de la fila i como $d_{i\bullet}^2 = \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2$, la media de los elementos de la columna j como $d_{\bullet j}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2$ y la media de los elementos de $\mathbf{D}$ como $d_{\bullet\bullet}^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2$.

Teorema 1.27. *Dada una matriz de distancias $\boldsymbol{D}$, se puede expresar la matriz de similitud $\boldsymbol{Q}$ definida como $q_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i\bullet}^2 - d_{\bullet j}^2 + d_{\bullet\bullet}^2)$.*

Demostración. En primer lugar, notar que se puede asumir sin pérdida de generalidad que las variables tienen media cero ya que las distancias al cuadrado entre dos puntos d_{ij} no varian si se expresan las variables en desviaciones a la media ya que

$$d_{ij}^2 = \sum_{s=1}^{p}(x_{is} - x_{ij})^2 = \sum_{s=1}^{p}[(x_{is} - \bar{x_s}) - (x_{js} - \bar{x_s})]^2. \tag{1.31}$$

Como se conocen las distancias entre puntos, se busca una matriz $\mathbf{X}$ con variables de media cero, $\mathbf{X}^T\mathbb{1} = 0$. De esto último se deduce $\mathbf{Q}\mathbb{1} = 0$, es decir, la suma de los elementos de las filas de $\mathbf{Q}$ tiene que ser cero, $\sum_{i=1}^{n} q_{ij} = 0$, y se suma por filas y por columnas en (1.24) obteniendo

$$\sum_{i=1}^{n} d_{ij}^2 = \text{traza}(\mathbf{Q}) + nq_{jj}, \qquad \sum_{j=1}^{n} d_{ij}^2 = \text{traza}(\mathbf{Q}) + nq_{ii} \tag{1.32}$$

y sumando de nuevo en la segunda expresión de (1.32) por filas se obtiene

$$\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 = 2n\,\text{traza}(\mathbf{Q}). \tag{1.33}$$

Sustituyendo en (1.24) los valores q_{ii} y q_{jj} obtenidos en (1.32) se obtiene

$$d_{ij}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 - \frac{\text{traza}(\mathbf{Q})}{n} + \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 - \frac{\text{traza}(\mathbf{Q})}{n} - 2q_{ij}. \tag{1.34}$$

Usando (1.33) en (1.34) se tiene

$$d_{ij}^2 = d_{i\bullet}^2 + d_{\bullet j}^2 - d_{\bullet\bullet}^2 - 2q_{ij}. \tag{1.35}$$

Finalmente, se despeja en (1.35) y se obtiene el resultado $q_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i\bullet}^2 - d_{\bullet j}^2 + d_{\bullet\bullet}^2)$. □

Teorema 1.28. *Dada una matriz de similitud semidefinida positiva $\boldsymbol{Q}$, se puede escribir según la siguiente expresión*

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{X}\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{X}^T \tag{1.36}$$

siendo $\boldsymbol{A}$ cualquier matriz ortogonal de dimensión $p \times p$.

Demostración. Se supone que la matriz de similitud $\mathbf{Q}$ es definida positiva de rango p. Entonces se puede descomponer como el siguiente producto de matrices

$$\mathbf{Q} = \mathbf{V}\Lambda\mathbf{V}^T \tag{1.37}$$

con $\mathbf{V}$ de dimensiones $n \times p$ formada por los vectores propios de $\mathbf{Q}$ y Λ matriz diagonal $p \times p$ de valores propios de $\mathbf{Q}$. Por tanto, se puede escribir (1.37) de la siguiente forma

$$\mathbf{Q} = (\mathbf{V}\Lambda^{1/2})(\Lambda^{1/2}\mathbf{V}^T). \tag{1.38}$$

Finalmente, se obtiene el resultado denotando $\mathbf{X}$=$\mathbf{V}\Lambda^{1/2}$. Se ha obtenido una matriz $\mathbf{X}$ de dimensiones $n \times p$ con p variables incorreladas que reproducen la métrica inicial.

Las distancias vienen dadas por los términos de la matriz de similitud $\mathbf{Q}$ según la proposición 1.24, y esta matriz es invariante ante rotaciones de las variables pues las distancias no varían si rotamos los puntos, es decir, no varían si se multiplica por una matriz ortogonal. □

1.5.2. Construcción de las coordenadas principales

La matriz de similitud $\mathbf{Q}$ obtenida a partir de la matriz de distancias $\mathbf{D}$ suele tener r valores propios positivos y considerablemente más grandes que el resto. Si los restantes $p-r$ valores propios no nulos son menores que los primeros, se puede obtener una representación aproximada de los puntos utilizando los mayores r valores propios positivos de la matriz de similitud $\mathbf{Q}$.

Para construir las coordenadas principales se calculan los valores propios de la matriz $\mathbf{Q}$. Se toman los r mayores valores propios positivos de manera que los restantes $p-r$ sean próximos a cero. De este modo se puede aproximar la matriz de similitud

$$\mathbf{Q} \approx (\mathbf{V}_r\Lambda_r^{1/2})(\Lambda_r^{1/2}\mathbf{V}_r^T) \tag{1.39}$$

donde $\mathbf{V}_r$ es la matriz formada por los vectores propios asociados a los r valores propios mayores y $\Lambda_r^{1/2}$ la matriz diagonal con las raíces de los r mayores propios $\sqrt{\lambda_i}$ con $i = 1, \cdots, r$.

Definición 1.29. Dada la matriz de distancias $\mathbf{D}$ se definen las coordenadas principales como la matriz $\mathbf{Y}_r$ =$\mathbf{V}_r\Lambda_r^{1/2}$ siendo $\mathbf{V}_r$ la matriz formada por los vectores propios asociados a los r valores propios mayores y $\Lambda_r^{1/2}$ la matriz diagonal con las raíces de los r mayores propios $\sqrt{\lambda_i}$ con $i = 1, \cdots, r$.

1.6. Relación entre componentes y coordenadas principales

Teorema 1.30. *Las coordenadas principales obtenidas a partir de la matriz de distancias **D** son equivalentes a las componentes principales de las variables de **X**.*

Demostración. Con las variables de $\mathbf{X}$, que se asumen de media cero, las componentes principales son los vectores propios de la matriz $\frac{1}{n}\mathbf{X}^T\mathbf{X}$ mientras que las componentes principales son los vectores propios de $\mathbf{Q}$=$\mathbf{X}\mathbf{X}^T$ multiplicados por la raíz de su valor propio asociado $\sqrt{\lambda_i}$. Demostrar que las componentes y las coordenadas principales son equivalentes es probar que $\mathbf{X}^T\mathbf{X}$ y $\mathbf{X}\mathbf{X}^T$ tienen los mismos valores propios no nulos y el mismo rango.

Sea $\mathbf{a}_i$ vector propio de $\mathbf{X}^T\mathbf{X}$ con valor propio asociado λ_i, entonces $\mathbf{X}^T\mathbf{X}\mathbf{a}_i = \lambda_i\mathbf{a}_i$ y multiplicando en esta expresión por la matriz $\mathbf{X}$ se tiene

$$\mathbf{X}\mathbf{X}^T\mathbf{X}\mathbf{a}_i = \lambda_i\mathbf{X}\mathbf{a}_i \tag{1.40}$$

de donde se deduce que $\mathbf{X}\mathbf{a}_i$ es vector propio de $\mathbf{X}\mathbf{X}^T$ con valor propio asociado λ_i.

Si $n > p$, la matriz $\mathbf{X}^T\mathbf{X}$ tiene rango máximo entonces tendrá p valores propios no nulos que serán los valores propios no nulos de $\mathbf{X}\mathbf{X}^T$.

Se pueden expresar las p componentes principales mediante la siguiente matriz de dimensiones $n \times p$

$$\mathbf{Y} = \mathbf{X}\mathbf{A} \tag{1.41}$$

con $\mathbf{A}$ la matriz formada por los p vectores propios asociados a los p valores propios mayores.

Por otro lado, la matriz de coordenadas principales se escribe

$$\mathbf{Z} = [\mathbf{v}_1, \cdots, \mathbf{v}_p] \begin{bmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_p} \end{bmatrix} = \mathbf{V}\Lambda^{1/2} \tag{1.42}$$

donde $\mathbf{V}$ es la matriz formada por los p vectores propios de $\mathbf{X}\mathbf{X}^T$ de dimensiones $n \times p$ y $\Lambda^{1/2}$ matriz diagonal $p \times p$.

Como las expresiones (1.41) y (1.42) son iguales, salvo un factor de escala $\sqrt{\lambda_i}$, ambos procedimientos conducen al mismo resultado.

□

Dada una matriz de datos centrados $\mathbf{X}$ de dimensión $n \times p$ se puede representar sobre un subespacio de dimensión r mediante la matriz de datos proyectados de dimensión $n \times r$ dada por

$$\mathbf{Y}_r = \mathbf{X}\mathbf{A}_r$$

siendo $\mathbf{A}_r$ la matriz formada por los coeficientes de las r primeras componentes principales de la matriz de datos $\mathbf{X}$.

Esta matriz de datos coincide, salvo por la matriz diagonal cuyos elementos son las raíces cuadradas de los r mayores valores propios de $\mathbf{S}$, con la matriz de coordenadas principales obtenida a partir de la matriz $\mathbf{D}$ de distancias o de la matriz $\mathbf{Q}$ de similitud.

Capítulo 2

t-Distributed Stochastic Neighbor Embedding (t-SNE)

El principal objetivo que comparten las técnicas de reducción de la dimensionalidad lineales es representar alejados en el espacio de dimensión pequeña aquellos puntos que sean más distintos. Sin embargo, el principal fallo radica en el caso en el que los puntos se encuentran dispuestos siguiendo un comportamiento no lineal donde el objetivo más importante es mantener las representaciones de los puntos similiares. Las técnicas de representación no lineales tienen como finalidad mantener la estructura local de los puntos, solucionando así el problema anterior.

Se describe a continuación una técnica no lineal conocida como t-Stochastic Neighbor Embedding ([3], [6] y [7]) que presenta un nuevo y popular método para la representación de datos de alta dimensión capaz de mantener la estructura local y global de los datos y de agruparlos en clústers.

A lo largo de este capítulo se buscará obtener una visualización de los datos en el plano, pero la técnica puede generalizarse para encontrar un espacio de dimensión r cualquiera.

2.1. Stochastic Neighbor Embedding

Para entender la técnica presentada en este capítulo es necesario comprender primero la técnica Stochastic Neighbor Embedding (SNE), ya que se trata de una mejora de la última. SNE es un método probabilístico cuyo objetivo es representar observaciones con muchas variables en el plano manteniendo la estructura de vecindad entre los puntos del espacio de dimensión grande. La idea principal es convertir la distancia euclídea entre puntos del espacio de dimensión grande en probabilidades condicionadas que presentan similitudes.

Definición 2.1. Dado un conjunto de observaciones $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ donde $\mathbf{x}_i \in \mathbb{R}^p \ \forall i = 1, \ldots, n$, en la técnica SNE se define la similitud entre un punto $\mathbf{x}_j$ con otro $\mathbf{x}_i$ del espacio de dimensión grande como la probabilidad

$$d_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2)}, \tag{2.1}$$

donde se considera $d_{i|i} = 0$.

Esta definición de similitud representa la estructura de vecindad que el método persigue mantener. Cuanto más lejos esté un punto $\mathbf{x}_j$ del punto $\mathbf{x}_i$ menor será la probabilidad de que éste sea escogido como vecino y, por tanto, de que pertenezca a su grupo.

En la definición 2.1 aparece la varianza σ_i de la distribución Gaussiana centrada en cada punto $\mathbf{x}_i$. No hay un único valor óptimo del parámetro para todos los puntos ya que la densidad de éstos puede variar. Cualquier valor de σ_i induce una distribución de probabilidad, P_i, sobre todos los demás puntos. Esta distribución tiene una entropía que aumenta del mismo modo que lo hace σ_i. El método busca el valor de σ_i que produce P_i con una perplejidad fijada por el usuario.

Definición 2.2. La perplejidad se define como $Perp(P_i)=2^{H(P_i)}$, donde $H(P_i)$ es la entropía de Shannon de P_i medida en bits $H(P_i) = -\sum_j d_{j|i} log_2 d_{j|i}$. La entropía de Shannon es una función matemática que corresponde a la cantidad de información que contiene una variable aleatoria. A mayor cantidad de información emitida mayor es la entropía, es decir, mayor es la incertidumbre sobre lo que se emite. En ausencia de consideraciones previas, la entropía será máxima si todos los símbolos del mensaje emitido son igual de probables.

La técnica SNE tiene como objetivo representar cada punto $\mathbf{x}_i \in \mathbb{R}^p$ mediante una variable bidimensional $\mathbf{y}_i \in \mathbb{R}^2$ manteniendo en el plano las posiciones de los puntos del espacio de dimensión grande.

Definición 2.3. Dado un conjunto de puntos $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ donde $\mathbf{y}_i \in \mathbb{R}^2 \ \forall i = 1, \ldots, n$, en la técnica SNE se define la similitud de un punto $\mathbf{y}_j$ con otro $\mathbf{y}_i$ del espacio de dimensión pequeña como la probabilidad

$$\hat{d}_{j|i} = \frac{\exp(-\|\mathbf{y}_i - \mathbf{y}_j\|^2)}{\sum_{k \neq i} \exp(-\|\mathbf{y}_i - \mathbf{y}_k\|^2)}, \tag{2.2}$$

donde se considera $\hat{d}_{i|i} = 0$.

Si el mapa de puntos $\mathbf{y}_i$, $\mathbf{y}_j$ modela correctamente la similitud entre los puntos $\mathbf{x}_i$, $\mathbf{x}_j$, entonces la similitud de los puntos del espacio de baja y de alta dimensión serán iguales, es decir, $d_{j|i} = \hat{d}_{j|i}$. El método SNE busca los puntos del plano $\mathbf{y}_i$ que minimizan la distancia entre $d_{j|i}$ y $\hat{d}_{j|i}$.

2.1.1. Similitud entre funciones de distribución: divergencia de Kullback-Leibler (KL)

Para estudiar la la similitud o diferencia entre dos funciones de distribución de probabilidad se hace uso de la *divergencia de Kullback-Leibler (KL)*[1] [8].

Definición 2.4. Dadas dos funciones de distribución de probabilidad P y Q se define la divergencia de Kullback-Leibler (KL) como el promedio ponderado de la diferencia logarítimca entre las probabilidades P y Q, donde el promedio se toma usando las probabilidades P,

$$KL(P||Q) = \sum_i P(i) \ln \left(\frac{P(i)}{Q(i)} \right). \tag{2.3}$$

Esta divergencia surge de la teoría de la información y su objetivo es cuantificar la cantidad de información presente en los datos. Puede ser interpretada como una medida de la información perdida cuando se hace uso de una función de probabilidad, Q, para aproximar otra, P.

La técnica SNE plantea un problema de optimización en el que se quiere encontrar los valores de los puntos $\mathbf{y}_i$ del plano que minimizan la distancia entre $d_{j|i}$ y $\hat{d}_{j|i}$, es decir, $J = \sum_i KL(d_{j|i}||\hat{d}_{j|i})$.

Definición 2.5. Para la técnica SNE se define la función coste del problema de optimización que plantea a partir de la siguiente expresión

$$J = \sum_i \sum_j d_{j|i} \ln \left(\frac{d_{j|i}}{\hat{d}_{j|i}} \right), \tag{2.4}$$

donde $d_{j|i}$ y $\hat{d}_{j|i}$ son las probabilidades (2.1) y (2.2) respectivamente.

2.1.2. Proceso de optimización: gradiente descendente

La búsqueda del valor óptimo que minimiza la función coste (2.4) se lleva a cabo por medio del método del *gradiente descendente*[2] cuyo procedimiento se basa en ver el cambio en el resultado de la función tras la variación de parámetros. Para ello, se necesita la expresión del gradiente de la función objetivo.

[1] *Divergencia de Kullback-Leibler (KL)* ver en Anexo A.

[2] *Método del gradiente descendente* ver en Anexo B.

Teorema 2.6. *La expresión del gradiente de la función coste* (2.4) *en la técnica SNE viene dada por la siguiente expresión*

$$\frac{\partial J}{\partial \mathbf{y}_i} = 2\sum_j (d_{j|i} - \hat{d}_{j|i} + d_{i|j} - \hat{d}_{i|j})(\mathbf{y}_i - \mathbf{y}_j), \tag{2.5}$$

donde $d_{i|j}$ *y* $\hat{d}_{i|j}$ *son las probabilidades* (2.1) *y* (2.2) *respectivamente.*

Demostración. Se definen las dos variables auxiliares siguientes:

$$\mathbf{e}_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\|, \qquad Z = \sum_{k\neq l} \exp(-\|\mathbf{y}_k - \mathbf{y}_l\|^2). \tag{2.6}$$

Derivando en la expresión de la función coste (2.4) respecto de $\mathbf{y}_i$ se obtiene

$$\frac{\partial J}{\partial \mathbf{y}_i} = \sum_j \left(\frac{\partial J}{\partial \mathbf{e}_{ij}} + \frac{\partial J}{\partial \mathbf{e}_{ji}}\right)(\mathbf{y}_i - \mathbf{y}_j). \tag{2.7}$$

A continuación, se calculan las derivadas que aparecen en el sumartorio de (2.7) introduciendo la variable auxiliar Z:

$$\frac{\partial J}{\partial \mathbf{e}_{ij}} = -\sum_{k\neq l} d_{k|l}\frac{\partial(log(\hat{d}_{i|j}))}{\partial \mathbf{e}_{ij}} = -\sum_{k\neq l} d_{kl}\left(\frac{\partial(log(\hat{d}_{k|l}Z))}{\partial \mathbf{e}_{ij}} - \frac{\partial log(Z)}{\partial \mathbf{e}_{ij}}\right) = -\sum_{k\neq l} d_{k|l}\left(\frac{1}{Z\hat{d}_{k|l}}\frac{\partial(\exp(-\mathbf{e}_{kl}^2))}{\partial \mathbf{e}_{ij}} - \frac{1}{Z}\frac{\partial Z}{\partial \mathbf{e}_{ij}}\right). \tag{2.8}$$

De forma análoga, se tiene

$$\frac{\partial J}{\partial \mathbf{e}_{ji}} = -\sum_{k\neq l} d_{k|l}\left(\frac{1}{Z\hat{d}_{k|l}}\frac{\partial(\exp(-\mathbf{e}_{kl}^2))}{\partial \mathbf{e}_{ji}} - \frac{1}{Z}\frac{\partial Z}{\partial \mathbf{e}_{ji}}\right).$$

Notar que $\frac{\partial(\exp(-\mathbf{e}_{kl}^2))}{\partial \mathbf{e}_{ij}}$ y $\frac{\partial(\exp(-\mathbf{e}_{kl}^2))}{\partial \mathbf{e}_{ji}}$ no se anulan si $k = i$ y $l = j$, luego continuando en la expresión anterior se tiene

$$\frac{\partial J}{\partial \mathbf{e}_{ij}} = -\left(-2\frac{d_{i|j}}{Z\hat{d}_{i|j}}\exp(-\mathbf{e}_{ij}^2) - \sum_{k\neq l} -2\frac{d_{k|l}\exp(-\mathbf{e}_{ij}^2)}{Z}\right) = \frac{2d_{i|j}\exp(-\mathbf{e}_{ij}^2)}{Z\hat{d}_{i|j}} - 2\sum_{k\neq l}\frac{d_{k|l}\exp(-\mathbf{e}_{ij}^2)}{Z} = 2d_{i|j} - 2\sum_{k\neq l} d_{k|l}\hat{d}_{i|j}.$$

Del mismo modo, al derivar con respecto a $\mathbf{e}_{ji}$ se obtiene

$$\frac{\partial J}{\partial \mathbf{e}_{ji}} = 2d_{j|i} - 2\sum_{k\neq l} d_{k|l}\hat{d}_{j|i}.$$

Por tanto, como $\sum_{k\neq l} d_{k|l} = 1$ se tiene $\frac{\partial J}{\partial \mathbf{e}_{ij}} = 2(d_{i|j} - \hat{d}_{i|j})$ y $\frac{\partial J}{\partial \mathbf{e}_{j|i}} = 2(d_{j|i} - \hat{d}_{j|i})$. Finalmente, se obtiene el resultado sustituyendo en (2.7) estas dos últimas expresiones de la derivada de la función coste J con respecto a $\mathbf{e}_{ij}$ y $\mathbf{e}_{ji}$.

□

El primer paso en el algoritmo del gradiente descendente se inicializa muestreando puntos del mapa al azar mediante una v.a. Gaussiana con $\sigma_i^2 = 0$ centrada en el origen. Luego, iterativamente, se le añade una suma exponencialmente decreciente de gradiantes anteriores para determinar los cambios en las coordenadas de los puntos del mapa.

Definición 2.7. Durante el proceso de optimización de la función coste (2.4) por medio del gradiente descendente con momento se define la solución en la iteración t como la matriz de dimensión $n \times 2$ dada por la siguiente expresión

$$\Upsilon^{(t+1)} = \Upsilon^{(t)} + \eta \left.\frac{\partial J}{\partial \Upsilon}\right|_{\Upsilon=\Upsilon^{(t)}} + \alpha(t)\left(\Upsilon^{(t)} - \Upsilon^{(t-1)}\right),$$

donde η indica la tasa de aprendizaje y $\alpha(t)$ representa el momento en la iteración t.

La matriz Υ obtenida en la última iteración proporciona el valor de los puntos del plano que minimizan la función coste (2.4). Para detener el algoritmo se presentan dos opciones: fijar un número T de iteraciones al inicio o realizar iteraciones hasta obtener un valor prefijado de la norma de la diferencia de dos iteraciones $\Upsilon^{(t)}$ y $\Upsilon^{(t+1)}$.

2.1.3. Inconvenientes del SNE

El primer inconveniente que presenta la técnica SNE es el coste computacional requerido para el proceso de optimización ya que su gradiente involucra todas las probabilidades de los puntos i y j en las que aparecen exponenciales. Esto provoca que el gradiente adopte comportamientos diferentes de manera rápida haciendo que el método no converja.

El segundo inconveniente que se observa es el llamado *crowding problem*. Para representar dos puntos que en el espacio de mayor dimensión se encuentran medianamente separados, el área disponible en el espacio de dimensión pequeña no será suficientemente grande y se tenderá a representar ambos puntos separados en el plano. Por tanto, los puntos que se encuentren separados en el mapa de dimensión grande se encontrarán demasiado lejos en el plano mientras que los que se encuentren próximos en el espacio grande, en el plano se aplastarán dando lugar a una aglomeración de puntos [3].

2.2. t-Distributed Stochastic Neighbor Embedding

En esta sección vamos a presentar una nueva técnica llamada t-Distributed Stochastic Neighbor Embedding (t-SNE) que mejora los inconvenientes del SNE. La técnica del t-SNE se diferencia de la presentada en la sección anterior en dos puntos: utiliza una versión simétrica de la función coste permitiendo simplificar el gradiente y una distribución t-Student en vez de la Gaussiana para calcular la similitud entre dos puntos del espacio de baja dimensión.

2.2.1. Simetrización del método SNE

Definición 2.8. Dado un conjunto de observaciones $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ donde $\mathbf{x}_i \in \mathbb{R}^p \ \forall i = 1, \ldots, n$, en la técnica SNE simetrizada se define la similitud entre un punto $\mathbf{x}_j$ con otro $\mathbf{x}_i$ del espacio de dimensión grande como la probabilidad

$$d_{ij} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2)}{\sum_{k\neq l}\exp(-\|\mathbf{x}_k - \mathbf{x}_l\|^2/2\sigma^2)}, \tag{2.9}$$

donde se considera $d_{ii} = 0$.

Esta definición puede presentar problemas si el punto $\mathbf{x}_i$ fuese un dato atípico ya que en ese caso los valores de d_{ij} serían muy pequeños para todo j y haría que esos valores no tuvieran peso en la función coste. Para solucionar este problema se definen probabilidades conjuntas d_{ij} como probabilidades condicionadas simetrizadas, es decir, $d_{ij} = \frac{d_{j|i} + d_{i|j}}{2}$. Con esta definición cada uno de los puntos $\mathbf{x}_i$ contribuye en la función objetivo.

Ahora, el valor de σ es el mismo para todas las observaciones $\mathbf{x}_i$ con $i = 1, \ldots, n$ según una perplejidad fijada por el usuario.

Definición 2.9. Dado un conjunto de puntos $\{\mathbf{y}_1,\dots,\mathbf{y}_n\}$ donde $\mathbf{y}_i \in \mathbb{R}^2\ \forall i = 1,\dots,n$, en la técnica SNE simetrizada se define la similitud de un punto $\mathbf{y}_j$ con otro $\mathbf{y}_i$ del espacio de dimensión pequeña como la probabilidad

$$\hat{d}_{ij} = \frac{\exp(-\|\mathbf{y}_i - \mathbf{y}_j\|^2)}{\sum_{k\neq l}\exp(-\|\mathbf{y}_k - \mathbf{y}_l\|^2)}, \tag{2.10}$$

donde se considera $\hat{d}_{ii} = 0$.

Definición 2.10. Para la técnica SNE simetrizada se define la función coste del problema de optimización que plantea a partir de la siguiente expresión

$$J = \sum_i \sum_j d_{ij} \ln\left(\frac{d_{ij}}{\hat{d}_{ij}}\right), \tag{2.11}$$

donde d_{ij} y $\hat{d}_{ij}$ son las probabilidades (2.9) y (2.10) respectivamente.

Teorema 2.11. *La expresión del gradiente de la función coste* (2.11) *en la técnica SNE simetrizada viene dada por la siguiente expresión*

$$\frac{\partial J}{\partial \mathbf{y}_i} = 4\sum_j (d_{ij} - \hat{d}_{ij})(\mathbf{y}_i - \mathbf{y}_j), \tag{2.12}$$

donde d_{ij} y $\hat{d}_{ij}$ son las probabilidades (2.9) *y* (2.10) *respectivamente.*

Demostración. El desarrollo de la derivada de la función coste con respecto a la variable $\mathbf{y}_i$ es el mismo que en el teorema 2.6 salvo que, debido a la simetrización de las probabilidades, podemos escribir (2.7) del siguiente modo

$$\frac{\partial J}{\partial \mathbf{y}_i} = \sum_j \left(\frac{\partial J}{\partial \mathbf{e}_{ij}} + \frac{\partial J}{\partial \mathbf{e}_{ji}}\right)(\mathbf{y}_i - \mathbf{y}_j) = 2\sum_j \frac{\partial J}{\partial \mathbf{e}_{ij}}(\mathbf{y}_i - \mathbf{y}_j). \tag{2.13}$$

De esta forma, sustituyendo $\frac{\partial J}{\partial \mathbf{e}_{ij}} = 2(d_{ij} - \hat{d}_{ij})$ en (2.13) tenemos la expresión del gradiente que se busca. □

2.2.2. Solución del *crowding problem*

Para la técnica t-SNE se propondrá usar la distribución t-Student con un grado de libertad (distribución de Cauchy) en lugar de la Gaussiana ya qué posee una cola más pesada que la última. De esta forma, los puntos involucrados en el *crowding problem* se encontrarán más dispersos porque una distancia moderada en el espacio de dimensión grande será representada por distancia mayor en el plano [3].

2.2.3. Método t-SNE

En este método se unen el método SNE simétrico y la idea de emplear la distribución t-Student con un grado de libertad.

Definición 2.12. Dado un conjunto de puntos $\{\mathbf{y}_1,\dots,\mathbf{y}_n\}$ donde $\mathbf{y}_i \in \mathbb{R}^2\ \forall i = 1,\dots,n$, en la técnica t-SNE se define la similitud de un punto $\mathbf{y}_j$ con otro $\mathbf{y}_i$ del espacio de dimensión pequeña como la probabilidad

$$\hat{d}_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k\neq l}(1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}, \tag{2.14}$$

donde se considera $\hat{d}_{ii} = 0$.

Definición 2.13. Para la técnica t-SNE se define la función coste del problema de optimización que plantea a partir de la siguiente expresión

$$J = \sum_i \sum_j d_{ij} \ln\left(\frac{d_{ij}}{\hat{d}_{ij}}\right), \tag{2.15}$$

donde d_{ij} y $\hat{d}_{ij}$ son las probabilidades (2.8) y (2.14) respectivamente.

Teorema 2.14. *La expresión del gradiente de la función coste* (2.15) *en la técnica t-SNE viene dada por la siguiente expresión*

$$\frac{\partial J}{\partial y_i} = 4\sum_j (d_{ij} - \hat{d}_{ij})(\boldsymbol{y}_i - \boldsymbol{y}_j)(1 + \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2)^{-1}, \tag{2.16}$$

donde d_{ij} y $\hat{d}_{ij}$ son las probabilidades (2.8) *y* (2.14) *respectivamente.*

Demostración. El desarrollo del gradiente es análogo al expuesto en el teorema 2.6 salvo que se toman las siguientes variables auxiliares:

$$\mathbf{e}_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\|, \qquad Z = \sum_{k \neq l} (1 + \|\mathbf{k}_i - \mathbf{y}_l\|^2)^{-1}). \tag{2.17}$$

Debido a la simetría de la técnica t-SNE se puede escribir (2.7) del siguiente modo

$$\frac{\partial J}{\partial \mathbf{y}_i} = 2\sum_j \frac{\partial J}{\partial \mathbf{e}_{ij}} (\mathbf{y}_i - \mathbf{y}_j). \tag{2.18}$$

Notar que $\frac{\partial((1+\mathbf{e}_{kl})^{-1})}{\partial \mathbf{e}_{ij}}$ no se anula si $k = i$ y $l = j$, luego continuando en la expresión (2.8) se tiene

$$\frac{\partial J}{\partial \mathbf{e}_{ij}} = -\left(-2\frac{d_{ij}}{Z\hat{d}_{ij}}(1+\mathbf{e}_{ij})^{-2} + \sum_{k\neq l} -2\frac{d_{kl}(1+\mathbf{e}_{ij})^{-2}}{Z}\right) = 2(d_{ij} - \hat{d}_{ij})(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}. \tag{2.19}$$

Finalmente, se obtiene el resultado sustituyendo (2.19) en (2.18). □

Algoritmo del método t-SNE

Se puede describir el algoritmo que sigue el método t-SNE mediante el siguiente pseudocódigo:

Algorithm 1 Algoritmo del método t-Distributed Stochastic Neighbor Embedding

Input: Conjunto de datos $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$
Parámetros de la función coste: Perplejidad *Perp p*
Parámetros de optimización: número de iteraciones T, tasa de aprendizaje η, momento $\alpha(t)$
Output: Representación en el espacio de baja dimensión $\Upsilon^{\mathrm{T}} = \{\mathbf{y}_1, \cdots, \mathbf{y}_n\}$
Begin Calcular las similitudes $d_{j|i}$ con perplejidad *Perp p*
Fijar $d_{ij} = \frac{d_{i|j} + d_{j|i}}{2}$
Inicializar la solución $\Upsilon^{(0)} = \{\mathbf{y}_1, \cdots, \mathbf{y}_n\}$
for t= 1 to T **do**
  Calcular las similitudes del espacio de baja dimensión $\hat{d}_{ij}$ (2.14)
  Calcular el gradiente $\frac{\partial J}{\partial \Upsilon}$ (2.16)
  Obtener $\Upsilon^{(t)} = \Upsilon^{(t-1)} + \eta \frac{\partial J}{\partial \Upsilon} + \alpha(t)(\Upsilon^{(t-1)} - \Upsilon^{(t-2)})$
end for

Capítulo 3

Experimentos

3.1. Introducción

En este capítulo se pondrá a prueba la técnica t-SNE introducida en el capítulo anterior y se comparará con la técnica lineal de componentes principales. Con el objetivo de visualizar ambas técnicas se realizarán varias representaciones bidimensionales de las mismas.

Con respecto a la técnica de t-SNE, se mostrarán las visualizaciones correspondientes al resultado de aplicarla para diferentes valores de los parámetros perplejidad y tasa de aprendizaje.

Además, para obtener una visión más general resultará interesante analizar el funcionamiento de ambas técnicas de reducción de la dimensionalidad ante muestras de dimensiones diferentes. Esto permitirá ver si las técnicas de reducción de dimensionalidad y sus correspondientes parámetros actúan diferente según el tamaño de la muestra a analizar. Para ello se considerarán dos conjuntos de datos de diferentes dimensiones.

3.2. Conjuntos de datos

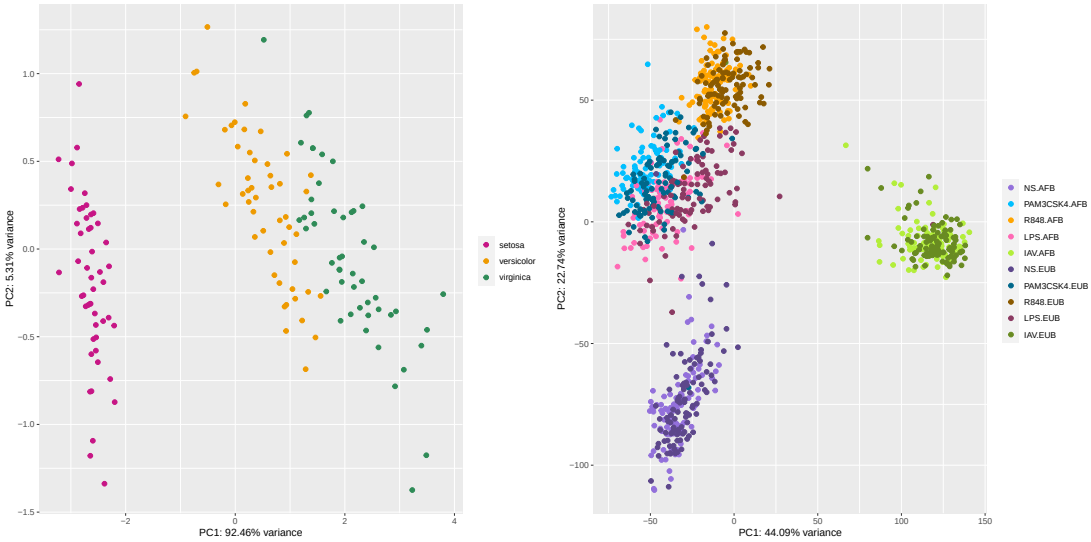
Como conjunto de datos de menor dimesión se va a analizar el dataset Iris [9] formado por 150 observaciones de la planta iris de las cuales 50 son del tipo virginica, 50 del tipo setosa y 50 del tipo versicolor. Las variables que aparecen son el tipo de flor como variable categórica, el ancho y largo del pétalo y el ancho y largo del sépalo en centímetros como variables numéricas.

Por otro lado, se considerarán los datos de un trabajo reciente sobre adaptación genética del sistema inmune de poblaciones humanas [10]. Este conjunto de datos (TMM) corresponde a 967 muestras de miocitos (glóbulos blancos del sistema inmunológico) que son sometidas a 4 estímulos víricos y bacterias o a la ausencia de estímulo. Para cada muestra de miocito se determina por secuenciación de ARN los genes que presentan expresiones genéticas diferenciadas y su correspondiente nivel de expresión. Esto dará un conjunto de 9483 variables para cada muestra de 967 miocitos. Los 4 estímulos a los que se somente la muestra de miocitos se denota por LPS, PAM3CS4, R848 y un virus de la gripe de tipo A, IAV; la ausencia de estímulo o grupo de referencia se denota por NS (*non-stimulated cells*). Cada muestra de células proviene de un sujeto de ascendencia europea (EUB) o africana (AFB) en un número aproximadamente igual. En consecuencia, se puede entender que las 967 muestras celulares están repartidas en 8 grupos preexistentes ascoiadas a las condiciones del estudio. La visualización en un plano de esta muestra permitirá comprobar el grado de separación entre estos grupos y la estructura local de sus relaciones.

3.3. Resultados

3.3.1. Componentes principales

Al aplicar la técnica de componentes principales al conjunto de datos IRIS se puede ver la figura 3.1a un clúster bien diferenciado correspondiente al tipo de iris setosa. Sin embargo, para los otros dos tipos de Iris, aunque sí se ven diferenciados, las componentes principales no consiguen separar estos dos clústers.



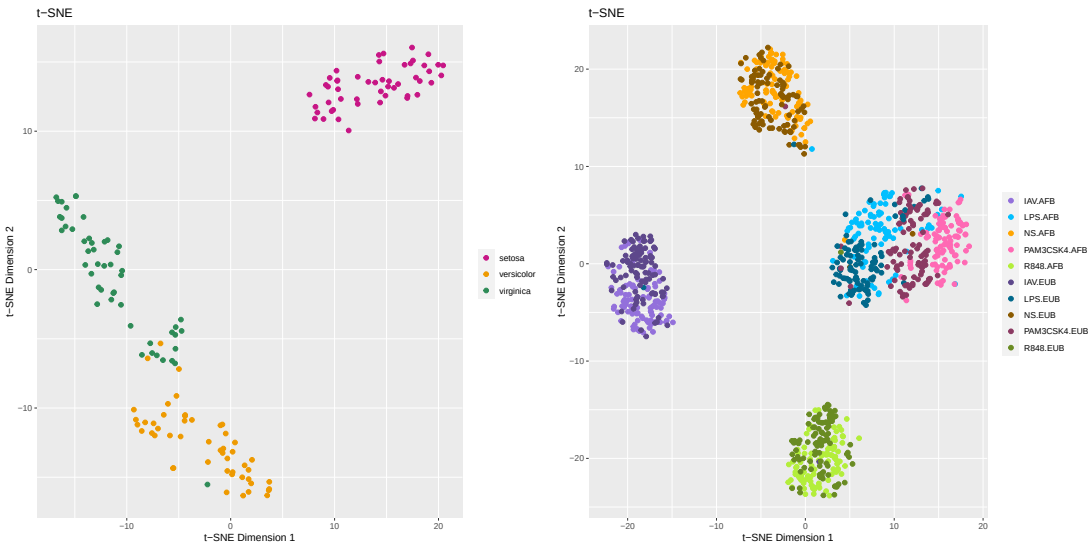
(a) IRIS dataset (b) TMM dataset

Figura 3.1: Análisis por componentes principales.

Como primer intento de analizar el número de clústers diferentes en el dataset TMM presentes según la condición y población se representan las dos primeras componentes principales (figura 3.1b). Se identifican los datos coloreándolos por condición, según la bacteria o estímulo viral, y por población, europeos y africanos. Se aprecian dos clústers bien diferenciados según la condición y un tercer clúster de células no estimuladas. Lo más destacable de este gráfico, es que solo se forman clústers según la condición y no según la población. Se puede interpretar como que la población no es determinante sobre la respuesta del monocito ante la presencia de bacterias y estímulos virales.

3.3.2. t-SNE

Se presentan ahora los resultados de aplicar la técnica t-SNE a ambos conjuntos de datos.



(a) IRIS dataset (perplexity = 15 - learning rate = 200) (b) TMM dataset (perplexity = 50 - learning rate = 200)

Figura 3.2: Gráficos t-SNE.

Con respecto al dataset Iris se pueden ver en la figura 3.2a los tres clústers con más nitidez que en el gráfico obtenido a través de componentes principales (figura 3.1a).

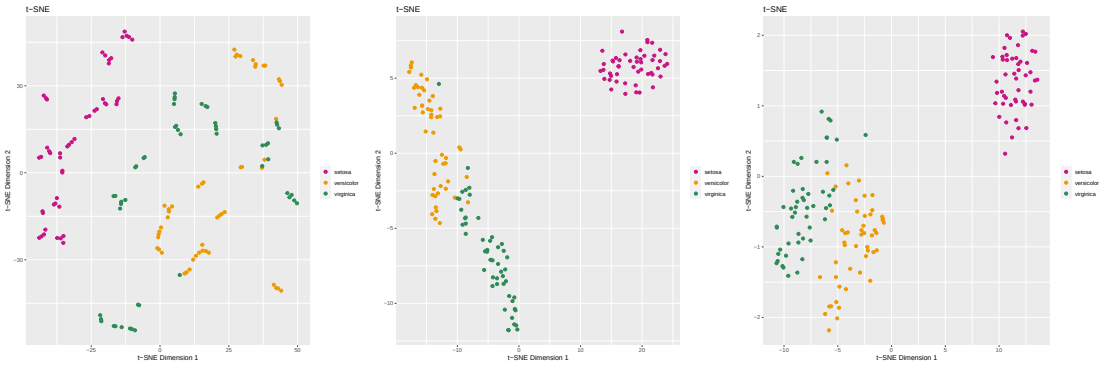
Para el conjunto de datos TMM, en la figura 3.2b, se presentan los diferentes clústeres según condición más separados unos de otros que como se obtenía mediante componentes principales (figura 3.1). Ahora se pueden apreciar cuatro clústers bien diferenciados mientras que con la técnica lineal solo se apreciaban tres.

Para comprender mejor la implementación de esta técnica resulta interesante poner a prueba de cambio algunos de sus parámetros.

Perplejidad

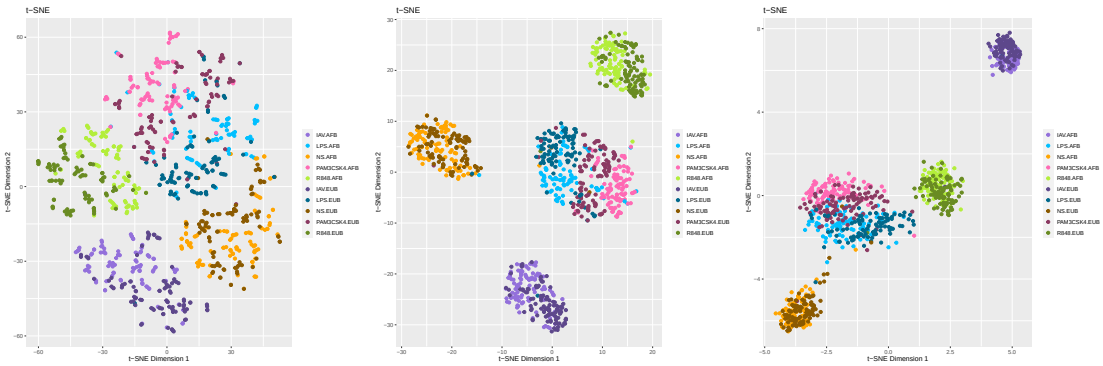
La perplejidad puede describirse como una aproximación del número de vecinos cercanos que tiene cada punto en el mapa de dimensión grande. Se trata de un parámetro difícil de ajustar [6].

A la vista de los resultados obtenidos en la figura 3.2, se puede apreciar la diferencia en la magnitud de este parámetro según la dimensión del conjunto de datos. Con objetivo de encontrar un valor adecuado de la perplejidad se va a aplicar t-SNE a ambos conjuntos de datos con diferentes valores del mismo.



(a) Perplexity = 2 (b) Perplexity = 20 (c) Perplexity = 40

Figura 3.3: Gráficos t-SNE para el dataset IRIS con distintas perplejidades y tasa de aprendizaje 200.



(a) Perplexity = 3 (b) Perplexity = 30 (c) Perplexity = 300

Figura 3.4: Gráficos t-SNE para el dataset TMM para distintas perplejidades con tasa de aprendizaje 200.

Tanto en la figura 3.3 como en la figura 3.4, se puede apreciar un cambio en la representación bidimensional de los datos ante distintos valores de la perplejidad.

Con respecto al dataset IRIS, se observa en la figura 3.3 que para valores muy pequeños (figura 3.3a) la visualización presenta una nube de puntos dispersados. Por otro lado, para valores más elevados de la perplejidad (figura 3.3c), aunque los clústers se mantienen, se observa que los puntos aparecen más alejados dentro de su mismo grupo que con una perplejidad moderada (figura 3.3b).

Para el conjunto de datos TMM, en la figura 3.4 con un valor de la perplejidad muy bajo (figura 3.4a), aunque se aprecian los diferentes clústers, no hay espacio entre ellos y los puntos se encuentran muy dispersos. Aumentando el valor del parámetro (figura 3.4b) disminuye la dispersión entre puntos facilitando

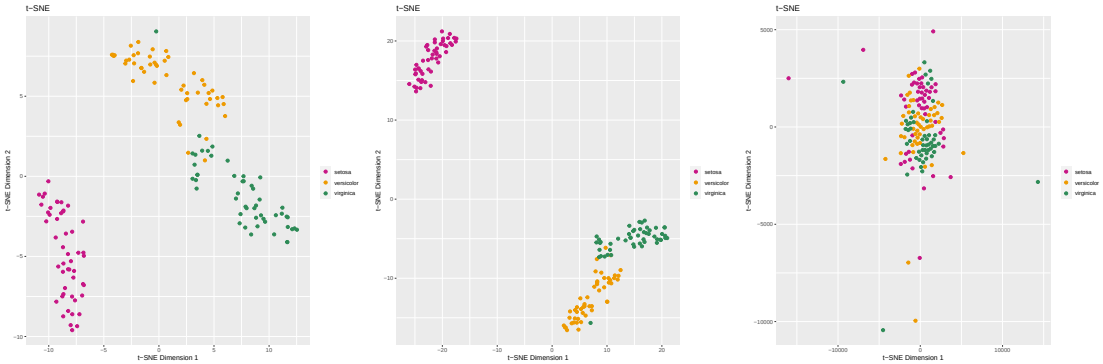
la distinción de los diferentes clústers. Con un valor de la perplejidad exagerado (figura 3.4c) los grupos tienden a converger perdiendo la estructura de clúster bien diferenciado.

La figura 3.3 y la figura 3.4 reflejan la sensibilidad de la técnica t-SNE ante cambios notables de la perplejidad: valores pequeños del parámetro suelen proporcionar mezclas homogéneas con datos dispersos y con valores exagerados puede ocurrir que los diferentes clústers se superpongan [6].

Tasa de aprendizaje

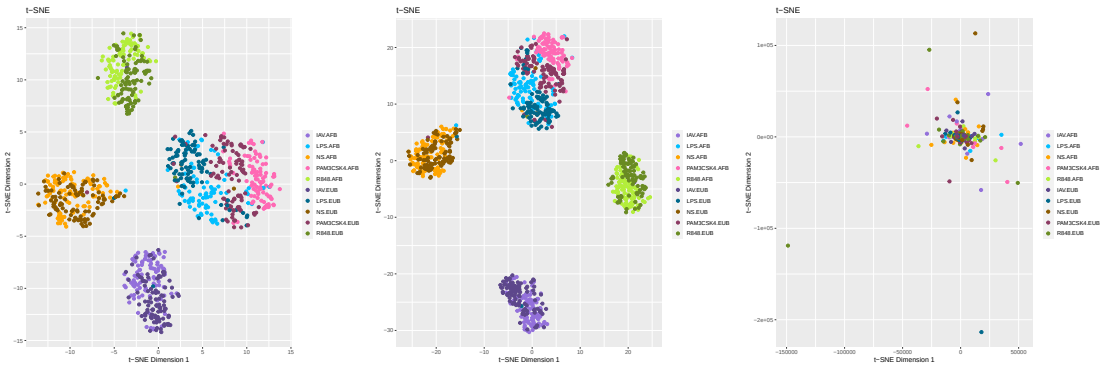
La tasa de aprendizaje o *learning rate* controla el tamaño del paso en el algoritmo del descenso del gradiente en la optimización. Por este motivo, se trata de un parámetro importante a la hora de minimizar la función coste del algoritmo t-SNE. Si toma valores muy pequeños, el modelo podría tardar mucho en converger o tomar como solución óptima un mínimo local. Por otro lado, con valores muy elevados del parámetro, el algoritmo podría no converger a la solución por haberla saltado [6].

Para ver cómo afectan estos cambios en el t-SNE, se van a mostrar varias representaciones bidimensionales de los resultado obtenidos con diferentes valores de la tasa de aprendizaje.



(a) Learning rate = 2 (b) Learning rate = 500 (c) Learning rate = 9.000

Figura 3.5: Gráficos t-SNE para el dataset IRIS para distintas tasas de aprendizaje con perplejidad 20.



(a) Learning rate = 2 (b) Learning rate = 800 (c) Learning rate = 10.000

Figura 3.6: Gráficos t-SNE para el dataset TMM para distintas tasas de aprendizaje con perplejidad 50.

Se observa en ambas visualizaciones (figura 3.5 y figura 3.6) que el cambio más notable se da al tomar una tasa de aprendizaje excesivamente grande ya que los puntos aparecen en el mapa dispersados sin seguir ningún patrón porque tienden a solaparse.

3.3.3. Número de clústers: coeficiente de silueta

Existen diversos criterios que permiten encontrar el número de clústers presentes en un conjunto de datos. Entre ellos se encuentra el método silhouette que se trata de una medida de lo próximo que está un dato a su grupo en comparación con otros grupos [11].

Se parte de un conjunto de datos agrupado mediante cualquier técnica, k-means[1] por ejemplo, en k clústers.

Definición 3.1. Se define la distancia media entre un punto $\mathbf{x}_i$ y los demás puntos dentro de su mismo clúster I como

$$a(\mathbf{x}_i) = \frac{1}{|I|-1} \sum_{\mathbf{x}_j \in I, j \neq i} d(\mathbf{x}_i, \mathbf{x}_j) \tag{3.1}$$

Se puede interpretar el valor de a($\mathbf{x}_i$) como lo bien que está el punto $\mathbf{x}_i$ asignado al clúster I. Cuanto menor sea su valor, mejor será la asignación a dicho clúster.

Definición 3.2. Se define la distancia media entre un punto $\mathbf{x}_i$ con todos los puntos de otro clúster J como

$$b(\mathbf{x}_i) = \min_{k \neq i} \frac{1}{|J|} \sum_{\mathbf{x}_j \in J} d(\mathbf{x}_i, \mathbf{x}_j) \tag{3.2}$$

Se dice que el clúster con menor valor de b($\mathbf{x}_i$) es el "clúster vecino" de $\mathbf{x}_i$ por ser el siguiente clúster que mejor se ajusta al punto $\mathbf{x}_i$.

Definición 3.3. Se define el valor de la silueta de un punto $\mathbf{x}_i$ del conjunto del datos como

$$s(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{\max\{a(\mathbf{x}_i), b(\mathbf{x}_i)\}}. \tag{3.3}$$

Si se da el caso en el que el número de clústers es igual a 1 se considera $s(\mathbf{x}_i) = 0$.

El valor de la silueta se encuentra entre -1 y 1. De acuerdo con la definición 3.3, un valor próximo a 1 de la silueta requiere $b(\mathbf{x}_i) \gg a(\mathbf{x}_i)$ y un valor grande de $b(\mathbf{x}_i)$ implica que un punto $\mathbf{x}_i$ está mal integrado con su clúster vecino. Por tanto, un valor próximo a 1 sugiere que los datos se presentan agrupados adecuadamente. El mismo razonamiento conduce a que un valor próximo a -1 de la silueta se debe a valores grandes de $a(\mathbf{x}_i)$, es decir, sería más apropiado considerar el punto $\mathbf{x}_i$ en el clúster vecino. Finalmente, un valor de la silueta próximo a cero implica $b(\mathbf{x}_i) \approx a(\mathbf{x}_i)$, es decir, que la distancia media entre los puntos de su mismo clúster y entre los puntos de otro clúster es similar. En este caso se considera que el punto $\mathbf{x}_i$ está en el borde de dos clústers.

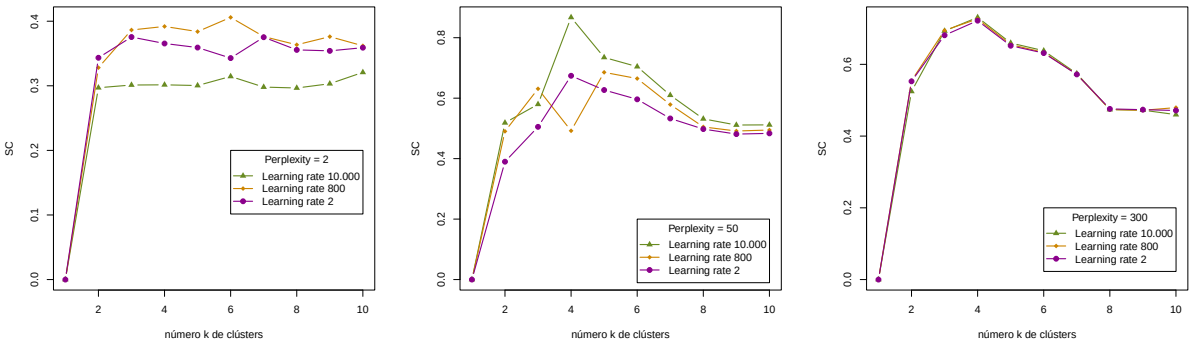
Para agrupar en k clústers se calcula la media de $s(\mathbf{x}_i)$ que denotamos por $\bar{s}$(k), y esto da un criterio sobre la calidad de esta división. El objetivo es encontrar el valor del número de clústers k_{opt} que maximiza este criterio

$$k_{opt} = \arg\max_k \bar{s}(\mathrm{k}) \tag{3.4}$$

Sensibilidad de k a los parámetros perplejidad y tasa de aprendizaje

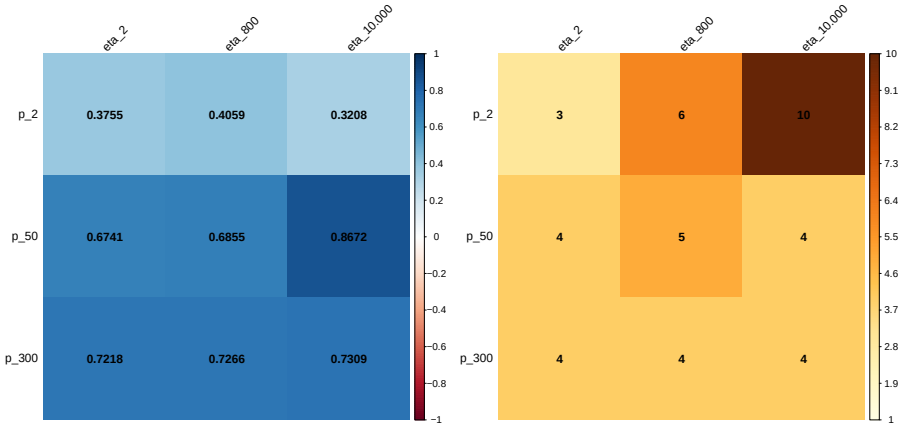
En este párrafo se busca analizar la sensibilidad del número de clústers k ante la variación de los parámetros perplejidad y tasa de aprendizaje a partir del coeficiente de silueta. Para ello, se propone la representación gráfica del coeficiente para distintos clústers y para diferentes valores de la perplejidad y de la tasa de aprendizaje (figura 3.7) con el conjunto de datos TMM.

[1] El algoritmo k-means es un algoritmo de clusterización que agrupa objetos en k grupos según sus características. La agrupación se realiza minimizando la suma de distancias (suele usarse la distancia cuadrática) entre cada objeto y el centroide de su grupo.



(a) Perplexity = 2 (b) Perplexity = 50 (c) Perplexity = 300

Figura 3.7: Coeficiente de silueta (SC) para distintos valores de k.



(a) Silhouette coefficient (SC) (b) Número de clústers k

Figura 3.8: Gráficos de calor entre perplejidad y tasa de aprendizaje.

Para un valor pequeño de la perplejidad (figura 3.7a) el número de clústers óptimo es 6 (figura 3.8b) ya que tiene el mayor coeficiente de silueta (figura 3.8a) y se obtiene con una tasa de aprendizaje igual a 800. Sin embargo, se observa que los coeficientes calculados con una perplejidad tan pequeña están muy por debajo del resto dando lugar a separaciones de los datos incorrectas, llegando a formar hasta 10 clústers (figura 3.8b).

El mayor coeficiente de silueta (figura 3.8a) se obtiene con perplejidad igual a 50 y tasa de aprendizaje igual a 10.000 distinguiéndose 4 clústers (figura 3.8b), coincidiendo con lo obtenido mediante la visualización gráfica de la técnica t-SNE (figura 3.2). En términos del coeficiente de silueta, se pueden considerar como parámetros óptimos en el dataset TMM una preplejidad igual a 50 y una tasa de aprendizaje igual a 10.000.

Visualización gráfica con los valores óptimos

Con perplejidad 50 y tasa de aprendizaje 10.000, se obtiene una visualización de la técnica t-SNE (figura 3.9) en la que se aprecian claramente los 4 clústers que indicaba el coeficiente de silueta (figura 3.8b). Aunque el t-SNE no consiga separar dos clústers diferenciando los estímulos virales LPS y PAM3CSK4, dentro del mismo clúster la técnica proporciona una visualización gráfica en la que se pueden diferenciar.

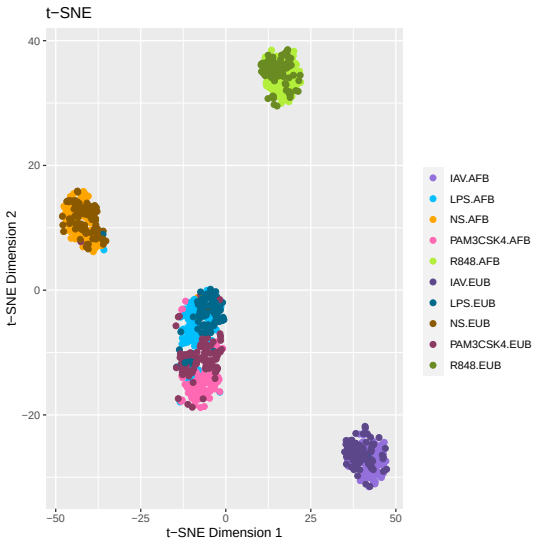
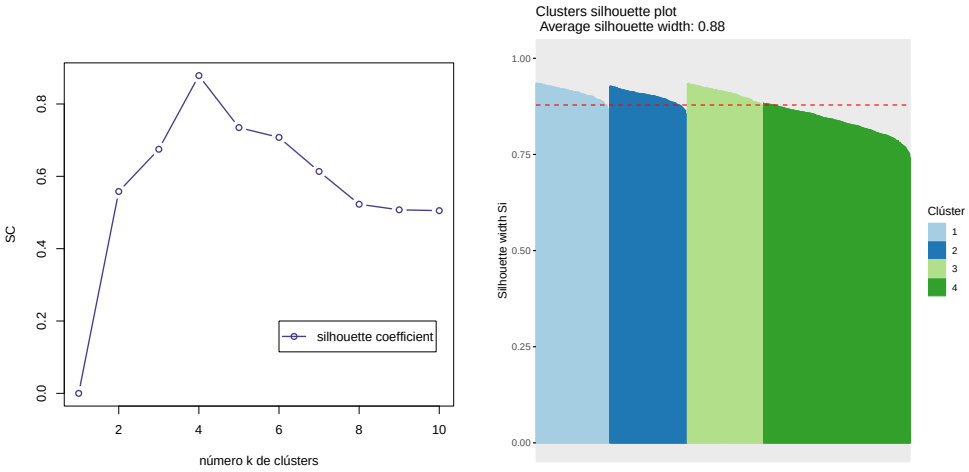


Figura 3.9: Gráfico t-SNE para valores óptimos (perplexity $= 50$ y $\eta = 10.000$).

Para estos valores óptimos, se obtiene un coeficiente de la silueta próximo a 0,88 y se consideran 4 los clústers en los que se divide el conjunto de datos TMM (figura 3.10).



(a) Silhouette coefficient

(b) perplexity = 50 - learning rate = 10.000

Figura 3.10: Coeficiente de silhouette para distintos valores de *k* y gráfico de silueta con valores óptimos.

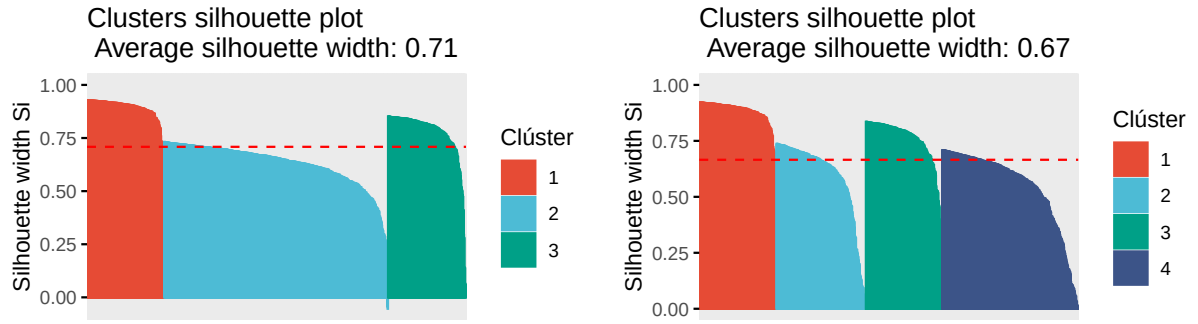
En la tabla cruzada (cuadro 3.1) según la condición experimental y el clúster al que ha sido asignado (figura 3.10b), se observa que la agrupación realizada es muy cercana a la realidad del dataset ya que en cada clúster aparecen muestras celulares bajo la misma condición (y de ambas poblaciones, africana y europea). Cabe destacar que las muestras del tipo LPS y PAM3CSK4 comparten clúster, hecho que se refleja desde el inicio del análisis (figura 3.2).

clúster/condición	1	2	3	4
IAV.AFB	0	98	0	0
IAV.EU	0	98	1	0
LPS.AFB	86	0	0	1
LPS.EU	94	1	0	1
NS.AFB	2	0	0	98
NS.EU	2	0	0	98
PAM3CSK4.AFB	96	0	0	0
PAM3CSK4.EU	99	0	0	1
R848.AFB	0	0	93	0
R848.EU	1	0	97	0

Cuadro 3.1: Tabla cruzada con condición.

Componentes principales y t-SNE

Por último, se presenta una comparativa del gráfico de silueta con la técnica de componentes principales para 3 y 4 clústers (figura 3.11). El promedio del coeficiente de la silueta es mayor con esta técnica considerando 3 clústers en lugar de 4 como ya se había apreciado en su visualización gráfica (figura 3.1b). Por este motivo, la técnica t-SNE proporciona un número de clústers (4 clústers) más cercano al número real de 5 condiciones experimentales que la técnica de componentes prinipales (3 clústers).



(a) ACP con $k = 3$

(b) ACP con $k = 4$

Figura 3.11: Gráficas de silhouette para distintos valores de k con el método de componentes principales.

3.3.4. Lectura e interpretación de los resultados

A la vista de los resultados, en la figura 3.3 y en la figura 3.4 se aprecia que, para valores de la perplejidad pequeños y con una tasa de aprendizaje moderada, la técnica t-SNE proporciona una nube de puntos homogénea con apenas espacio entre los clústers. Aumentando el valor del parámetro se consigue una mayor separación entre clústers, aunque con valores excesivamente elevados se llegan a solapar unos con otros.

Con respecto a la tasa de aprendizaje, se refleja en la figura 3.5 y la figura 3.6 que en un conjunto de datos de menor dimensión el cambio de este parámetro de un valor pequeño a uno más elevado, con valores moderados de la perplejidad, es mayor que en datasets de grandes dimensiones. Para valores excesivamente elevados de la tasa de aprendizaje los puntos se dispersan sin ningún patrón impidiendo distinguir ningún clúster.

El coeficiente de silueta permite encontrar el número óptimo de clústers k existentes en el conjunto de datos. Sabiendo que las muestras pertenecen a 5 tipos diferentes de condiciones, la técnica t-SNE se

acerca más a dicho valor de clústers ofreciendo 4 diferentes bien separados, mientras que con la técnica de componentes principales solo se obtienen 3. Se ha alcanzado un valor promedio de la silueta próximo a 1 con valores de perplejidad 50 y tasa de aprendizaje 10.000 que se han considerado óptimos para este ejemplo. Además, en el gráfico de calor (figura 3.8) se observa que el número de clústers k se aleja del esperado con valores pequeños de la perplejidad mientras que para valores pequeños de la tasa de aprendizaje se obtiene un número de clústers esperado si se acompaña de una perplejidad adecuada.

3.4. Implementación en R

La implementación de la técnica t-SNE se lleva a cabo en R mediante la función Rtsne disponible en un paquete con el mismo nombre [9].

```
Rtsne(
  X, dims = 2, perplexity = 30, theta = 0.5, pca = TRUE, momentum = 0.5, eta = 200
  ... )
```

3.5. Conclusiones

A partir de los resultados obtenidos, se ha comprobado que la técnica no lineal de reducción de la dimensionalidad t-Distributed Stochastic Neighbor Embedding (t-SNE) ha proporcionado mejores resultados que la técnica lineal de componentes principales. Además, se ha estudiado la sensibilidad del número de clústers ante los dos parámetros más significantes en esta ténica no lineal.

- Valores pequeños de la perplejidad provocan una nube de puntos en el plano creando pequeños clústers inexistentes en el dataset. Aumentando su valor los puntos se juntan definiendo mejor los diferentes clústers. Sin embargo, un valor excesivamente grande provoca que los grupos se solapen impidiendo ser distinguidos. Esto hechos también se reflejan en los valores del coeficiente de silueta ya que los asociados a perplejidades pequeñas son muy bajos y proporcionan divisiones del dataset no correspondientes a las 5 condiciones experimentales.

- Con valores pequeños de la tasa de aprendizaje el número de clústers se distinguirá siempre que se realicen un número de iteraciones mínimo que permitan encontrar el óptimo. Aumentando moderadamente el valor se consigue agrupar los puntos y separar los diferentes clústers. Si la tasa de aprendizaje toma valores excesivamente grandes los puntos se dispersan en mapa formando un círculo indefinido.

En cuanto a la magnitud de ambos parámetros se ha apreciado sus valores dependen de la dimensión del conjunto de datos que se está tratando, pues conjuntos con muchas observaciones admiten perplejidades y tasas de aprendizaje mayores que datasets con menos observaciones.

Aunque la técnica t-SNE es conocida como una técnica de visualización, ésta puede ser aplicada como un camino para conocer el número de clústers de un dataset y para evaluar la disposición de los datos en cada uno de los clústers existentes, es decir, para realizar un análisis clúster. Éste consiste en agrupar puntos que presenten características similares y a su vez separarlos de aquellos que tengan propiedades diferentes. La técnica t-SNE proporciona una visualización de los datos en el espacio de dimensión pequeña que puede ser útil para obtener una primera aproximación del número de clústers o para comprobar si el número de clústers obtenido por otros procedimientos es correcto. A pesar de conocer estas ventajas, no hay que olvidar que t-SNE no se trata de una técnica enfocada a la agrupación sino a la visualización de los datos en el espacio de dimensión pequeña.

Este estudio para encontrar la mejor representación gráfica que proporciona el método t-SNE ha servido para comprender la importancia de conocer el conjunto de datos con el que se está tratando. Por este motivo, el t-SNE puede ser considerada como una técnica de reducción de la dimensionalidad de análisis presupervisado, en lugar de no supervisado, ya que su mejor salida aparece cuando se tiene conocimiento sobre el conjunto de datos.

Bibliografía

[1] J. LEVER, M. KRZYWINSKI Y N. ALTMAN, *Principal Component Analysis*, Nat Methods 14, 641–642 (2017). Disponible en https://doi.org/10.1038/nmeth.4346.

[2] C. SÁNCHEZ, *Análisis Multivariante*, Máster en Técnicas Estadísticas (2008-2009). Disponible en http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP/MATERIALESMASTER/Mat_14_master0809multi-tema7.pdf.

[3] L. VAN DER MAATEN Y GEOFFREY HINTON, *Visualizing Data using t-SNE*, Journal of Machine Learning Research 9, 2008.

[4] D. PEÑA, *Análisis de Datos Multivariantes*, McGraw Hill, 2002.

[5] JAN R. MAGNUS Y HEINZ NEUDECKER, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley, 1999.

[6] C. ROBLEDO, *Visualización de conjuntos de datos de alta dimensionalidad con t-SNE*, Universidad de Oviedo, Máster en Análisis de Datos e Inteligencia de Negocios, 2019.

[7] E. FETAYA, J. LUCAS Y E. ANDREWS, *CSC 411 Lecture 13: t-SNE*, University of Toronto.

[8] *Kullback–Leibler divergence*. (9 de junio de 2022). En *Wikipedia*. Disponible en https://en.wikipedia.org/w/index.php?title=Kullback%E2%80%93Leibler_divergence&oldid=1092039611.

[9] RSTUDIO TEAM (2022), *RStudio: Integrated Development Environment for R*. Boston, MA. Disponible en http://www.rstudio.com/.

[10] H. QUACH, *Genetic Adaptation and Neardental Admixture Shaped the Inmune System of Humans Populations*. Disponible en http://dx.doi.org/10.1016/j.cell.2016.09.024.

[11] F. WANG, H. FRANCO-PENYA, J. KELLEHER, J. PUGH Y R. ROSS, *An Analysis of the Application of Simplified Silhouette to the Evaluation of k-means Clustering Validity*. School of Computing, Dublin Institute of Technology, Ireland.

[12] A. SÁNCHEZ, *El Algoritmo del Gradiente Descendente*. Fundamental Nerve, 2020.

Anexos

Anexo A: Divergencia de Kullback-Leibler (KL)

Dadas dos distribuciones de probabilidad, $\eta \ll \mu$, se define la divergencia de Kullback-Leibler (KL) de η a μ como

$$KL(\mu||\eta) = -\mathbf{E}_\mu \left[\log \frac{\partial \eta}{\partial \mu}\right].$$

Lema. *La divergencia de Kullback-Leibler (KL) verifica las siguientes propiedades:*

1. *KL*≥ 0 *y se da la igualdad si* $\eta = \mu$.
2. *No es simétrica (no es una distancia).*

Demostración. 1. Se deduce aplicando la *desigualdad de Jensen*.

$$\mathbf{E}_\mu \left[\log \frac{\partial \eta}{\partial \mu}\right] \leq \log \mathbf{E}_\mu \left[\frac{\partial \eta}{\partial \mu}\right] = \log 1 = 0.$$

La segunda parte se deduce aplicando de nuevo la *desigualdad de Jensen* para el caso de la igualdad.

$$0 = \mathbf{E}_\mu \left[\log \frac{\partial \mu}{\partial \mu}\right] \leq \log \mathbf{E}_\mu \left[\frac{\partial \mu}{\partial \mu}\right] = \log 1 = 0.$$

2. Se sigue inmediatamente de su definición.

□

Una de las aplicaciones más útiles de la *divergencia de Kullback-Leibler* (KL) en estádistica está relacionada con el método de ajuste de distribuciones por máxima verosimilitud.

Dadas $x_1, \cdots, x_n$ observaciones independientes de una variable aleatoria de función de densidad f desconocida y se tratan de ajustar dentro de una familia de funciones de densidad f_λ. Aplicando máxima verosimilitud el objetivo es encontrar el parámetro λ que maximiza la función

$$\mathbf{L}_\lambda = \sum_i log\ f_\lambda(x_i),$$

que considerando valores de n grandes puede aproximarse como $\mathbf{L}_\lambda = \int f(x) log\ f_\lambda(x)\, dx$. Si a esta última expresión le restamos $\int f(x) log f(x)\, dx$ se tiene

$$\int f(x) log\ f(x)\, dx - \int f(x) log\ f_\lambda(x)\, dx = \int f(x) log \left(\frac{f_\lambda(x)}{f(x)}\right) dx,$$

que es la *divergencia de Kullback-Leibler* (KL) entre f_λ y la verdadera distribución f. Buscar el parámetro λ que maximiza la verosimilitud es (aproximadamente) equivalente a encontrar λ que minimiza la *divergencia de Kullback-Leibler* (KL) entre la distribución real y la familia de distribuciones parametrizadas por el parámetro.

Anexo B: Método del gradiente descendente

El método del gradiente descendente (GD) es un algoritmo de optimización genérico muy utilizado en algoritmos de *Machine Learning* que mide cuánto varía la salida de la función objetivo si modificamos alguno de sus parámetros. Su objetivo es encontrar el valor óptimo de los parámetros de forma iterativa que minimizan la función objetivo.

Se puede considerar el gradiente como una generalización de la derivada que representa la pendiente en el punto en el que nos encontramos de la función objetivo.

Empezamos explicando el método considerando una función objetivo $J : \mathbb{R} \to \mathbb{R}$ que depende de un solo parámetro ω. El algortimo comienza evaluando la función objetivo en un punto arbirario y actualiza el parámetro siguiendo la siguiente regla

$$\omega_{t+1} = \omega_t - \eta \frac{\partial}{\partial \omega} J(\omega_t),$$

donde η es un hiperparámetro que se denomina tasa de aprendizaje y debido a su importancia en el algoritmo se profundizará en él más adelante.

De acuerdo con esta expresión, si la pendiente de la tangente es positiva el valor del parámetro aumenta y también aumentaría la función objetivo por lo que deberíamos movernos en dirección opuesta. Por el contrario, si la pendiente es negativa el valor de ω disminuye y, en consecuencia, el de la función objetivo, por lo que el algoritmo se mantiene en esa dirección.

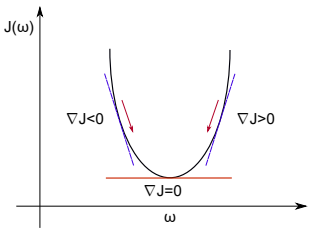
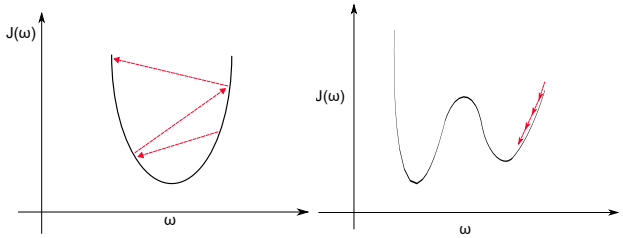


Figura 12: Gradiente descendente para una función $J : \mathbb{R} \to \mathbb{R}$ dependiente de un parámetro.

En general, dada una función convexa $J : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ se pretende encontrar el parámetro $\omega \in \Omega$ que minimiza $J(\omega)$. Para ello se plantea el siguiente algoritmo que se suele inicializar en cero y cada iteración se hace en la dirección negativa del gradiente $\nabla J(\omega) = \left(\frac{\partial J}{\partial x_1}(\omega), \cdots, \frac{\partial J}{\partial x_n}(\omega)\right)$

$$\omega_{t+1} = \omega_t - \eta_t \nabla J(\omega_t),$$

donde $\eta_t > 0$ es la tasa de aprendizaje. Su principal función es controlar la velocidad de búsqueda ya que determina el paso del descenso del gradiente durante el proceso de optimización. Determinar este parámetro no es trivial ya que si se toma demasiado pequeño (figura 13b) se deberán realizar varias iteraciones del algoritmo y puede quedarse atrapado en mínimos locales mientras que si se toma muy grande (figura 13a) puede ocurrir que no converja.



(a) Elevada tasa de aprendizaje (b) Pequeña tasa de aprendizaje

Figura 13: Gradiente descendente para valores diferentes de η.

Se puede determinar el valor de η resolviendo el siguiente problema de optimización

$$\eta_t \in \arg\min_{\eta>0} J(x_t - \eta \nabla J(x_t))$$

Resolver este problema es común en problemas de carácter cuadrático y donde la evaluación del gradiente es costosa, en otro caso no es aconsejable [12].

Debido a la importancia de este algoritmo de optimización existen variantes del mismo entre las que destacan: gradiente descendente estocástico y gradiente descendente con momento.

Gradiente descendente estocástico

Para esta variante del método se considera un problema de optimización en el que la función objetivo pueda escribirse como un sumatorio, esto es

$$\arg\min J(x) = \frac{1}{|\Omega|} \sum_{i \in \Omega} J_i(x).$$

Así el algoritmo queda como

$$\omega_{t+1} = \omega_t - \frac{1}{|\Omega|} \sum_{i \in \Omega} \nabla J_i^T,$$

donde el segundo término puede ser interpretado como un valor esperado $\sum \nabla J_i^T = \mathbb{E}\{\nabla J_i^T\}$.

El término estocástico proviene del hecho de considerar en cada iteración, en vez de la suma sobre toda la población Ω, sumar sobre una muestra $S \subset \Omega$.

De esta forma, el algoritmo que se plantea es

$$\omega_{t+1} = \omega_t - \frac{1}{|S|} \sum_{i \in S} \nabla J_i^T.$$

La principal ventaja del método es la disminución de cálculo en iteraciones realizadas al seleccionar una muestra en vez de toda la población. Este algoritmo es útil cuando la función objetivo es la suma de costos individuales sobre un conjunto muy grande ya que la muestra suele ser muy representativa y reproduce un valor próximo al de la población.

Sin embargo, hay que tener en cuenta la existencia de datos atípicos (outliers) ya que las muestras pueden ser robustas a grandes desviaciones.

Si la función objetivo tiene muchos mínimos locales pequeños, el gradiente estocástico permite suavizar la función objetivo y reduce el riesgo de tener una convergencia temprana.

Gradiente descendente con momento

El gradiente descendente (GD) se caracteriza por su aproximación constante hacia el mínimo local. Esta propiedad tiene como principal inconveniente presentar una convergencia lenta en caso de estar inicialmente lejos del óptimo.

El método del gradiente descendente con momento (MGD) propone incluir un término adicional que tenga en cuenta el valor de la actualización aplicada en la iteración anterior, de esta forma se estarán teniendo en cuenta los gradientes anteriores además del actual. El gradiente actual se multiplicará por la tasa de aprendizaje (η) y el valor de la actualización anterior por una constante conocida como coeficiente del momentum (α)

$$\omega_{t+1} = \omega_t + \nabla \omega_t,$$

siendo $\nabla \omega_t = \alpha(t) \nabla \omega_{t-1} - \eta_t \nabla J(x_t)$.