

Feature Extraction (1)

(Extracción de características)

Reducción de la Dimensionalidad

Dos estrategias:

Selección de características: se selecciona un subconjunto de los atributos originales.

- **Algoritmos de Filtrado:** la calidad de los atributos se mide usando alguna medida estadística general (Chi Cuadrado, Correlación, Información Mutua, etc.)
- **Algoritmos de Wrapping:** la calidad de un conjunto de atributos se evalúa calculando la calidad promedio de un modelo entrenado con esos atributos
- **Algoritmos embebidos:** la selección es parte del algoritmo de aprendizaje

Extracción / construcción de características: la idea es construir nuevas características que condensen / resuman la información relevante de los atributos originales.
Por ejemplo PCA, LDA

Reducción de la dimensionalidad con transformaciones de los atributos (1)

- Idea: encontrar una transformación $y=f(x)$ que conserve la información acerca del problema, minimizando el número de componentes
- En general, la función óptima $y=f(x)$ será no lineal
- Sin embargo, no hay una forma de generar sistemáticamente transformaciones no lineales:
 - ▣ La selección de un subconjunto particular de transformaciones depende del problema
 - ▣ Por esta razón, la limitación a transformaciones lineales ha sido ampliamente aceptada, $y = W^T x$
➔ **y es una proyección lineal de x**

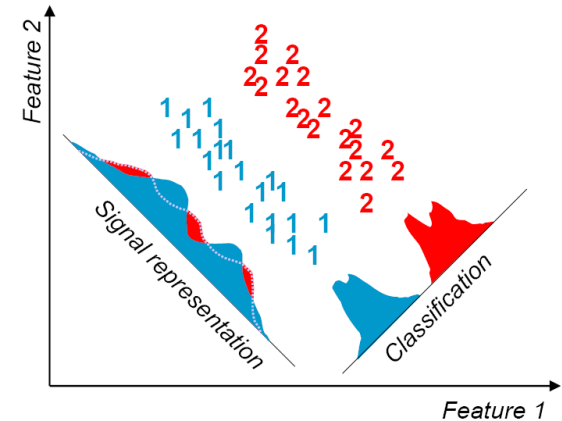
Reducción de la dimensionalidad con transformaciones de los atributos (2)

$$\begin{array}{ccc}
 \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} & \xrightarrow{\text{Transformación lineal}} & \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \cdots \\ w_{21} & w_{22} & \cdots \\ \vdots & \vdots & \ddots \\ w_{M1} & w_{M2} & \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \\
 & & \begin{array}{cc} \text{M-dimensional} & \text{N-dimensional} \\ M < N & \end{array}
 \end{array}$$

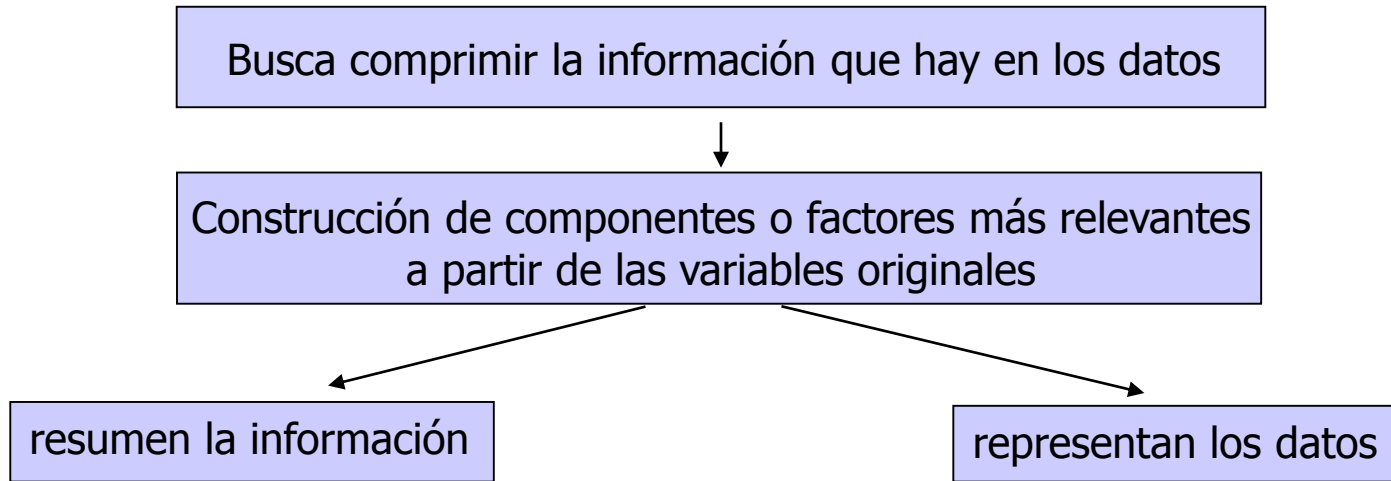
- Por el momento nos centraremos en transformaciones lineales

Representación de la señal versus clasificación (PCA vs. LDA)

- La selección de la transformación extractora de características, $\mathbf{y}=\mathbf{f}(\mathbf{x})$, está guiada por una función objetivo que buscamos maximizar (o minimizar)
- Dependiendo del criterio usado por la función objetivo, las técnicas de extracción de características se dividen en dos categorías:
 - **Clasificación:** El objetivo de la transformación extractora de características es resaltar en un espacio de menos dimensiones la información discriminante de clases
 - **Representación de la señal:** El objetivo de la transformación extractora de características es representar los vectores de atributos de manera precisa en un espacio de menos dimensiones
- Hay dos técnicas principales en la extracción lineal de características:
 - Análisis Discriminante Lineal (LDA), que utiliza el criterio de clasificación
 - Análisis de Componentes Principales (PCA), que usa el criterio de representación de la señal



PCA: Definición y objetivo



Simplificar la estructura de los datos transformando las variables originales en otras llamadas componentes principales a través de combinaciones lineales de las mismas:

$$y = W^T x$$

Definición formal de PCA (1)

- El objetivo de **PCA** es realizar una **reducción de la dimensionalidad** preservando lo máximo posible la información contenida en los datos originales **sin tener en cuenta la clase**.
- PCA busca reducir la dimensionalidad proyectando los datos originales en M ejes $\{ \mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M \}$ ($M < N$, y cada uno de los vectores \mathbf{w} tiene longitud 1). Los datos comprimidos \mathbf{y} tienen M dimensiones, el componente i es $\mathbf{y}_i = \mathbf{w}_i^T \mathbf{x}$

Definición formal de PCA (2)

- Supongamos sin pérdida de generalidad que la media de los datos \mathbf{x} es $\mathbf{0}$. En el caso de que nos interese comprimir los datos a una sola dimensión ($M=1$) la proyección de cada punto \mathbf{x} en el eje \mathbf{w}_1 es:

$$\tilde{\mathbf{x}} = (\mathbf{w}_1^T \cdot \mathbf{x})$$

- ¿Cuál es el eje \mathbf{w}_1 que maximiza la información que se mantiene de los datos originales? \Rightarrow el que minimiza el promedio del error $\|\tilde{\mathbf{x}} - \mathbf{x}\|$ calculado a lo largo de la base de datos de entrenamiento
- Se puede demostrar que el \mathbf{w}_1 óptimo es el autovector de Σ (matriz de covarianza de los datos originales) con mayor autovalor, y normalizado para que tenga longitud 1.

Definición formal de PCA (3)

- ¿Qué es un autovector de Σ ? Es un vector \mathbf{u} que cumple:

$$\Sigma \cdot \mathbf{u} = \lambda \cdot \mathbf{u}$$
- El número λ es el “autovalor” correspondiente al autovector \mathbf{u}

Propiedades:

1. El autovalor λ_i es exactamente igual a la varianza de la componente y_i .
Su valor no puede ser negativo.
2. Si se asume que la nube de puntos es Gaussiana, la nube tiene una forma “elipsoidal”. Los autovectores de Σ representan los ejes de simetría de esta elipsoide
3. Si \mathbf{w}_i y \mathbf{w}_j son dos autovectores que tienen diferentes autovalores, entonces son perpendiculares

Definición formal de PCA (4)

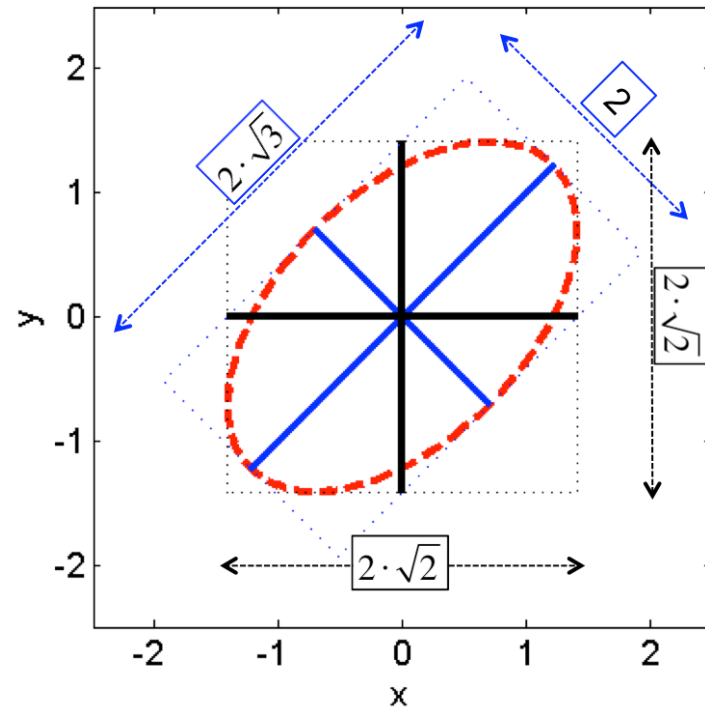
Ejemplo: $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$

Varianza de atributo 1 = varianza de atributo 2 = 2

- $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \cdot \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 3 \\ 3 \end{bmatrix} = 3 \cdot \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$
- $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \cdot \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = 1 \cdot \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$
- $\Rightarrow \mathbf{w}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ autovector con autovalor 3
- $\Rightarrow \mathbf{w}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ autovector con autovalor 1

Varianza en eje principal 1 = 3

Varianza en eje principal 2 = 1



Definición formal de PCA (5)

- Supongamos que queremos comprimir los datos a M dimensiones. ¿Cuáles son los ejes \mathbf{w}_i que maximizan en conjunto la información que se mantiene de los datos originales? \Rightarrow los que minimizan el promedio del error $\|\tilde{\mathbf{x}} - \mathbf{x}\|$ calculado a lo largo de la base de datos de entrenamiento
- Se puede demostrar que el conjunto $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M\}$ óptimo es el formado por los M autovectores de Σ (matriz de covarianza de los datos originales) con autovalores más grandes, y normalizados para que tengan longitud 1

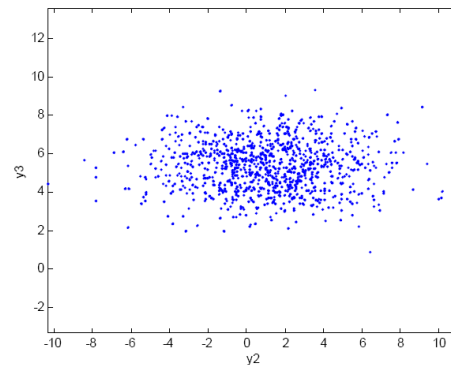
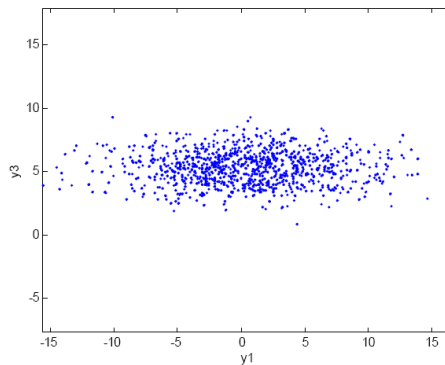
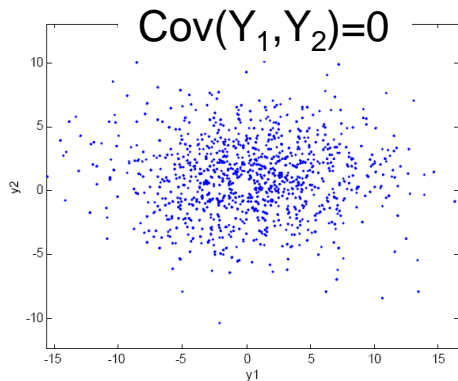
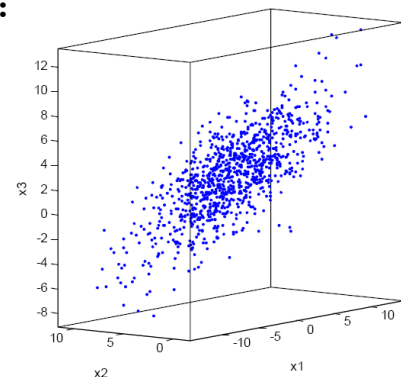
PCA: Ejemplo 1

- Ejemplo: distribución gaussiana en tres dimensiones con los siguientes parámetros:

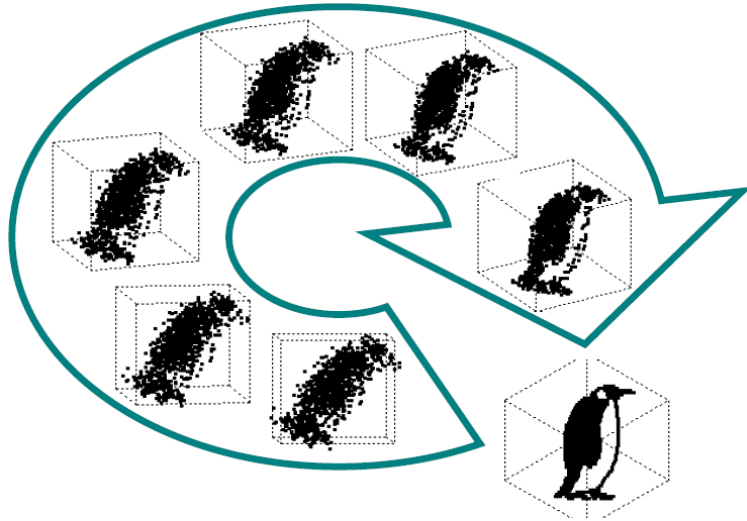
$$\mu = [0 \ 5 \ 2]^T \quad \Sigma = \begin{bmatrix} 25 & -1 & 7 \\ -1 & 4 & -4 \\ 7 & -4 & 10 \end{bmatrix}$$

- Ahora mostramos los tres pares de proyecciones en los componentes principales

- La primera proyección tiene la mayor varianza, seguida por la segunda
- Las proyecciones PCA “decorrelacionan” los ejes. $\text{Cov}(Y_i, Y_k) = 0$

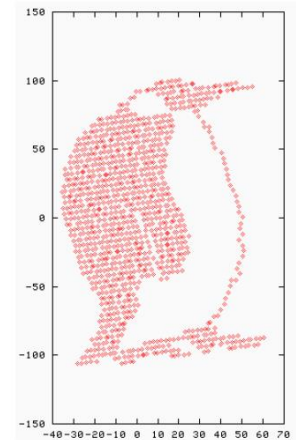


PCA: Ejemplo 2



- Ahora tenemos una nube de datos en 3 dimensiones
- Inicialmente, excepto por un alargamiento en la nube de puntos, no hay estructura aparente
- Elegir una rotación apropiada nos permite descubrir la estructura que hay por debajo (podemos pensar en esta rotación como el “caminar” en 3 dimensiones, buscando el mejor punto de vista).

- PCA nos puede ayudar en encontrar la estructura implícita en nuestros datos. Selecciona una rotación tal que casi toda la variabilidad de los datos es representada en las primeras componentes principales.
 - En nuestro ejemplo no parece de mucha ayuda.
 - Sin embargo, cuando tenemos docenas de dimensiones, PCA es muy potente



PCA: Ejemplo 3

- Tenemos los siguientes datos en dos dimensiones:

$$X=(x_1,x_2)=\{(1,2),(3,3),(3,5),(5,4),(5,6),(6,5),(8,7),(9,8)\}$$

- La estimación del promedio y de la matriz de covarianza es:

$$\mu = \begin{bmatrix} 5 \\ 5 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 7.14 & 4.86 \\ 4.86 & 4.00 \end{bmatrix}$$

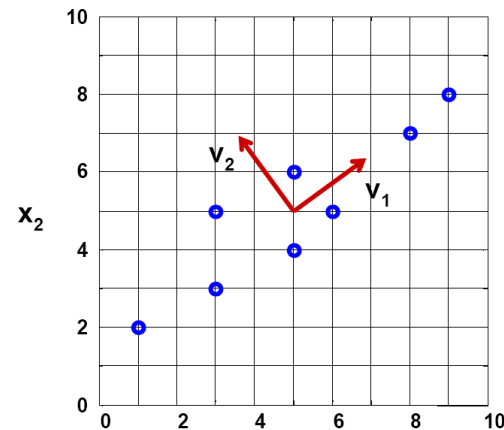
Los autovectores son:

$$\Sigma \omega = \lambda \omega \Rightarrow |\Sigma - \lambda I| = 0$$

$$\mathbf{w}_1 = \begin{bmatrix} 0.81 \\ 0.59 \end{bmatrix} \text{ con autovalor } 10.68$$

$$\mathbf{w}_2 = \begin{bmatrix} -0.59 \\ 0.81 \end{bmatrix} \text{ con autovalor } 0.47$$

- La primera componente explica el $10.68/(10.68+0.47) \cdot 100 = 95.8\%$ de los datos originales



Comentarios sobre PCA (1)

- Ya que PCA elige los autovectores de la matriz de covarianza Σ_x , es capaz de encontrar los ejes independientes de los datos cuando éstos están distribuidos gaussianamente
 - Para distribuciones no Gaussianas (multimodales, por ejemplo), PCA simplemente *decorrelaciona* los ejes (las nuevas variables tienen correlación 0 entre ellas).
- La principal limitación de PCA es que no tiene en cuenta la separabilidad de las clases ya que no tiene en cuenta las clases de los vectores \mathbf{x} ➔ **Método no supervisado**
 - PCA simplemente realiza una rotación de coordenadas que alinea los ejes transformados con las direcciones de máxima varianza.
 - **No hay garantía alguna de que los ejes de máxima varianza contengan una buena información para la clasificación**

Comentarios sobre PCA (2)

- Comentarios históricos:
 - PCA es la técnica más antigua de análisis multivariable
 - Se conoce también como “transformada de Karhunen-Loève” en otros campos como la Teoría de la Comunicación y la Física.

Uso práctico de PCA (1)

Matemáticamente: Partiendo de N variables iniciales

Construcción de una matriz a partir de los datos de partida

Matriz de Covarianza si los datos no están estandarizados
Matriz de Correlaciones si los datos están estandarizados

Cálculo de los autovectores y autovalores de la matriz

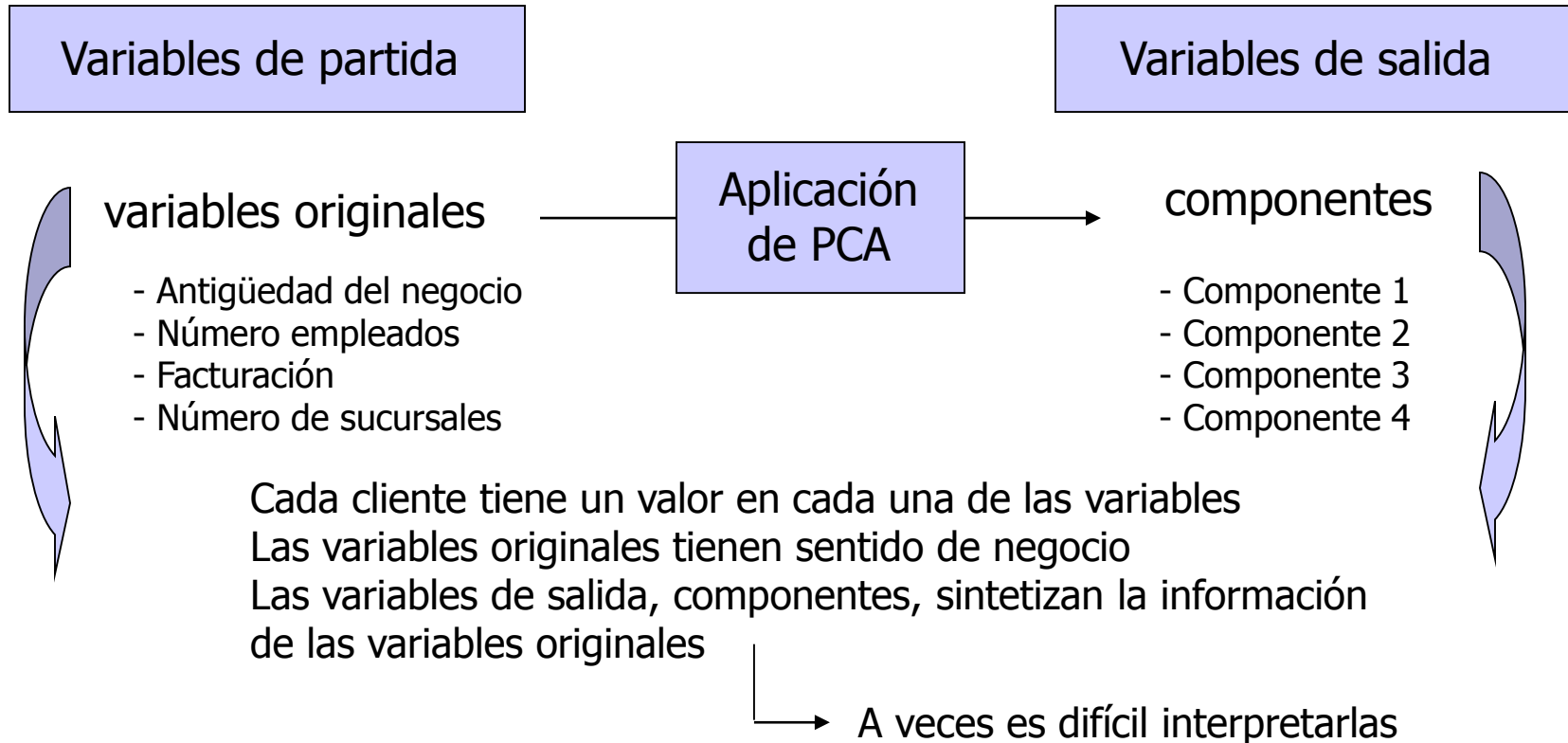
λ_i y $(w_{i1}, w_{i2}, \dots, w_{iN})^T$ para $i = 1, \dots, N$

Cálculo de los nuevos atributos ("componentes")

Las componentes sintetizan la información de las variables

$$\begin{cases} y_1 = w_{11} \cdot x_1 + \dots + w_{1N} \cdot x_N \\ \vdots \\ y_M = w_{M1} \cdot x_1 + \dots + w_{MN} \cdot x_N \end{cases}$$

Uso práctico de PCA (2)



Uso práctico de PCA (3)

Tipo de las variables

Por las propiedades y características del modelo **sólo está permitido el uso de variables numéricas**

Si se dispone de variables categóricas que describan directamente el problema:



Para que intervengan en el análisis de componentes principales, es necesario convertirlas en numéricas



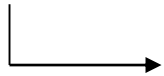
Transformar la variable creando variables dicotómicas

Uso práctico de PCA (4)

Escala de medida → PCA no es invariante a la escala!!!

Las escalas en la que estén medidas las variables influyen en esta técnica

Si no se quiere dar importancia a una variable por la escala en la que viene medida


 Normalización (o “estandarización”) de los datos

Correlaciones

PCA es una técnica que tiene sentido aplicarse en el caso de existir **correlaciones altas** entre las variables (indicio de que existe información redundante).

⇒ Como consecuencia, pocos factores explicarán gran parte de la variabilidad total.

Uso práctico de PCA (5)

- Sea \mathbf{X} la matriz de datos de entrenamiento (cada columna un patrón, cada fila un atributo), y consideremos que los atributos han sido previamente estandarizados.
- En este caso la matriz de covarianza equivale a la de correlación, y es igual a

$$\Sigma = \frac{1}{N_{\text{tr}} - 1} \cdot \mathbf{X} \cdot \mathbf{X}^T$$

- En algunas situaciones el número de dimensiones N de los datos de partida es mucho mayor que el número de patrones de entrenamiento N_{tr} . Por ejemplo:
 - En bases de datos de genomas
 - En bases de datos de imágenes
 - En bases de datos de audios
- Esto hace que el tamaño de la matriz Σ sea muy grande, y por tanto el cálculo de los autovectores sea muy costoso. Por otra parte, muchos de los autovalores serán nulos (una condición necesaria para que no salga ningún autovalor nulo es que $N_{\text{tr}} > N$)

Uso práctico de PCA (6)

Características de los componentes o factores

La correlación entre factores diferentes es 0, $\text{Cov}(y_i, y_j)=0$

Los primeros tienen más relevancia (más información) que los últimos

Los factores sintetizan la información de las variables originales

Si los atributos antiguos estaban estandarizados:

Los nuevos atributos ("componentes" o "factores") tienen media 0

La varianza de cada factor es exactamente igual al autovalor asociado

El Análisis de Componentes Principales estudia las relaciones que las variables tienen entre sí, descubriendo grupos de variables muy correlacionadas entre sí

Uso práctico de PCA (7)

Elección de componentes

Si nos quedamos con N componentes no reducimos dimensionalidad

Criterios para la ayuda a la elección de factores o componentes

- 1.- Seleccionar los componentes necesarios para sumar al menos el 80% del valor total de la varianza
- 2.- En el caso de utilizar la matriz de correlación, seleccionar todos los factores con autovalor ≥ 1 .
En caso de utilizar la matriz de covarianza, seleccionar aquellas componentes cuyos autovalores son mayor o igual al promedio de todos los autovalores $\frac{1}{N} \sum_{i=1}^N \lambda_i$
- 3.- Representar en una gráfica los valores propios y seleccionar el número de componentes en función de un cambio brusco

Ejemplo: Calificaciones de 15 alumnos en distintas asignaturas

Se dispone de las calificaciones de 15 alumnos en 8 materias: lengua, matemáticas, física, inglés, filosofía, historia, química y gimnasia.

Componente	Varianza explicada (autovalor / suma de los 8 autovalores)	Varianza acumulada
1	0.464	0.464
2	0.358	0.821 (0.464+0.358)
3	0.119	0.941 (0.464+0.358+0.119)
4	0.027	0.968 (0.464+ ... + 0.027)
...
8	0.002	1

Ejemplo: Calificaciones de 15 alumnos en distintas asignaturas

Construcción de 8 componentes como mucho porque hay 8 variables independientes

Con 8 componentes se consigue explicar toda la información

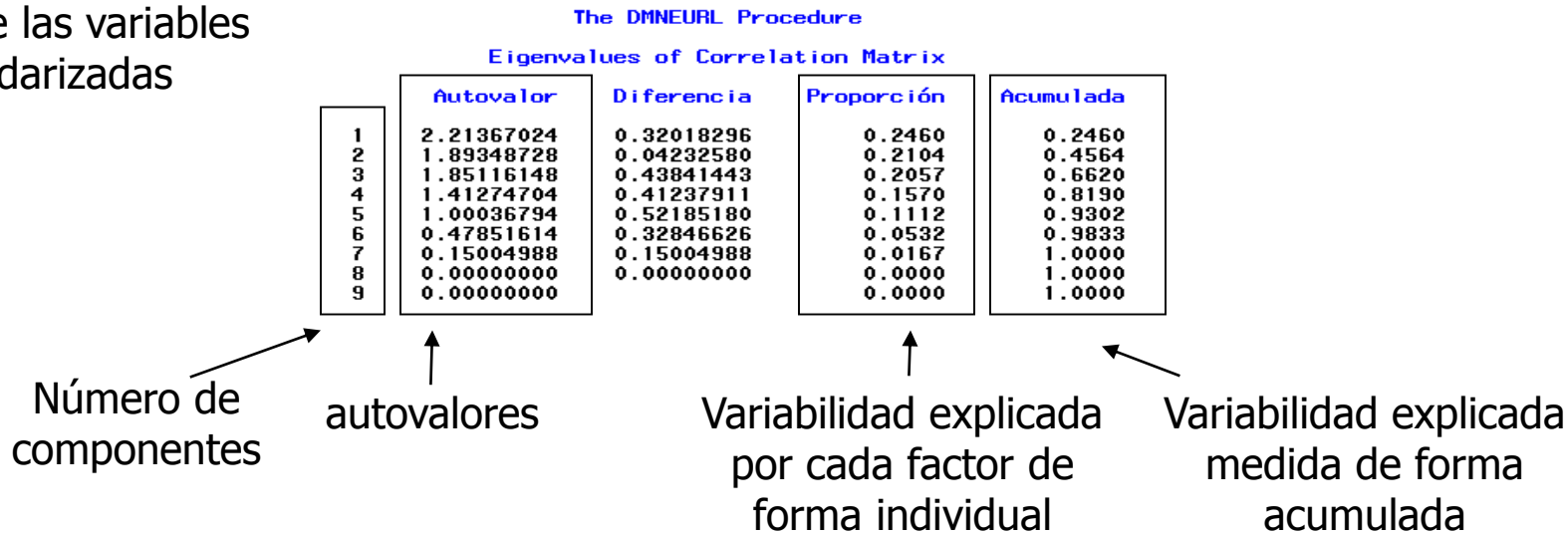
Cada componente aporta menos información que el anterior

El primer componente explica casi el 50% de la información

Con 3 componentes se consigue explicar el 94% de la información

Ejemplo: eliminación de variables redundantes en problema de impago en PYMES

PCA sobre las variables estandarizadas



¿Por qué hay 9 componentes? →

Porque hay 9 variables independientes

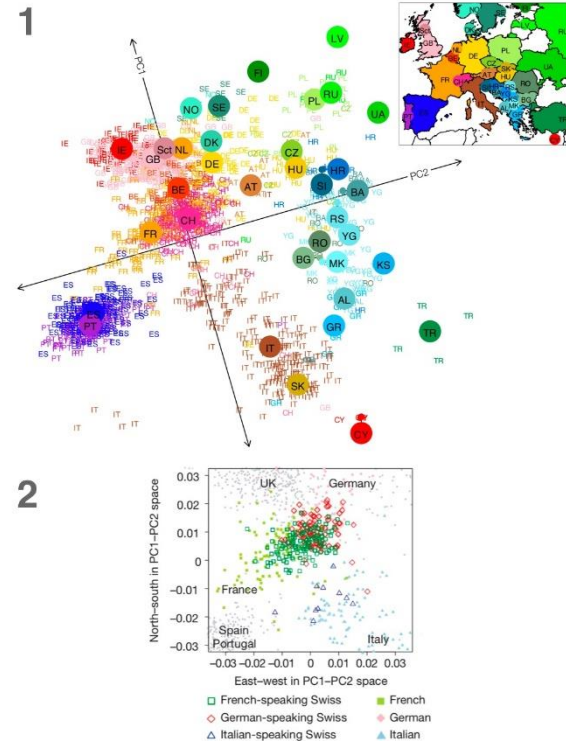
¿PCA detecta redundancia en las variables? → Sí

Ejemplo: estudio de datos genéticos

En el artículo “Genes mirror geography within Europe”, Novembre et al. 2008, Nature vol. 456, pp. 98-10, Novembre y colaboradores toman los datos genéticos de 1387 personas europeas y realizan PCA, mostrando la proyección en los dos primeros componentes principales.

Cada punto: una persona. Cada círculo: medianas en cada país.

Resultado: hay una correspondencia muy fuerte entre el mapa genético y el mapa geográfico.



courtesy: John Novembre, UCLA

Aplicaciones de PCA (1)

1. Reducir el número de variables del problema mediante extracción de características: $y = W^T x$

Nos olvidamos de las antiguas variables y trabajamos en adelante con las variables sintéticas obtenidas de los componentes principales.

Por ejemplo, entrenamos un árbol de decisión con ellas.

2. Reducir el número de variables del problema mediante selección de variables:

Usamos PCA para que nos “diga” qué variables antiguas podemos “tachar” (no aportan información relevante o son redundantes).

Aquellas características o variables cuyos coeficientes sean elevados para los primeros componentes principales, serán las que aporten la mayor capacidad discriminatoria

A partir de ese momento, trabajamos sólo con las variables antiguas relevantes.

Aplicaciones de PCA (2)

3. Detección de "outliers":

Usamos PCA para detectar qué ejemplos son atípicos. Diagramas de dispersión en los factores.

4. Descubrimiento de "clusters":

Usamos PCA para detectar grupos diferentes de ejemplos.