Aprendizaje no supervisado

VC09: Análisis de Componentes Principales

Félix José Fuentes Hurtado
felixjose.fuentes@campusviu.es

Universidad Internacional de Valencia

Cuestiones previas

- Datos originales vs. datos transformados
  - Interpretabilidad vs. utilidad
  - Datos completos vs. pérdida de información
  - Gran cantidad de datos vs. Cantidad manejable de datos

Cuestiones previas

Valor medio y valor esperado

Dada una variable aleatoria X, el **valor esperado** de X es:

$$E[X] = \sum_{x \in \mathcal{X}} x \cdot P(X = x)$$

Dada una variable aleatoria X, el **valor esperado** de X es:

$$E[X] = \sum_{x \in \mathcal{X}} x \cdot P(X = x)$$

Dada una muestra S de X, el **valor medio** de S es:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} S_i$$

Dada una variable aleatoria X, la **varianza** de X es:

$$Var(X) = E[(X - E[X])^2]$$

donde $E[X] = \sum_{x \in \mathcal{X}} x \cdot P(X = x)$

Cuestiones previas
Variance

Dada una variable aleatoria X, la **varianza** de X es:

$$Var(X) = E[(X - E[X])^2]$$

donde $E[X] = \sum_{x \in \mathcal{X}} x \cdot P(X = x)$

Dada una muestra S de X, la **varianza** de S es:

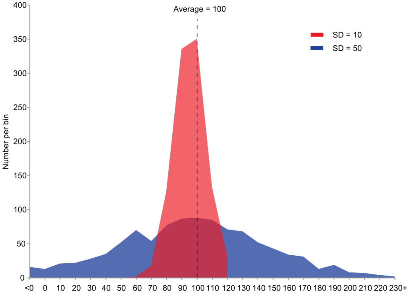
$$\sigma_n^2 = \frac{1}{n} \sum_{i=1}^{n} (S_i - \bar{X})^2$$

Ejemplos:

X : Altura de estudiantes

X : Edad

X : Horas de estudio



Justificación, necesidad de la transformación

Probablemente, el principal uso del análisis de componentes es la reducción de dimensionalidad

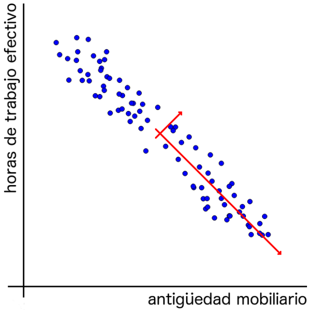
Expresar los mismos datos, con la menor pérdida de información posible, a través de un menor número de variables.

Otra info: descubrimiento de relaciones ocultas entre variables, espacio más apropiado para la aplicación de ciertas técnicas de análisis, etc.

Ejemplo

Estudio del rendimiento de trabajadores/as

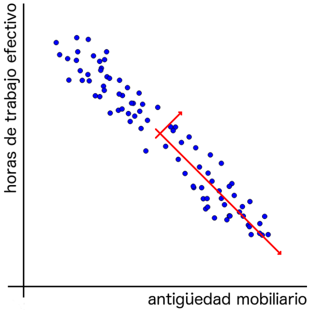
- Variables: No. horas trabajadas, Antigüedad del material, Comodidad, Movilidad en el puesto de trabajo, Rendimiento, etc.
- Si el número de horas de trabajo real está directamente relacionado con la antigüedad del material, la relación puede quedar escondida a simple vista



Ejemplo

Estudio del rendimiento de trabajadores/as

- Variables: No. horas trabajadas, Antigüedad del material, Comodidad, Movilidad en el puesto de trabajo, Rendimiento, etc.
- Si el número de horas de trabajo real está directamente relacionado con la antigüedad del material, la relación puede quedar escondida a simple vista
- Mediante análisis de componentes, se descubriría la relación entre ellas y la presencia de información redundante



Análisis de Componentes Principales (PCA)

Idea

- La idea es crear un conjunto de variables nuevo (reducido) que representen la misma información
- Serie de componentes (variables) ortogonales que explican, cada vez en menor medida, una porción de la información

  Podríamos decir: PCA obtiene representaciones comprimidas de los datos

Análisis de Componentes Principales (PCA)

Idea

- La idea es crear un conjunto de variables nuevo (reducido) que representen la misma información
- Serie de componentes (variables) ortogonales que explican, cada vez en menor medida, una porción de la información
  Podríamos decir: PCA obtiene representaciones comprimidas de los datos
- Las componentes que explican en menor medida los datos se eliminan para conseguir la reducción de dimensionalidad

Análisis de Componentes Principales (PCA)

Idea

- La idea es crear un conjunto de variables nuevo (reducido) que representen la misma información
- Serie de componentes (variables) ortogonales que explican, cada vez en menor medida, una porción de la información
  Podríamos decir: PCA obtiene representaciones comprimidas de los datos
- Las componentes que explican en menor medida los datos se eliminan para conseguir la reducción de dimensionalidad
- Efectivo contra el ruido y los valores extraños
- La representación en espacios alternativos puede ser útil para ciertos tipos de técnicas de análisis

PCA

- CPs: serie de proyecciones de los datos **mutuamente no correlacionadas**, **ordenadas** según la cantidad de varianza de los datos originales que explican
- Cada CP es el eje que mejor explica la mayor porción de varianza no explicada

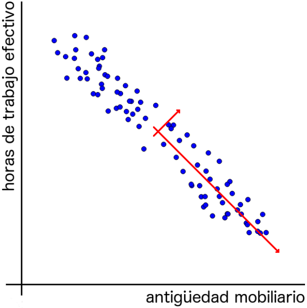
CP.1: Explica la mayor cantidad de varianza

CP.2: Ortogonal a CP.1, es el eje que explica la mayor cantidad de varianza no explicada por CP.1

CP.3: Ortogonal a CP.1 y CP.2, es el eje que explica la mayor cantidad de varianza no explicada por CP.1 ni por CP.2

...

\# CPs = # Variables originales



PCA

Paso previo: Estandarización de los datos

Objetivo

Conseguir que todas las variables originales tengan el mismo rango.

1. Centrar las variables (media $= 0$):

$$\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i - \frac{1}{n}\sum_{i'=1}^{n} \boldsymbol{x}_{i'} \ , \forall i \in \{1, \ldots, n\}$$

2. Re-escalar las variables (varianza $= 1$):

$$x_i j \leftarrow x_{ij} / \sqrt{\frac{1}{n}\sum_{i'=1}^{n} (x_{i'j})^2} \ , \forall i \in \{1, \ldots, n\} \wedge j \in \{1, \ldots, v\}$$

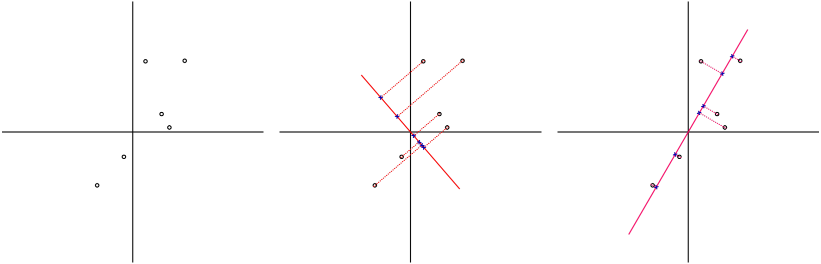
** Evitar que las variables de mayor rango dominen las de menor rango **

PCA

Objetivo

Encontrar la dirección sobre la que mejor se expresan los datos

Buscar un vector $\boldsymbol{u}$ tal que si los datos se proyectan en esa dirección, la varianza de la proyección es máxima



PCA

Varianza de una proyección

Buscar el vector $\boldsymbol{u}$ que maximiza la varianza sobre todo el conjunto de datos, $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$:

$$\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^t \boldsymbol{u})^2 = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}^t \boldsymbol{x}_i \boldsymbol{x}_i^t \boldsymbol{u} = \boldsymbol{u}^t \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i \boldsymbol{x}_i^t\right) \boldsymbol{u} = \boldsymbol{u}^t \Sigma \boldsymbol{u}$$

donde $\Sigma = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i \boldsymbol{x}_i^t$ es la matriz de covarianza.

PCA

Varianza de una proyección

Buscar el vector $\boldsymbol{u}$ que maximiza la varianza sobre todo el conjunto de datos, $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$:

$$\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^t\boldsymbol{u})^2 = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}^t\boldsymbol{x}_i\boldsymbol{x}_i^t\boldsymbol{u} = \boldsymbol{u}^t\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^t\right)\boldsymbol{u} = \boldsymbol{u}^t\Sigma\boldsymbol{u}$$

donde $\Sigma = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^t$ es la matriz de covarianza.

Tarea

El problema se define como:

$$\arg\max_{\boldsymbol{u}} \boldsymbol{u}^t\Sigma\boldsymbol{u}$$

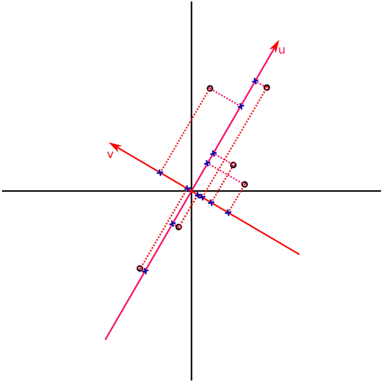
Respuesta: El vector propio principal de Σ

PCA

¡Los **vectores propios** de Σ son los vectores ortogonales que buscamos!

Procedimiento

1. Calcular la matriz de covarianzas, Σ
2. Descomponer Σ en vectores propios
   (descomposición en valores singulares)
3. Seleccionar los q vectores propios principales como CPs
   (los q vectores propios con mayor valor propio asociado)



PCA

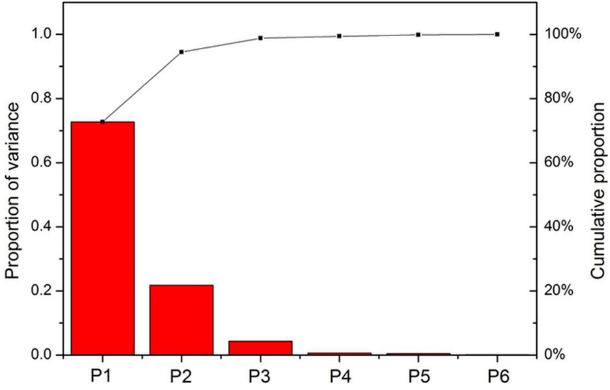
- Cada componente principal $\boldsymbol{u}_j$ es una combinación lineal de las variables originales
- El nuevo conjunto de datos Z en el espacio transformado es:

$$z_i = \begin{bmatrix} \boldsymbol{u}_1^t \boldsymbol{x}_i \\ \boldsymbol{u}_2^t \boldsymbol{x}_i \\ \vdots \\ \boldsymbol{u}_q^t \boldsymbol{x}_i \end{bmatrix}, \quad \forall i \in \{1, \dots, n\}$$

- La reducción de la dimensionalidad depende del número de componentes principales (q).

La reducción de la dimensionalidad depende de q
(número de componentes principales)

- Si $q = v$, no hay pérdida de información ni reducción de dimensionalidad
- A menor q, mayor reducción de dimensionalidad y pérdida de información
- Ritmo de pérdida de información: depende de la redundancia de las variables y las relaciones ocultas en los datos



1. Fijar un umbral s (ej., 95) en el acumulado de varianza explicada
2. Seleccionar las q CPs que expliquen al menos el s % de la varianza total de los datos

Ventajas

- Técnica no paramétrica
- Único parámetro ajustable (posterior): número de componentes q
- En el espacio de optimización no existen máximos locales donde el método pudiese quedar atrapado

Desventajas

- El nuevo espacio puede no ser intuitivo
- La interpretabilidad de las variables se pierde (oscurece)
- Se limita a momentos muestrales de orden 2 (varianza) y a proyecciones lineales

Aprendizaje no supervisado

VC09: Análisis de Componentes Principales

Félix José Fuentes Hurtado

felixjose.fuentes@campusviu.es

Universidad Internacional de Valencia