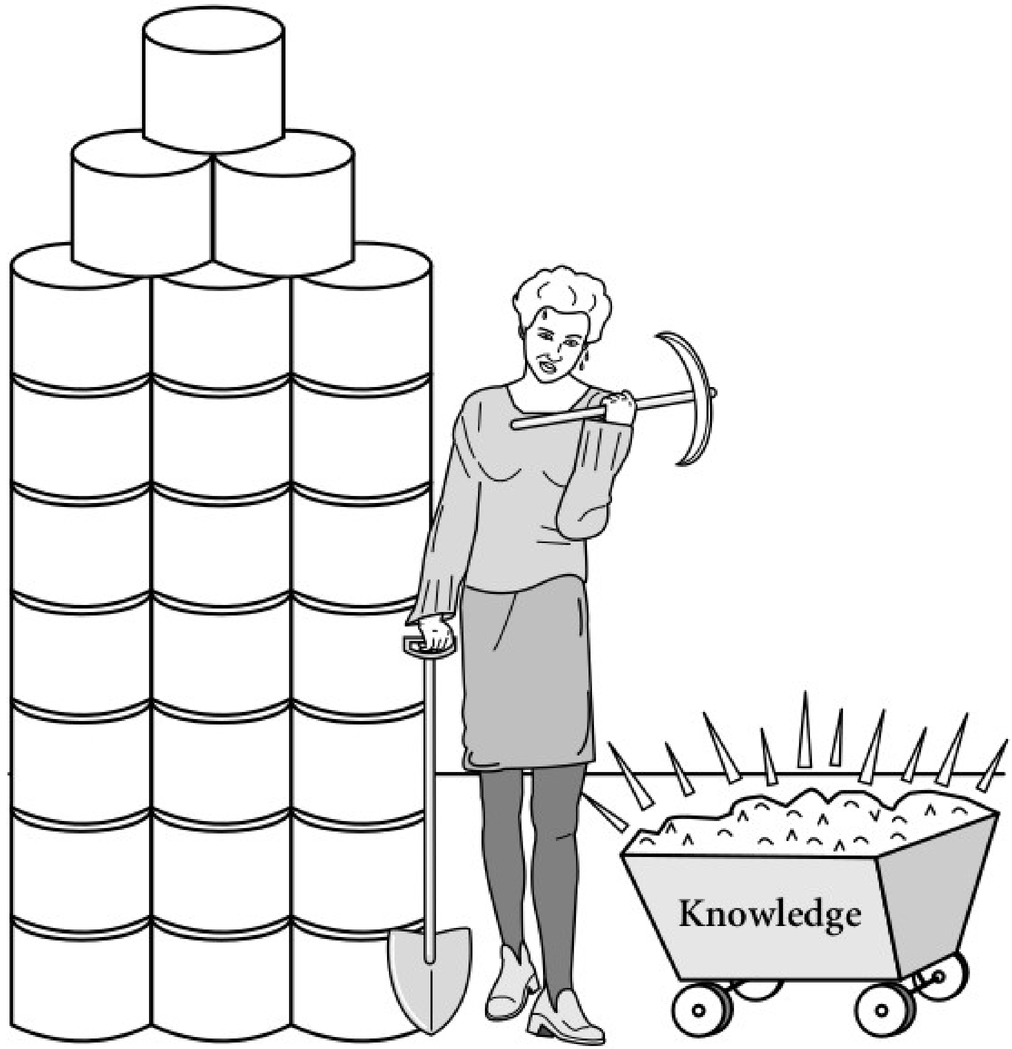
PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

Minería de Datos

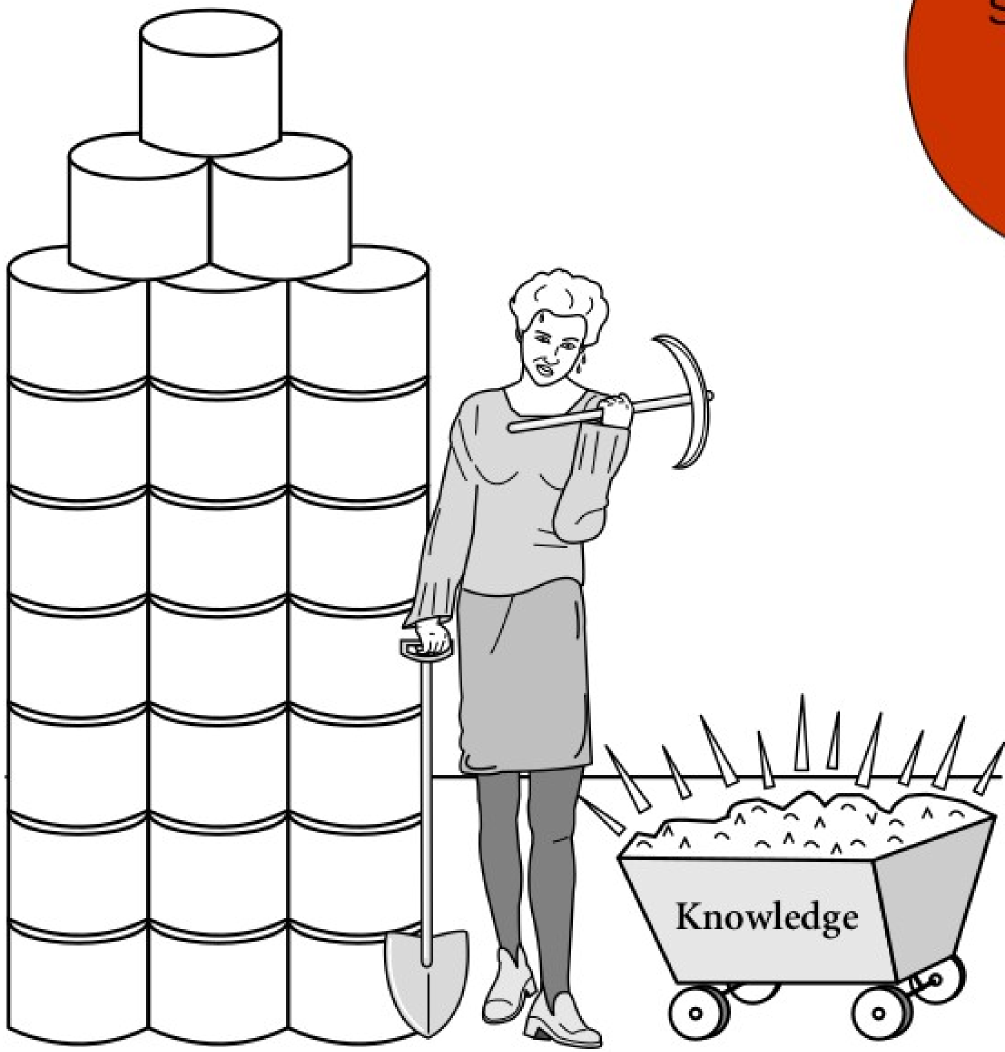
¿Qué vamos a ver?

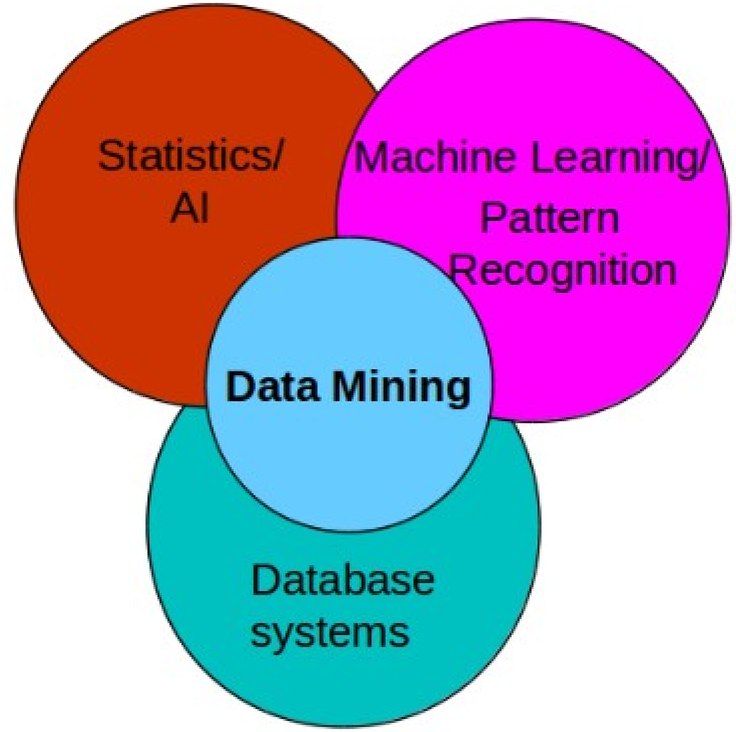
- Representación (PCA, t-SNE, ...)
- Asociación (apriori)
- Clustering (k-means, HAC)
- Clasificación
- Auto-encoders

Data Mining

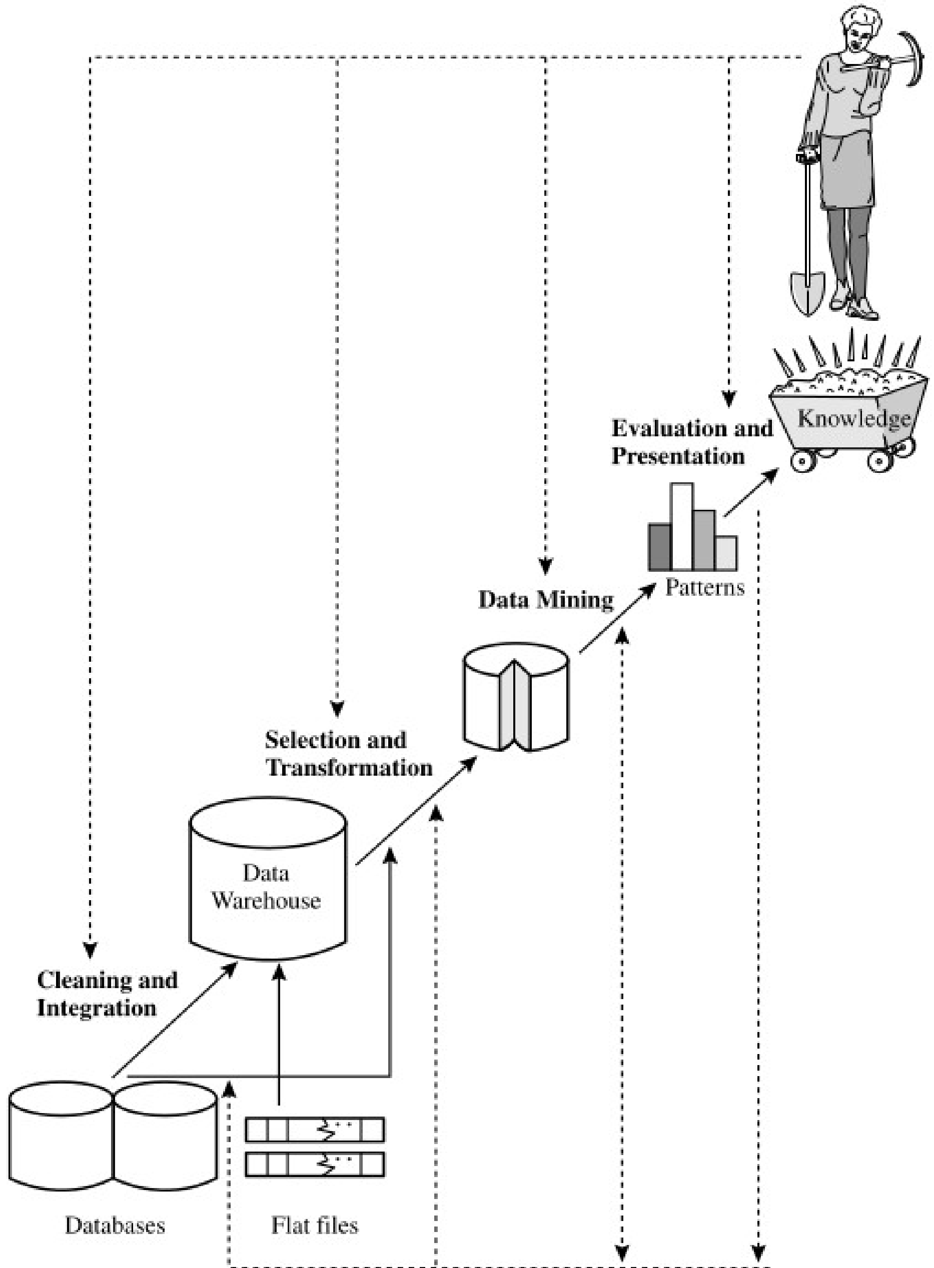


Data Mining

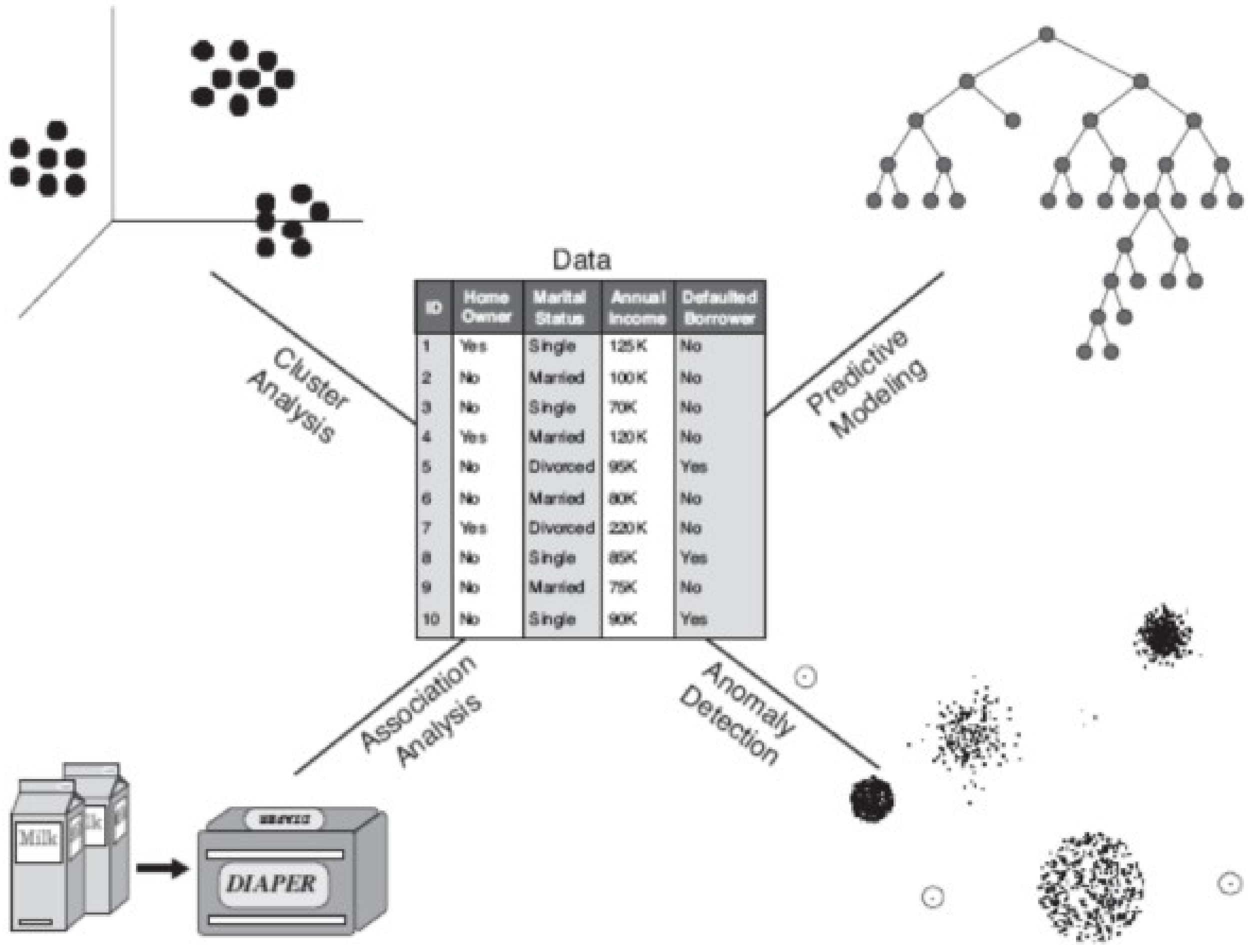

Knowledge


Statistics/
AI
Machine Learning/
Pattern
Recognition
Data Mining
Database
systems

Data Mining

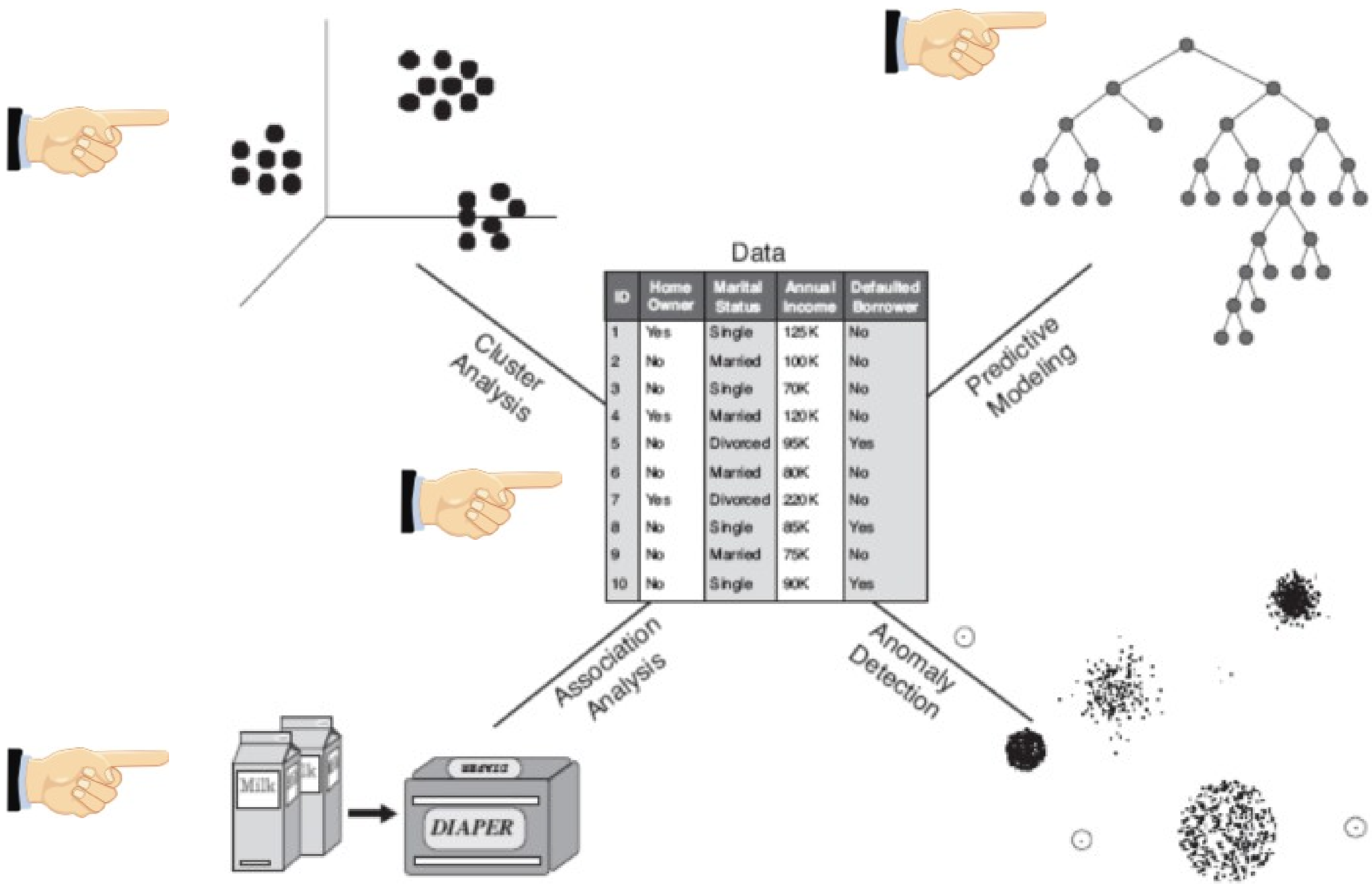

Knowledge
Evaluation and Presentation
Patterns
Data Mining
Selection and Transformation
Data Warehouse
Cleaning and Integration
Databases
Flat files

Data Mining



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Data Mining



Data
Cluster Analysis
Predictive Modeling
Association Analysis
Anomaly Detection
Milk
DIAPER

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Fuentes de Datos

Tid	Refund	Marital Status	Taxable Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

(a) Record data.

TID	*ITEMS*
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

(b) Transaction data.

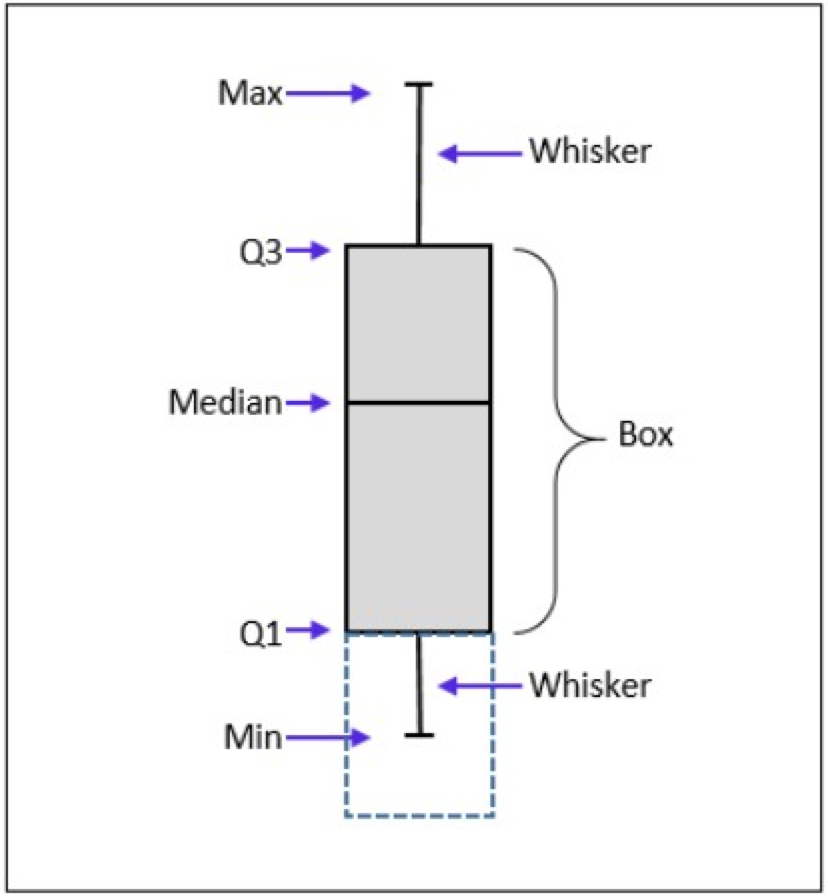
Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

(c) Data matrix.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

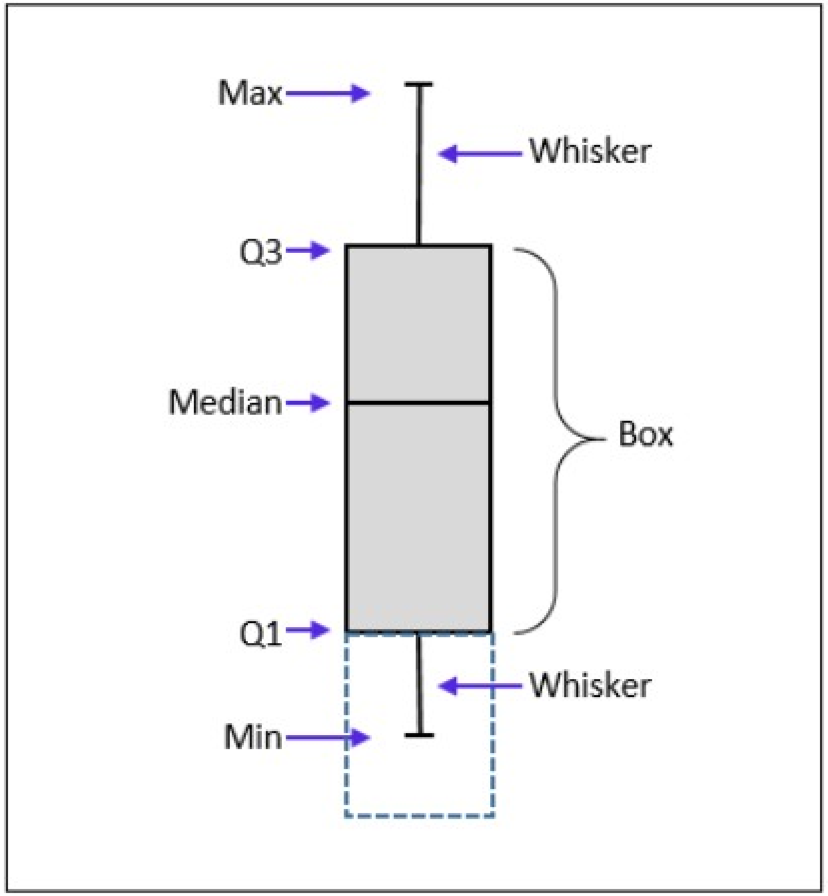
(d) Document-term matrix.

Herramientas de análisis exploratorio

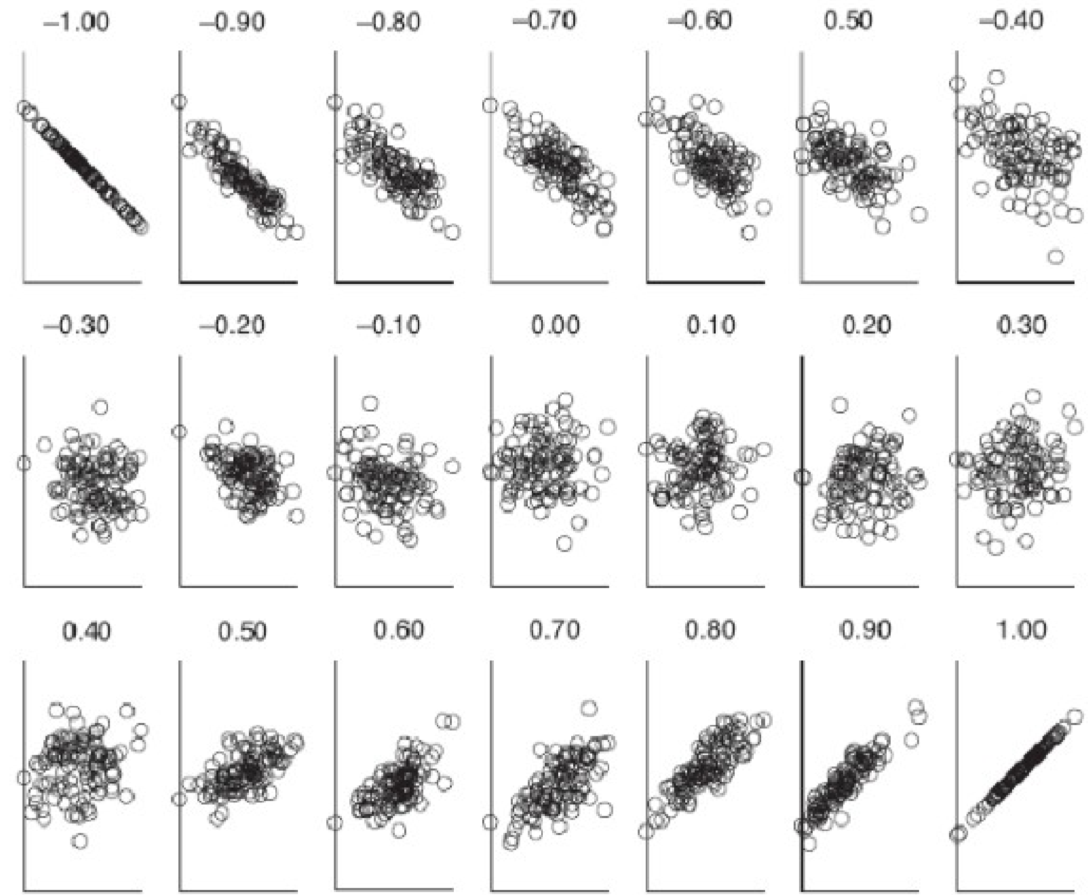


Box Plots

Herramientas de análisis exploratorio



Box Plots

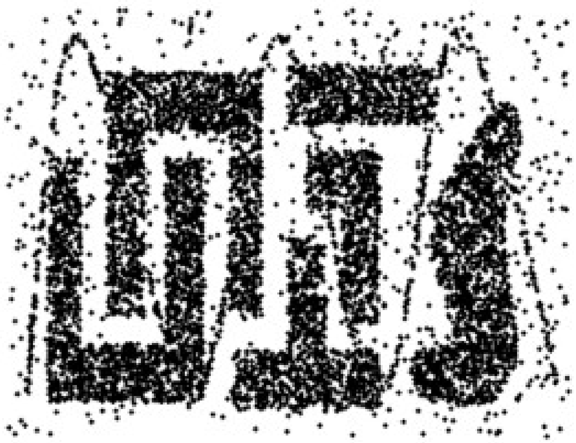


Scatter Plots

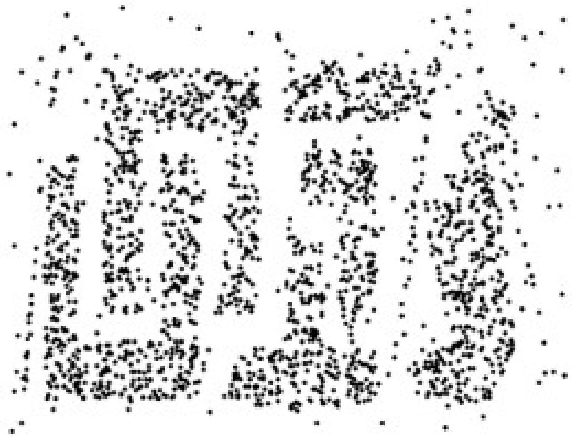
Captura de Datos

- Muestreo aleatorio: equiprobable en la selección de datos.
- Muestreo con reemplazo: los datos no son removidos de la colección (puede ser muestreado más de una vez).
- Muestreo sin reemplazo: los datos son removidos de la colección si son muestreados (puede ser muestreado solo una vez).
- Muestreo estratificado: los datos son segmentados de acuerdo a algún criterio. Luego, en cada segmento se conduce muestreo aleatorio.

Captura de Datos



8000 points



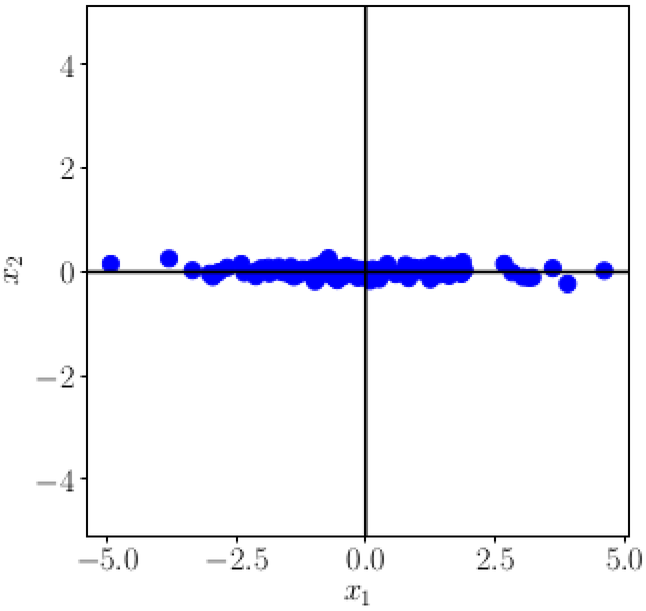
2000 Points

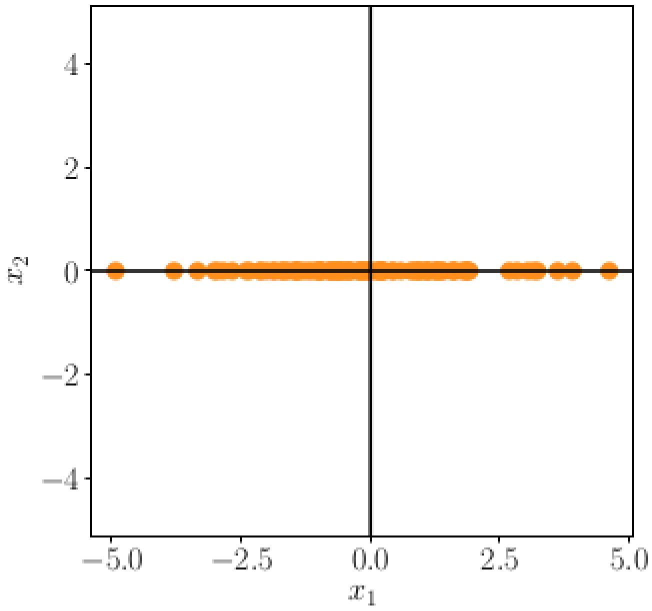
500 Points

Riesgo empírico en muestreo

Representaciones y reducción de dimensionalidad

Análisis de Componentes Principales (PCA)





X1 retiene la mayor parte de la varianza por lo que remover x2 es neutro en términos de compresión.

Análisis de Componentes Principales (PCA)

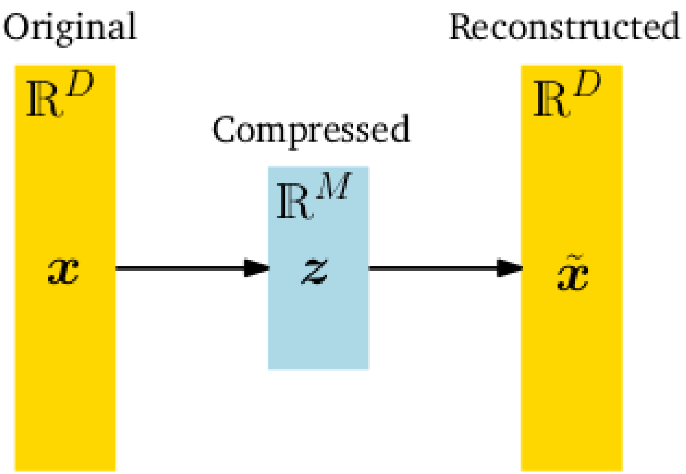
dataset $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, $\boldsymbol{x}_n \in \mathbb{R}^D$

Análisis de Componentes Principales (PCA)

dataset $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, $\boldsymbol{x}_n \in \mathbb{R}^D$

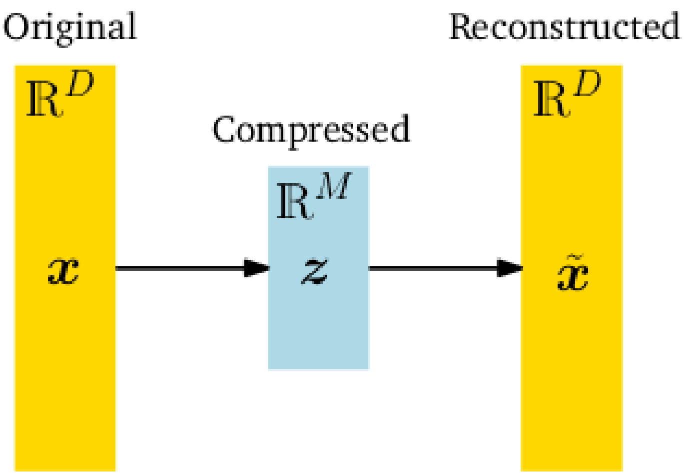
Análisis de Componentes Principales (PCA)

dataset $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}, \boldsymbol{x}_n \in \mathbb{R}^D$



Análisis de Componentes Principales (PCA)

dataset $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}, \boldsymbol{x}_n \in \mathbb{R}^D$

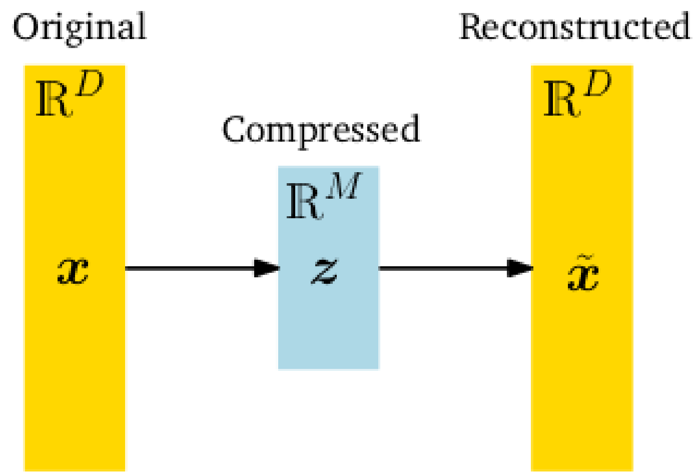


$\boldsymbol{z}_n = \boldsymbol{B}^\top \boldsymbol{x}_n \in \mathbb{R}^M$ → Baja dimensionalidad

→ Base de la descomposición $\boldsymbol{B} := [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M] \in \mathbb{R}^{D \times M}$.

Análisis de Componentes Principales (PCA)

dataset $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, $\boldsymbol{x}_n \in \mathbb{R}^D$



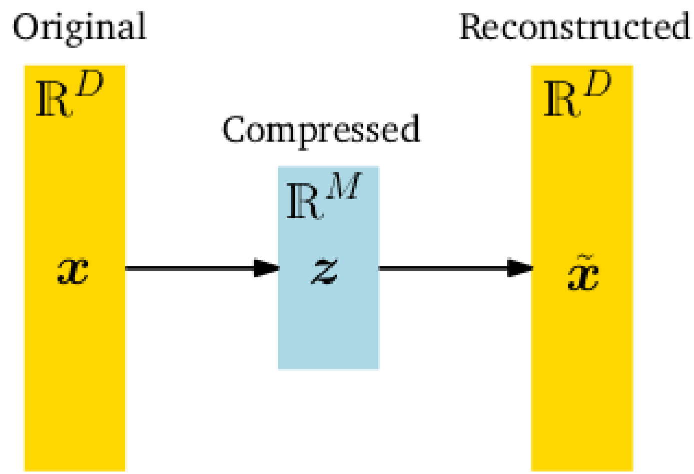
$\boldsymbol{z}_n = \boldsymbol{B}^\top \boldsymbol{x}_n \in \mathbb{R}^M$ → Baja dimensionalidad

→ Base de la descomposición $\boldsymbol{B} := [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M] \in \mathbb{R}^{D \times M}$.

→ $\boldsymbol{b}_i^\top \boldsymbol{b}_j = 0$ y $\boldsymbol{b}_i^\top \boldsymbol{b}_i = 1$

Análisis de Componentes Principales (PCA)

dataset $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, $\boldsymbol{x}_n \in \mathbb{R}^D$



$\boldsymbol{z}_n = \boldsymbol{B}^\top \boldsymbol{x}_n \in \mathbb{R}^M$ → Baja dimensionalidad

→ Base de la descomposición $\boldsymbol{B} := [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M] \in \mathbb{R}^{D \times M}$.

→ $\boldsymbol{b}_i^\top \boldsymbol{b}_j = 0$ y $\boldsymbol{b}_i^\top \boldsymbol{b}_i = 1$

Minimizar la pérdida de información implica capturar la mayor cantidad de varianza en la descomposición.

Análisis de Componentes Principales (PCA)

Importante: la varianza de la descomposición es independiente de la media.

$$\mathbb{V}_{\boldsymbol{z}}[\boldsymbol{z}] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}^\top(\boldsymbol{x} - \boldsymbol{\mu})] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}^\top\boldsymbol{x} - \boldsymbol{B}^\top\boldsymbol{\mu}] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}^\top\boldsymbol{x}]$$

Análisis de Componentes Principales (PCA)

Importante: la varianza de la descomposición es independiente de la media.

$$\mathbb{V}_{\boldsymbol{z}}[\boldsymbol{z}] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}^\top(\boldsymbol{x} - \boldsymbol{\mu})] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}^\top\boldsymbol{x} - \boldsymbol{B}^\top\boldsymbol{\mu}] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}^\top\boldsymbol{x}]$$

Por lo tanto, corregimos los datos a media 0.

Análisis de Componentes Principales (PCA)

Importante: la varianza de la descomposición es independiente de la media.

$$\mathbb{V}_{\boldsymbol{z}}[\boldsymbol{z}] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}^\top(\boldsymbol{x}-\boldsymbol{\mu})] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}^\top\boldsymbol{x} - \boldsymbol{B}^\top\boldsymbol{\mu}] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}^\top\boldsymbol{x}]$$

Por lo tanto, corregimos los datos a media 0.

PCA (enfoque secuencial): Comenzamos buscando $\boldsymbol{b}_1 \in \mathbb{R}^D$:

Análisis de Componentes Principales (PCA)

Importante: la varianza de la descomposición es independiente de la media.

$$\mathbb{V}_{\boldsymbol{z}}[\boldsymbol{z}] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}^\top(\boldsymbol{x} - \boldsymbol{\mu})] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}^\top \boldsymbol{x} - \boldsymbol{B}^\top \boldsymbol{\mu}] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}^\top \boldsymbol{x}]$$

Por lo tanto, corregimos los datos a media 0.

PCA (enfoque secuencial): Comenzamos buscando $\boldsymbol{b}_1 \in \mathbb{R}^D$:

$$V_1 := \mathbb{V}[z_1] = \frac{1}{N} \sum_{n=1}^{N} z_{1n}^2$$

Análisis de Componentes Principales (PCA)

Importante: la varianza de la descomposición es independiente de la media.

$$\mathbb{V}_{\boldsymbol{z}}[\boldsymbol{z}] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}^\top(\boldsymbol{x} - \boldsymbol{\mu})] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}^\top\boldsymbol{x} - \boldsymbol{B}^\top\boldsymbol{\mu}] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}^\top\boldsymbol{x}]$$

Por lo tanto, corregimos los datos a media 0.

PCA (enfoque secuencial): Comenzamos buscando $\boldsymbol{b}_1 \in \mathbb{R}^D$:

$$V_1 := \mathbb{V}[z_1] = \frac{1}{N}\sum_{n=1}^{N} z_{1n}^2$$. Dado que $z_{1n} = \boldsymbol{b}_1^\top \boldsymbol{x}_n$,

$$V_1 = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{b}_1^\top \boldsymbol{x}_n)^2 = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{b}_1^\top \boldsymbol{x}_n \boldsymbol{x}_n^\top \boldsymbol{b}_1$$

$$= \boldsymbol{b}_1^\top \left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n \boldsymbol{x}_n^\top\right)\boldsymbol{b}_1 = \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1 \, ,$$

Análisis de Componentes Principales (PCA)

Importante: la varianza de la descomposición es independiente de la media.

$$\mathbb{V}_{\boldsymbol{z}}[\boldsymbol{z}] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}^\top(\boldsymbol{x}-\boldsymbol{\mu})] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}^\top\boldsymbol{x} - \boldsymbol{B}^\top\boldsymbol{\mu}] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}^\top\boldsymbol{x}]$$

Por lo tanto, corregimos los datos a media 0.

PCA (enfoque secuencial): Comenzamos buscando $\boldsymbol{b}_1 \in \mathbb{R}^D$:

$V_1 := \mathbb{V}[z_1] = \frac{1}{N}\sum_{n=1}^{N} z_{1n}^2$. Dado que $z_{1n} = \boldsymbol{b}_1^\top \boldsymbol{x}_n$,

$$\begin{aligned} V_1 &= \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{b}_1^\top\boldsymbol{x}_n)^2 = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{b}_1^\top\boldsymbol{x}_n\boldsymbol{x}_n^\top\boldsymbol{b}_1 \\ &= \boldsymbol{b}_1^\top\left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top\right)\boldsymbol{b}_1 = \boldsymbol{b}_1^\top\boldsymbol{S}\boldsymbol{b}_1 \,, \end{aligned}$$

Por lo tanto, maximizar la varianza corresponde a:

$$\max_{\boldsymbol{b}_1} \boldsymbol{b}_1^\top\boldsymbol{S}\boldsymbol{b}_1 \quad \text{, sujeto a } \|\boldsymbol{b}_1\|^2 = 1 \,.$$

Análisis de Componentes Principales (PCA)

$$\max_{\boldsymbol{b}_1} \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1$$, sujeto a $\|\boldsymbol{b}_1\|^2 = 1$.

Lagrangiano

$$\mathfrak{L}(\boldsymbol{b}_1, \lambda) = \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1 + \lambda_1 (1 - \boldsymbol{b}_1^\top \boldsymbol{b}_1)$$

Análisis de Componentes Principales (PCA)

$$\max_{\boldsymbol{b}_1} \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1$$, sujeto a $\|\boldsymbol{b}_1\|^2 = 1$.

Lagrangiano

$$\mathfrak{L}(\boldsymbol{b}_1, \lambda) = \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1 + \lambda_1 (1 - \boldsymbol{b}_1^\top \boldsymbol{b}_1)$$

$$\frac{\partial \mathfrak{L}}{\partial \boldsymbol{b}_1} = 2\boldsymbol{b}_1^\top \boldsymbol{S} - 2\lambda_1 \boldsymbol{b}_1^\top , \qquad \frac{\partial \mathfrak{L}}{\partial \lambda_1} = 1 - \boldsymbol{b}_1^\top \boldsymbol{b}_1$$

Análisis de Componentes Principales (PCA)

$$\max_{\boldsymbol{b}_1} \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1$$, sujeto a $\|\boldsymbol{b}_1\|^2 = 1\,.$

Lagrangiano

$$\mathfrak{L}(\boldsymbol{b}_1, \lambda) = \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1 + \lambda_1 (1 - \boldsymbol{b}_1^\top \boldsymbol{b}_1)$$

$$\frac{\partial \mathfrak{L}}{\partial \boldsymbol{b}_1} = 2\boldsymbol{b}_1^\top \boldsymbol{S} - 2\lambda_1 \boldsymbol{b}_1^\top \,, \qquad \frac{\partial \mathfrak{L}}{\partial \lambda_1} = 1 - \boldsymbol{b}_1^\top \boldsymbol{b}_1$$

= 0

$$\boldsymbol{S}\boldsymbol{b}_1 = \lambda_1 \boldsymbol{b}_1 \,,$$
$$\boldsymbol{b}_1^\top \boldsymbol{b}_1 = 1\,.$$

Análisis de Componentes Principales (PCA)

$$\max_{\boldsymbol{b}_1} \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1$$, sujeto a $\|\boldsymbol{b}_1\|^2 = 1$.

Lagrangiano

$$\mathfrak{L}(\boldsymbol{b}_1, \lambda) = \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1 + \lambda_1 (1 - \boldsymbol{b}_1^\top \boldsymbol{b}_1)$$

$$\frac{\partial \mathfrak{L}}{\partial \boldsymbol{b}_1} = 2\boldsymbol{b}_1^\top \boldsymbol{S} - 2\lambda_1 \boldsymbol{b}_1^\top , \qquad \frac{\partial \mathfrak{L}}{\partial \lambda_1} = 1 - \boldsymbol{b}_1^\top \boldsymbol{b}_1$$

= 0

$$\boldsymbol{S}\boldsymbol{b}_1 = \lambda_1 \boldsymbol{b}_1 ,$$
$$\boldsymbol{b}_1^\top \boldsymbol{b}_1 = 1 .$$

Reescribimos: $V_1 = \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1 = \lambda_1 \boldsymbol{b}_1^\top \boldsymbol{b}_1 = \lambda_1$

Análisis de Componentes Principales (PCA)

$$\max_{\boldsymbol{b}_1} \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1$$, sujeto a $\|\boldsymbol{b}_1\|^2 = 1$.

Lagrangiano

$$\mathfrak{L}(\boldsymbol{b}_1, \lambda) = \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1 + \lambda_1 (1 - \boldsymbol{b}_1^\top \boldsymbol{b}_1)$$

$$\frac{\partial \mathfrak{L}}{\partial \boldsymbol{b}_1} = 2\boldsymbol{b}_1^\top \boldsymbol{S} - 2\lambda_1 \boldsymbol{b}_1^\top , \qquad \frac{\partial \mathfrak{L}}{\partial \lambda_1} = 1 - \boldsymbol{b}_1^\top \boldsymbol{b}_1$$

= 0

$$\boldsymbol{S}\boldsymbol{b}_1 = \lambda_1 \boldsymbol{b}_1 ,$$
$$\boldsymbol{b}_1^\top \boldsymbol{b}_1 = 1 .$$

Se calcula usando la SVD

Reescribimos: $V_1 = \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1 = \lambda_1 \boldsymbol{b}_1^\top \boldsymbol{b}_1 = \lambda_1$

Análisis de Componentes Principales (PCA)

$$\max_{\boldsymbol{b}_1} \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1$$, sujeto a $\|\boldsymbol{b}_1\|^2 = 1$.

Lagrangiano

$$\mathfrak{L}(\boldsymbol{b}_1, \lambda) = \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1 + \lambda_1 (1 - \boldsymbol{b}_1^\top \boldsymbol{b}_1)$$

$$\frac{\partial \mathfrak{L}}{\partial \boldsymbol{b}_1} = 2\boldsymbol{b}_1^\top \boldsymbol{S} - 2\lambda_1 \boldsymbol{b}_1^\top , \qquad \frac{\partial \mathfrak{L}}{\partial \lambda_1} = 1 - \boldsymbol{b}_1^\top \boldsymbol{b}_1$$

= 0

$$\boldsymbol{S}\boldsymbol{b}_1 = \lambda_1 \boldsymbol{b}_1 ,$$
$$\boldsymbol{b}_1^\top \boldsymbol{b}_1 = 1 .$$

Se calcula usando la SVD

Reescribimos: $V_1 = \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1 = \lambda_1 \boldsymbol{b}_1^\top \boldsymbol{b}_1 = \lambda_1$

La reconstrucción es: $\tilde{\boldsymbol{x}}_n = \boldsymbol{b}_1 z_{1n} = \boldsymbol{b}_1 \boldsymbol{b}_1^\top \boldsymbol{x}_n \in \mathbb{R}^D$

Análisis de Componentes Principales (PCA)

$$\max_{\boldsymbol{b}_1} \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1$$, sujeto a $\|\boldsymbol{b}_1\|^2 = 1$.

Lagrangiano

$$\mathfrak{L}(\boldsymbol{b}_1, \lambda) = \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1 + \lambda_1 (1 - \boldsymbol{b}_1^\top \boldsymbol{b}_1)$$

$$\frac{\partial \mathfrak{L}}{\partial \boldsymbol{b}_1} = 2\boldsymbol{b}_1^\top \boldsymbol{S} - 2\lambda_1 \boldsymbol{b}_1^\top , \qquad \frac{\partial \mathfrak{L}}{\partial \lambda_1} = 1 - \boldsymbol{b}_1^\top \boldsymbol{b}_1$$

= 0

$$\boldsymbol{S}\boldsymbol{b}_1 = \lambda_1 \boldsymbol{b}_1 ,$$
$$\boldsymbol{b}_1^\top \boldsymbol{b}_1 = 1 .$$

Se calcula usando la SVD

Reescribimos: $V_1 = \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1 = \lambda_1 \boldsymbol{b}_1^\top \boldsymbol{b}_1 = \lambda_1$

La reconstrucción es: $\tilde{\boldsymbol{x}}_n = \boldsymbol{b}_1 z_{1n} = \boldsymbol{b}_1 \boldsymbol{b}_1^\top \boldsymbol{x}_n \in \mathbb{R}^D$

Análisis de Componentes Principales (PCA)

Proceso iterativo:

$$\hat{\boldsymbol{X}} := \boldsymbol{X} - \sum_{i=1}^{m-1} \boldsymbol{b}_i \boldsymbol{b}_i^\top \boldsymbol{X} = \boldsymbol{X} - \boldsymbol{B}_{m-1} \boldsymbol{X} \,,$$

con $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{D \times N}$

y $\boldsymbol{B}_{m-1} := \sum_{i=1}^{m-1} \boldsymbol{b}_i \boldsymbol{b}_i^\top$

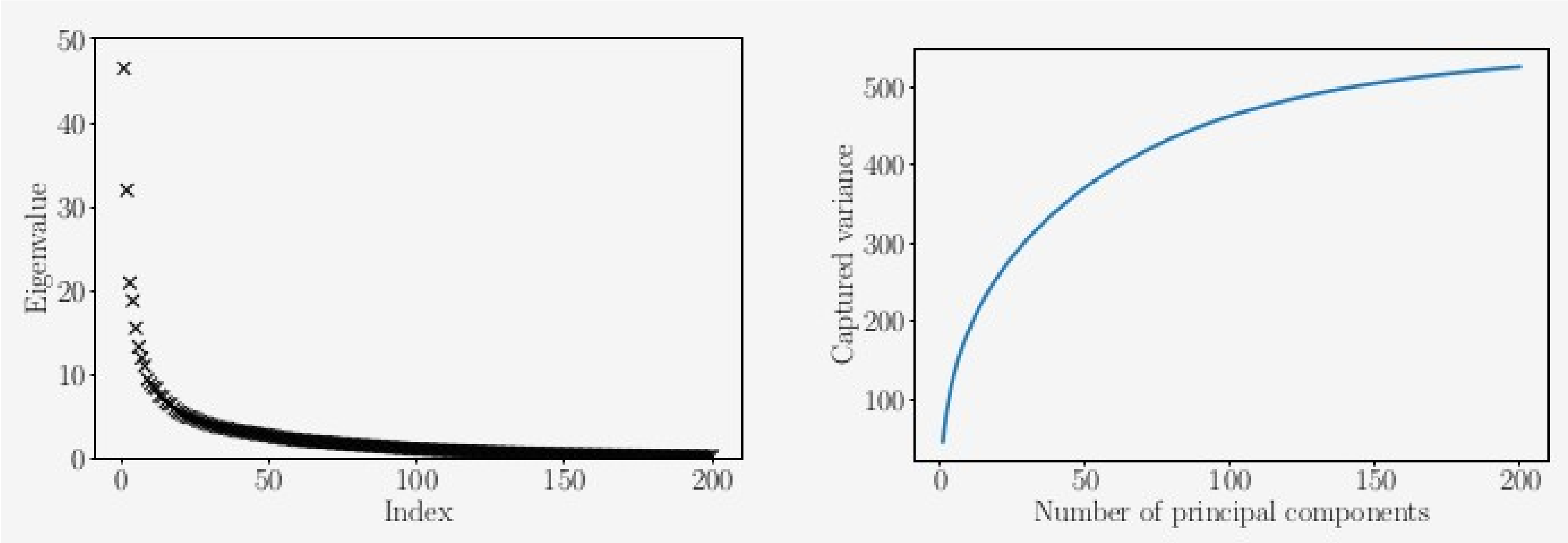
Análisis de Componentes Principales (PCA)

Proceso iterativo:

$$\hat{\boldsymbol{X}} := \boldsymbol{X} - \sum_{i=1}^{m-1} \boldsymbol{b}_i \boldsymbol{b}_i^\top \boldsymbol{X} = \boldsymbol{X} - \boldsymbol{B}_{m-1} \boldsymbol{X}\,,$$

con $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{D \times N}$

y $\boldsymbol{B}_{m-1} := \sum_{i=1}^{m-1} \boldsymbol{b}_i \boldsymbol{b}_i^\top$



Análisis de Componentes Principales (PCA)

- Aspectos prácticos:

    - Usa la full SVD (LAPACK) para datos densos.
    - Usa la SVD truncada (ARPACK) para datos dispersos.
    - Se puede usar MLE (reconstrucción – data) para estimar el # de componentes.

- Implementaciones:

    - Python: sklearn

https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html