Selección de atributos

Richard Weber

Francisco Cisternas

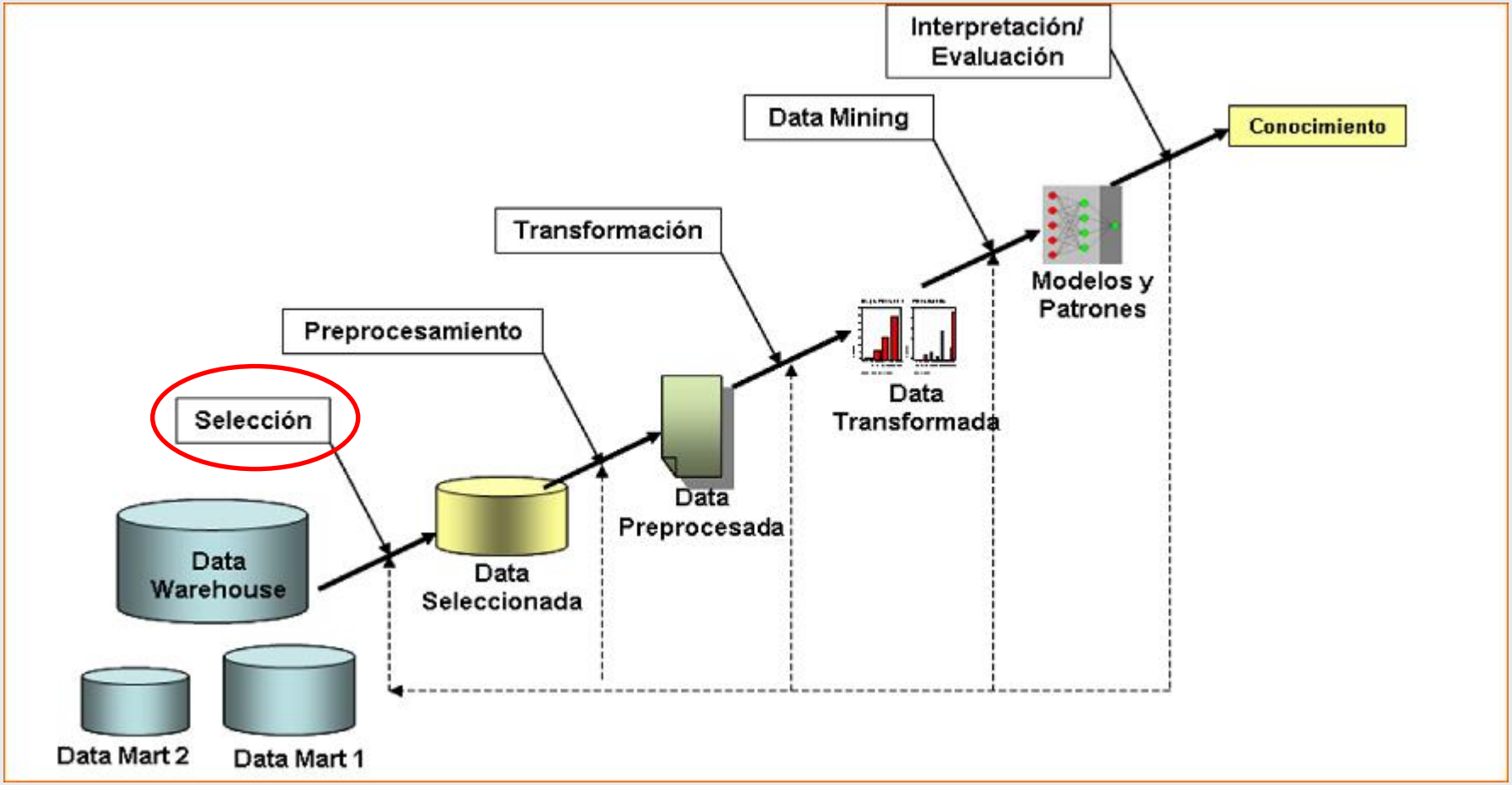
(frcister@ing.uchile.cl)

Departamento de Ingeniería Industrial

Universidad de Chile

PROCESO DE KDD

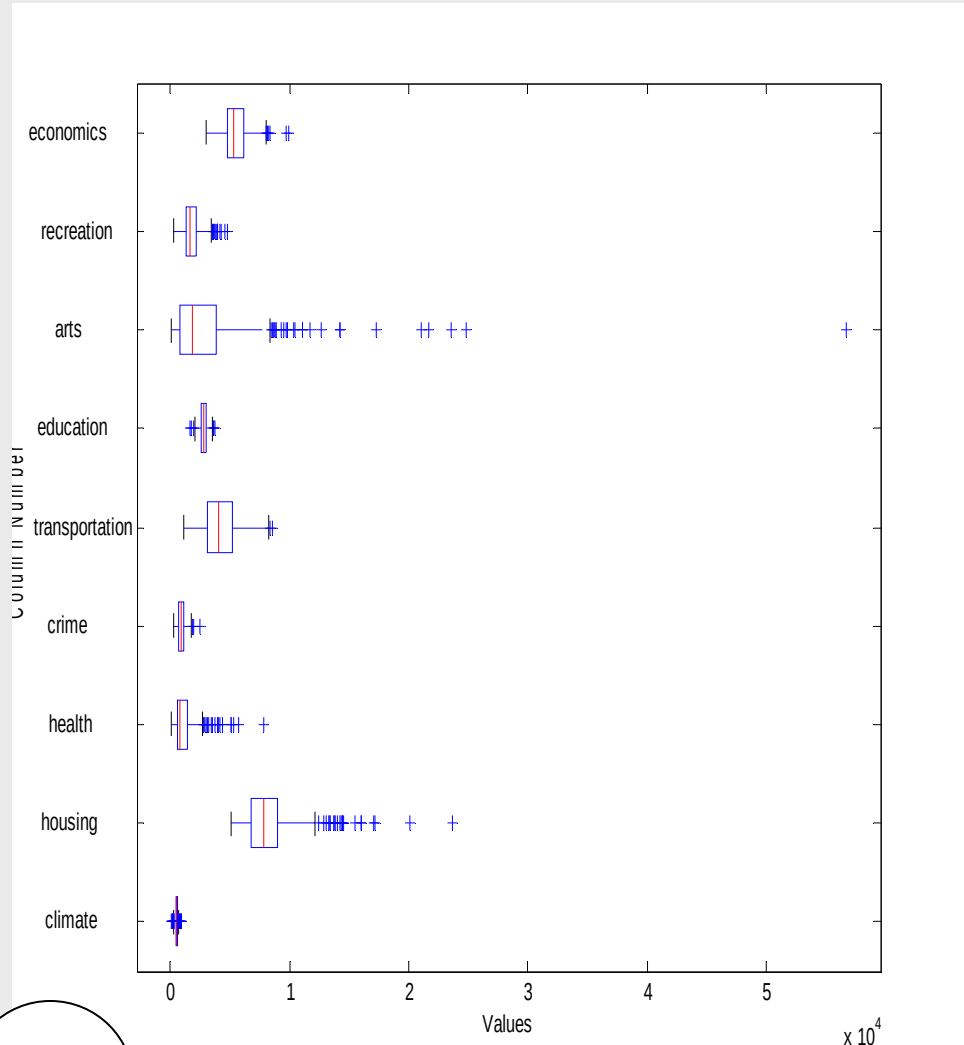
KNOWLEDGE DISCOVERY IN DATABASES



"KDD es el proceso no-trivial de identificar patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles dentro de los datos"

Motivación (1/3)

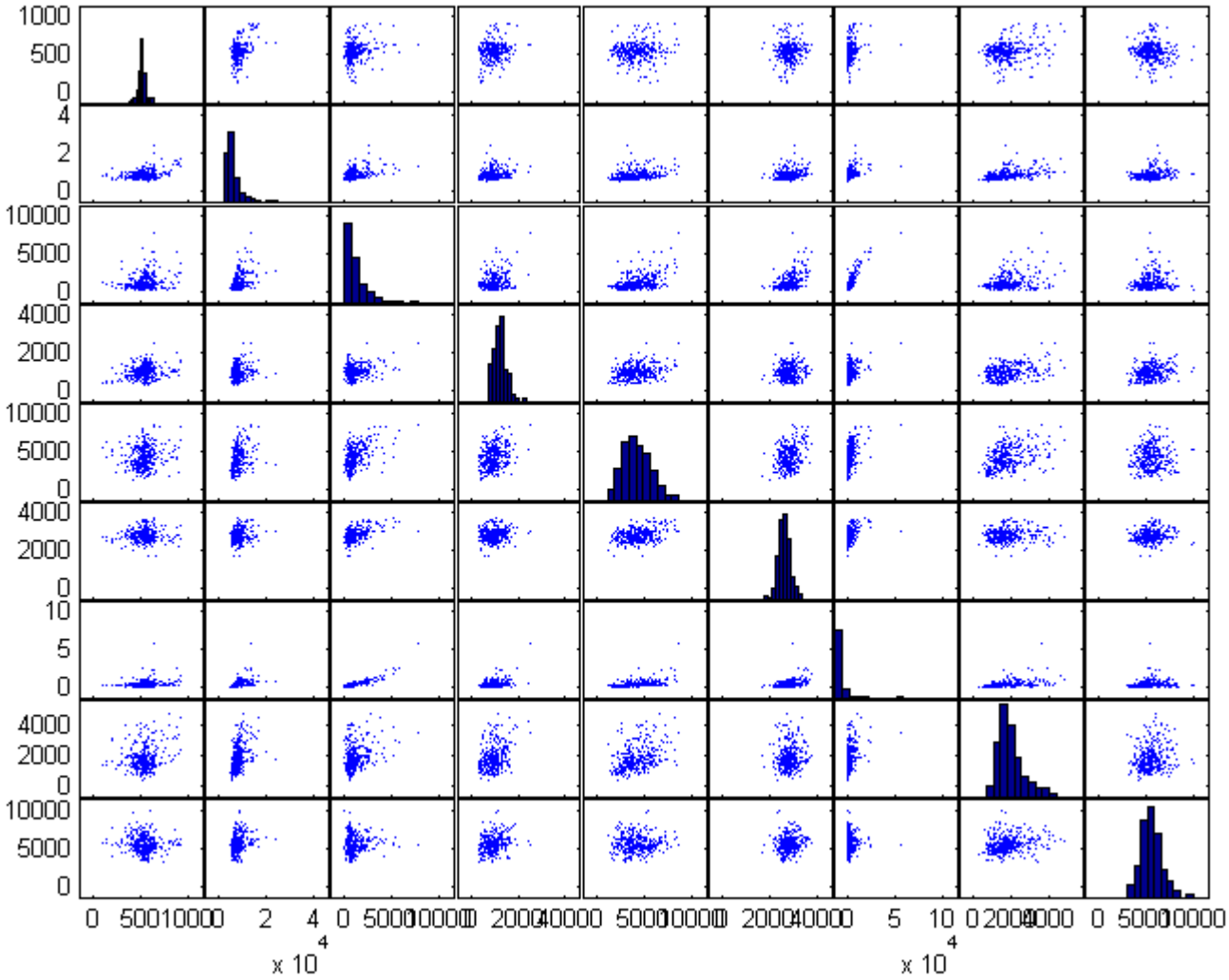
- Problema: 9 índices de calidad de vida en 329 ciudades de USA (Ejemplo de MatLab).
  - Índices: climate, housing, health, crime, transportation, education, arts, recreation, and economics. Siempre más es mejor, alto crimen ➔ baja tasa de criminalidad.
- ¿Que hacemos?, hagamos una exploración simple.
  - 'Boxplot'



Motivación (2/3)

- Veamos las relaciones

Tal vez se puede mirar,
¿pero que pasa con 20 variables?

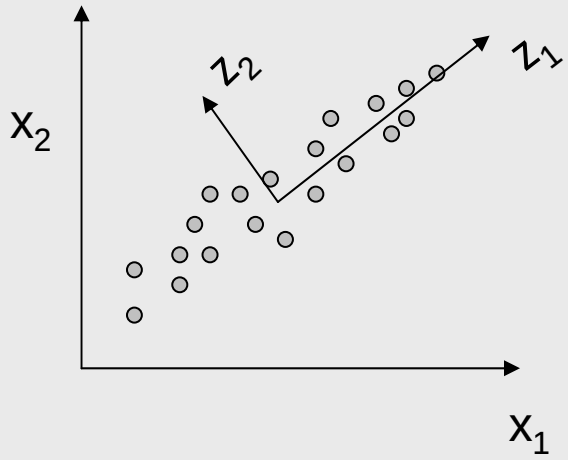


Motivación (3/3)

- Que nos gustaría:
  - Reducir el Espacio de Atributos con la menor pérdida de información posible.
- La Solución

Idea Básica

- Rotación de ejes para maximizar varianza.
- Idea de Fondo: tal vez sólo me baste con z_1.



- Planteamiento del Problema:
  - Hay *p*-variables con valores medidos.
  - *n* elementos de que son las medidas.
  - X matriz de *n*×*p*
  - Supuesto: X es centrado en la media (cada variable tiene restada su media)

Siguiendo la idea

- Posible Solución:
  - Sea a_1 el vector de pesos de la proyección de dimensión de p×**1** (desconocido por ahora).
  - Entonces podemos escribir la primera componente buscada como:

$$z_1 = Xa_1$$

  - La media de z_1 será cero y su varianza es:

$$\frac{1}{n} z_1^T z_1 = \frac{1}{n} a_1^T X^T X a_1 = a_1^T S a_1$$

Matriz de Varianza-Covarianza

Entonces Maximicemos la Varianza

Muy Fácil aumenta a_1

Solución

- Queremos un ponderador bien comportado exijámosle norma 1 ➔

$$a_1^T a_1 = 1$$

- Ahora tenemos un problema de optimización, ocupemos Lagrange:

$$M = a_1^T S a_1 - \lambda\left(a_1^T a_1 - 1\right)$$

- Maximizamos derivando

$$\frac{\delta M}{\delta a_1} = 2Sa_1 - 2\lambda a_1 = 0 \qquad \rightarrow Sa_1 = \lambda a_1$$

- O sea

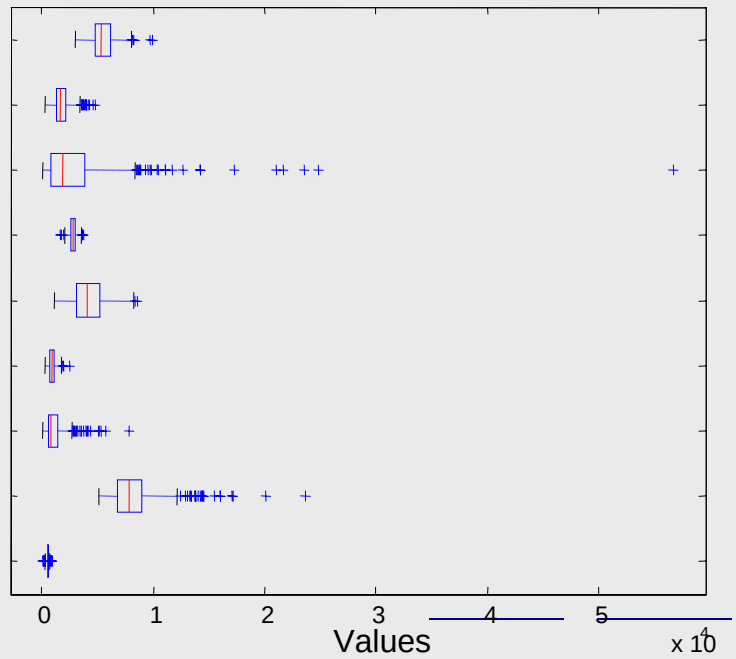
$$\left(S - I\lambda\right)a_1 = 0$$

Valores y Vectores propios

Solución

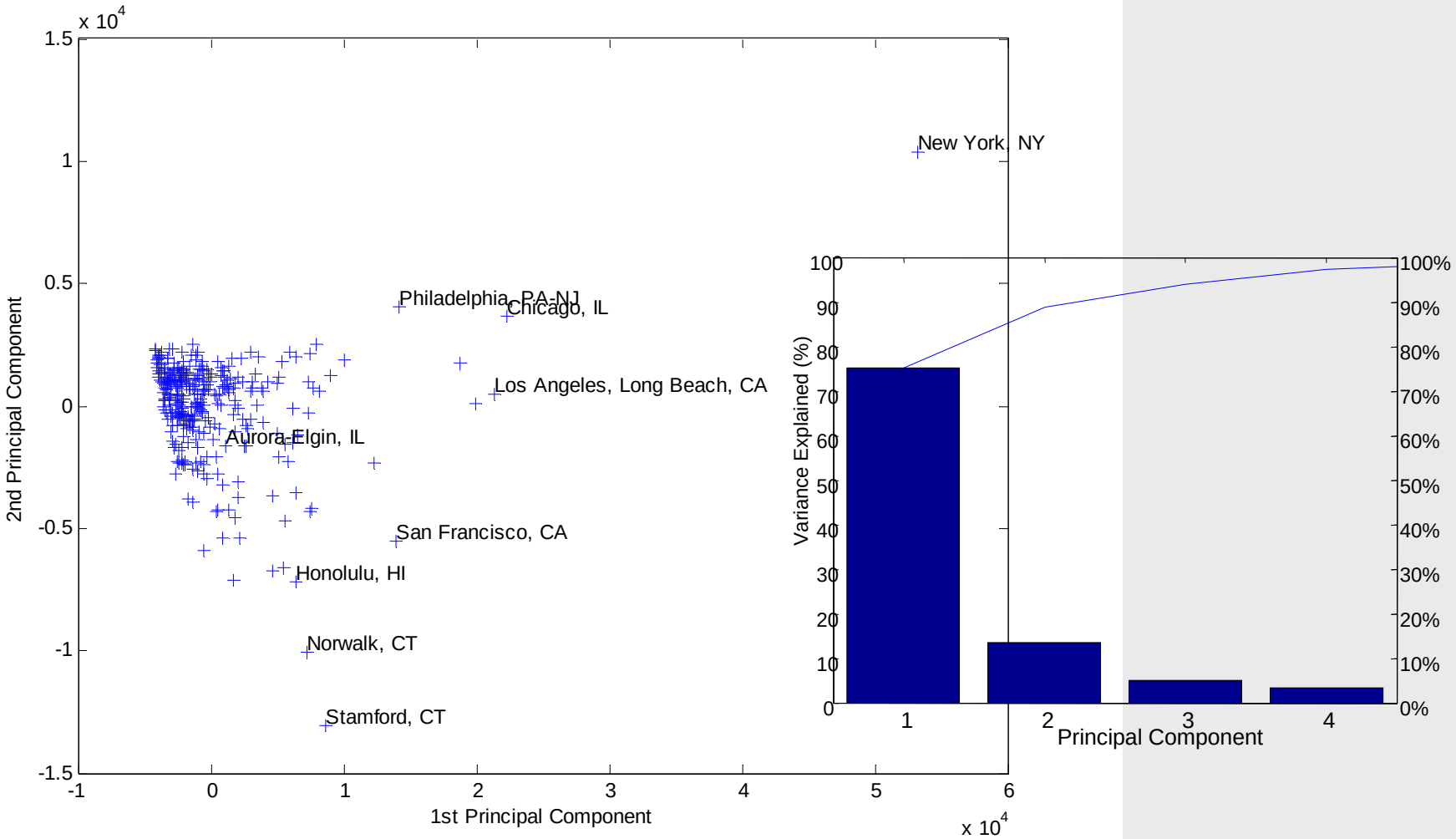
- El mayor valor del vector propio corresponde a la primera componente y así sucesivamente.
- De regreso a nuestro problema:

	a_1	a_2	a_3	a_4
Economics	0.0064	-0.0155	0.0067	0.0263
Recreation	0.2691	-0.9372	0.0826	0.1778
Arts	0.1783	0.0205	-0.0278	0.0266
Education	0.0281	0.0109	-0.0376	-0.0990
Transportation	0.1493	-0.0188	-0.9715	0.0384
Crime	0.0252	0.0014	-0.0415	-0.0216
Health	0.9309	0.2823	0.1510	-0.0278
housing	0.0698	-0.1038	-0.1496	-0.0690
climate	0.0251	-0.1734	-0.0127	-0.9745



¿Esto es lo que queremos?

Solución


x 10^4
1.5
1
0.5
0
-0.5
-1
-1.5
2nd Principal Component
-1
0
1
2
3
4
5
6
1st Principal Component
x 10^4
New York, NY
Philadelphia, PA-NJ
Chicago, IL
Los Angeles, Long Beach, CA
Aurora-Elgin, IL
San Francisco, CA
Honolulu, HI
Norwalk, CT
Stamford, CT
Variance Explained (%)
Principal Component
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%

Covarianza vs. Correlación

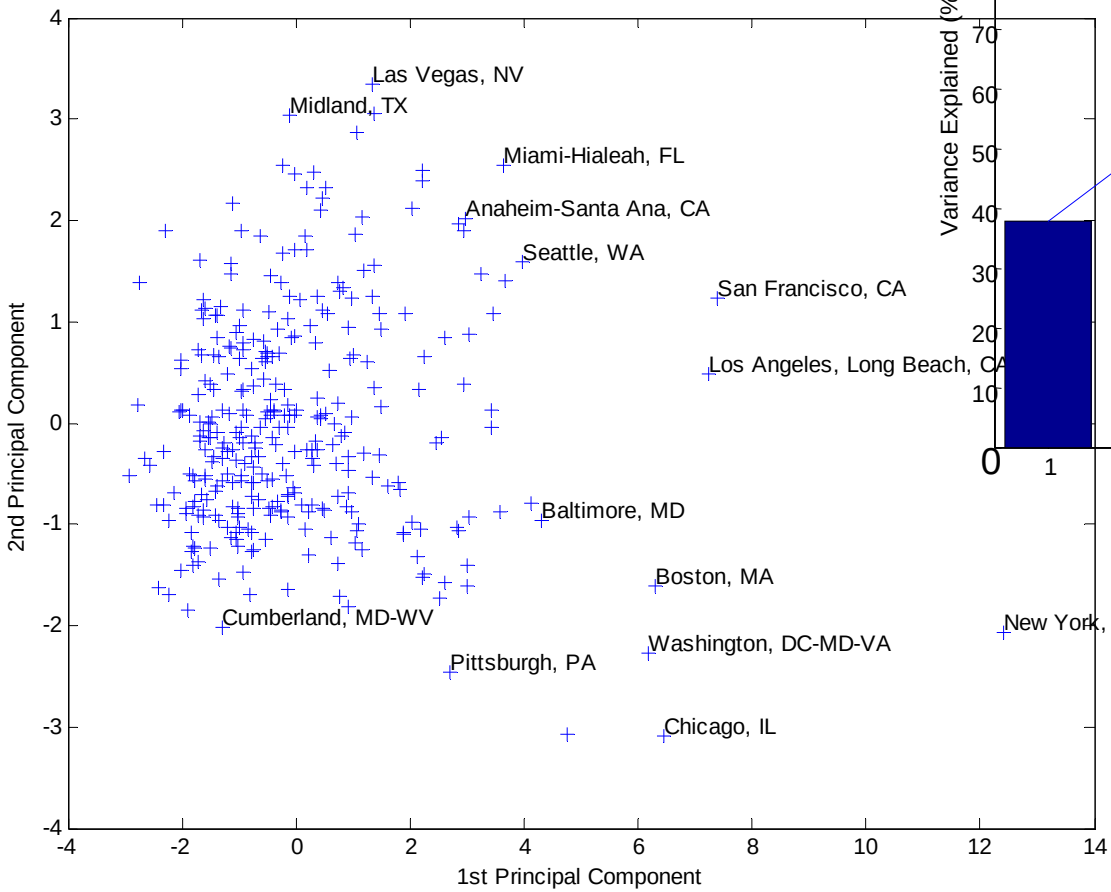
- Da lo mismo si los atributos tienen mucha diferencia entre sus varianzas ➔ ¡NO!
- Quiero que la varianza importe o no. Si la varianza es informativa si
- La opción entonces es estandarizar las variables.

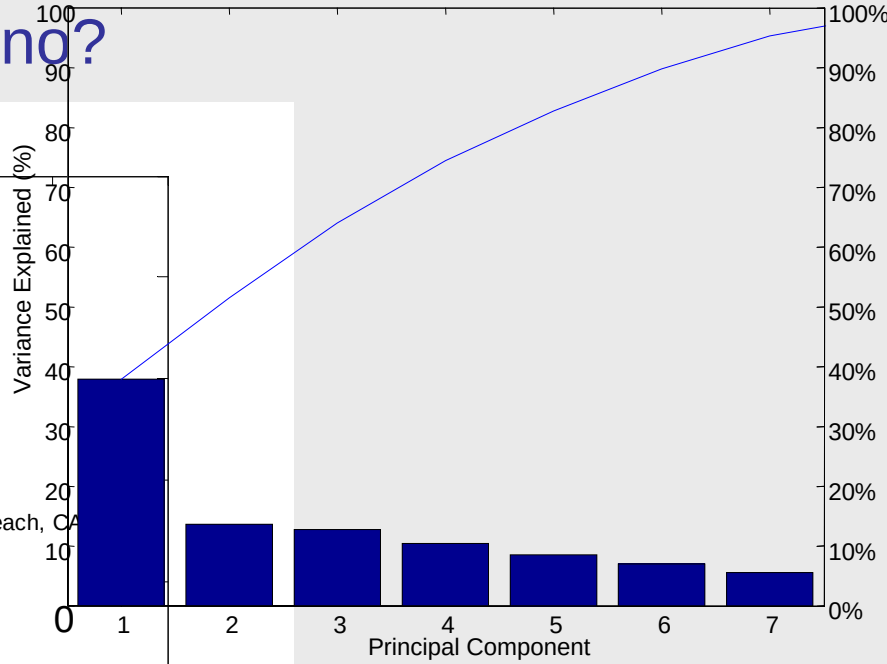
Economics	0.2064	0.2178	-0.6900	0.1373
Recreation	0.3565	0.2506	-0.2082	0.5118
Arts	0.4602	-0.2995	-0.0073	0.0147
Education	0.2813	0.3553	0.1851	-0.5391
Transportation	0.3512	-0.1796	0.1464	-0.3029
Crime	0.2753	-0.4834	0.2297	0.3354
Health	0.4631	-0.1948	-0.0265	-0.1011
housing	0.3279	0.3845	-0.0509	-0.1898
climate	0.1354	0.4713	0.6073	0.4218
	a_1	a_2	a_3	a_4

Esto parece más razonable

Covarianza vs. Correlación

- ¿Pero reduje más atributos o no?





Signos de las Componentes

- Interpretación de los valores de los ponderadores
  - Ejemplo Huba et al. (1981)
    - Muestra 1684 estudiantes en los Angeles.
    - 13 drogas
    - 5 categorías ordinales de respuesta

	a_1	a_2
cigarettes,	(0.278,	(0.280,
beer,	0.286,	0.396,
wine,	0.265,	0.392,
spirits,	0.318,	0.325,
cocaine,	0.208,	-0.288,
tranquilizers,	0.293,	-0.259,
drug storemedications used to get high,	0.176,	-0.189,
heroin and other opiates,	0.202,	-0.315,
marijuana,	0.339,	0.163,
hashish,	0.329,	-0.050,
inhalants (such as glue),	0.276,	-0.169,
hallucinogenics,	0.248,	-0.329,
and amphetamines	0.329).	-0.232).

Notas para el Data Miner

- Los datos no deben tener la media incluida (media 0)
- La variancia relativa de los atributos si importa en el análisis. (puede jugar en contra o a favor)
- Es intensiva en el uso de recursos computacionales. ($O(np^2+p^3)$)
- Sólo se puede aplicar con resultados satisfactorios a datos de valores ordinales y continuos.
- El método no garantiza que la información desechada no sea relevante.
- Los problemas de Clasificación y ‘Feature selection’ pueden no ser el mismo problema.
- Por otra parte los problemas en los problemas de regresión puede ser útil.

No Confundir

- Análisis de Componentes Principales (PCA):
  - Objetivo: Transformar las variables o atributos existentes en nuevas variables.

Análisis de Factores

- Análisis de Factores (‘Factor Analysis’):
  - Crea un modelo de los datos donde *p* variables medidas se pueden expresar como combinaciones lineales de un número pequeño de *m* variables ‘latentes’ que no puede ser medido explícitamente.
  - Trata de representar la varianza total de la base de datos.

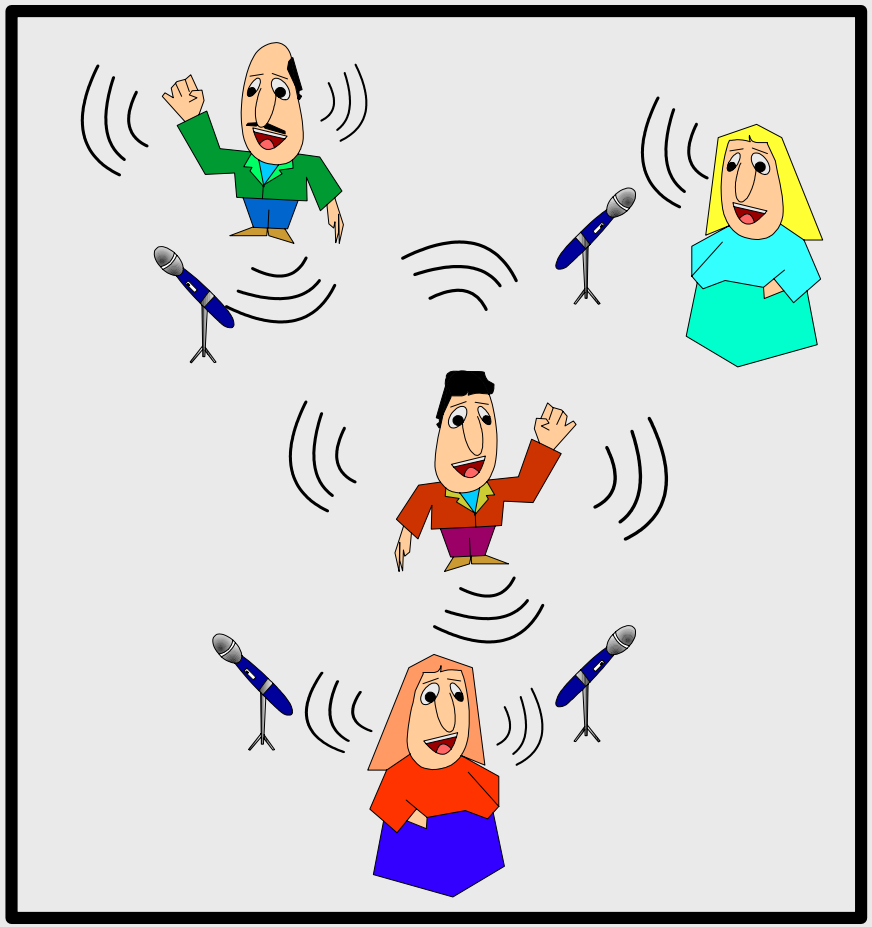
Componentes principales v/s Factorial

- El Análisis de Componentes Principales trata de hallar componentes (factores) que sucesivamente expliquen la mayor parte de la varianza total. Por su parte el Análisis Factorial busca factores que expliquen la mayor parte de la varianza común.
- El Análisis Factorial supone que existe un factor común subyacente a todas las variables, el Análisis de Componentes Principales no hace tal asunción.
- En el Análisis de Componentes Principales, el primer factor o componente sería aquel que explica una mayor parte de la varianza total, el segundo factor sería aquel que explica la mayor parte de la varianza restante, es decir, de la que no explicaba el primero y así sucesivamente. De este modo sería posible obtener tantos componentes como variables originales aunque esto en la práctica no tiene sentido.

Independent Component Analysis

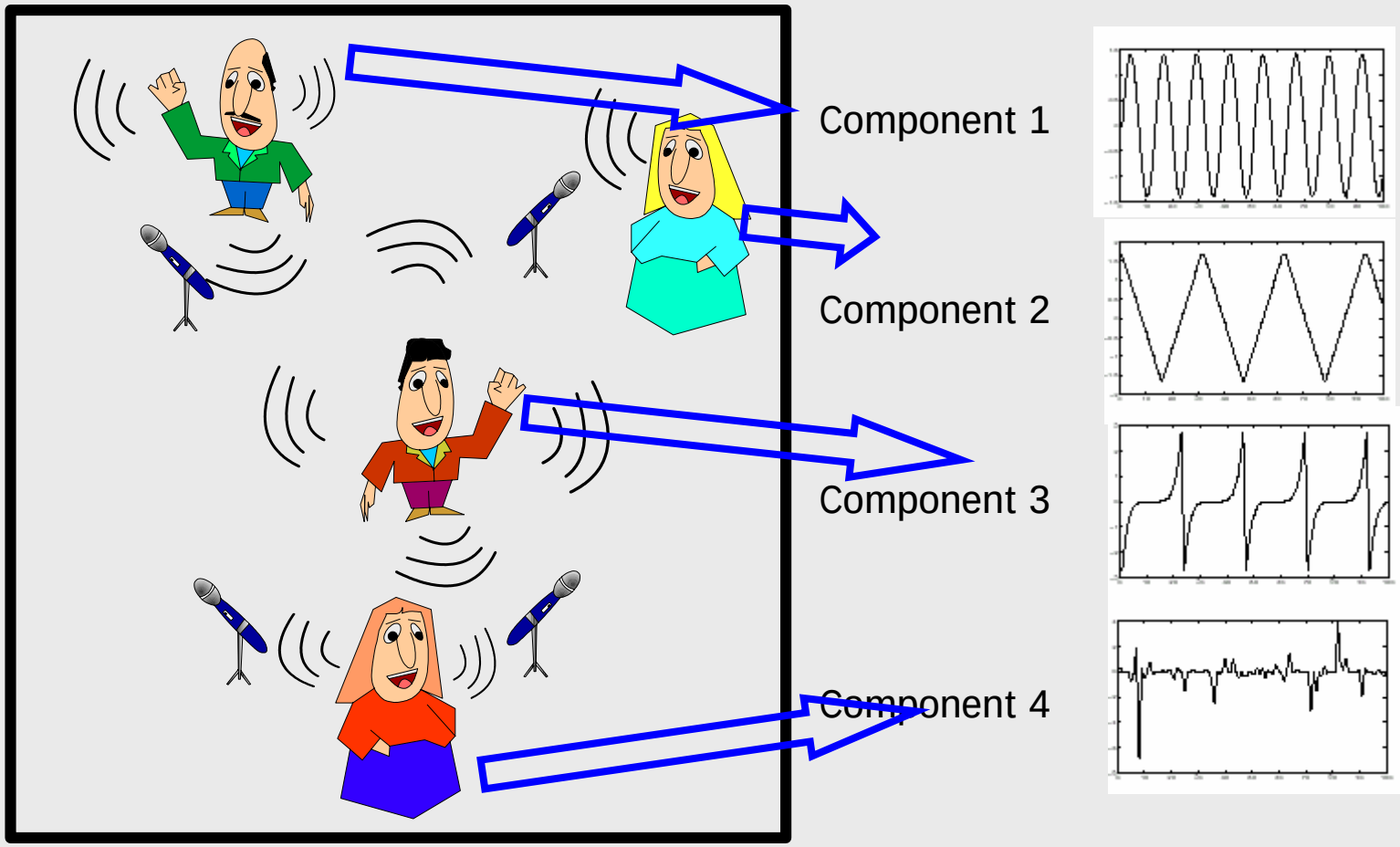
- ICA: Independent Component Analysis:
  - Used to separate statistically independent signals.
- Example:
  - Cocktail Party Problem

Cocktail Party Problem



Independent
Component Analysis

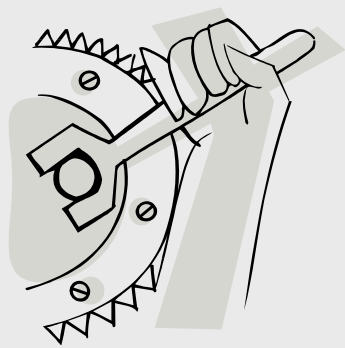
ICA for Cocktail Party Problem



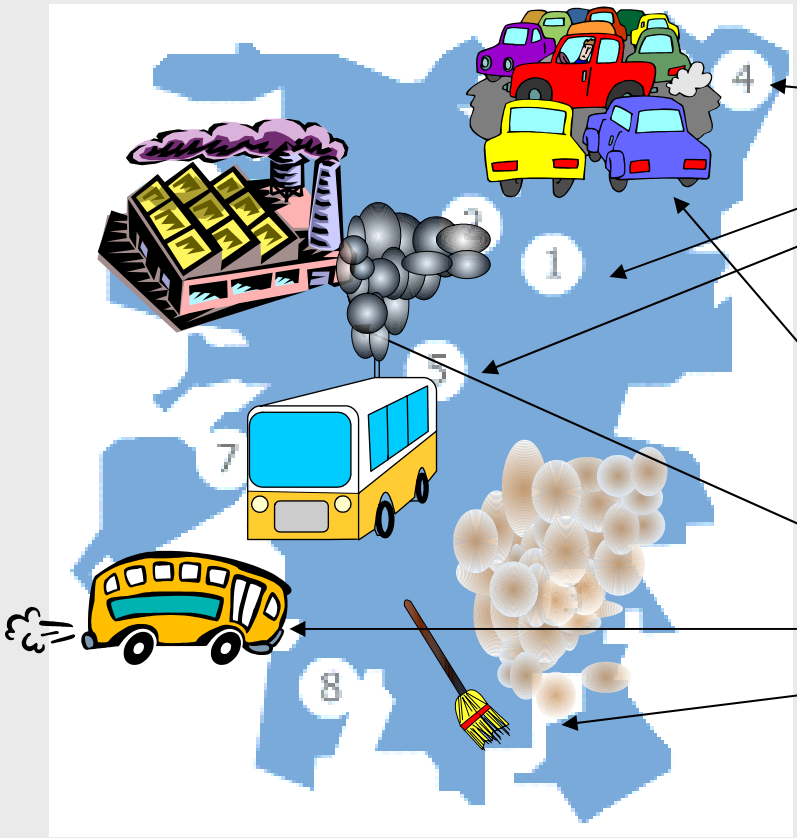
Applications of ICA

- Financial Time Series
- Image Processing
- … …

- http://www.cis.hut.fi/projects/ica/



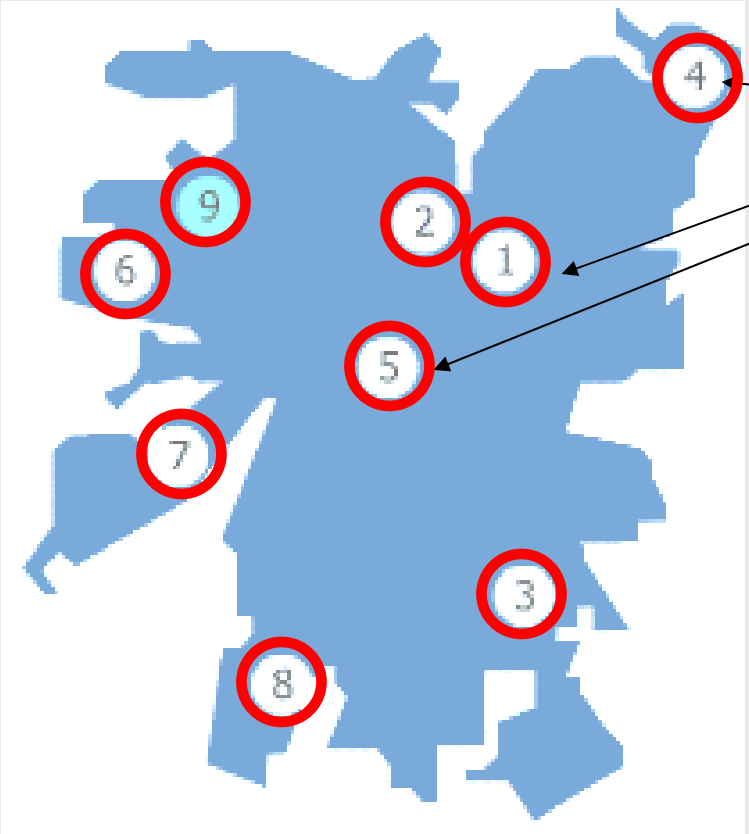
Application in Santiago



Monitoring stations = Microphons

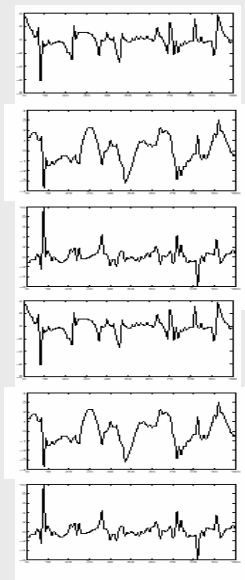
Sources of Contamination = Persons

Application: Prediction of Smog



Independent Component Analysis

Monitoring Stations in Santiago



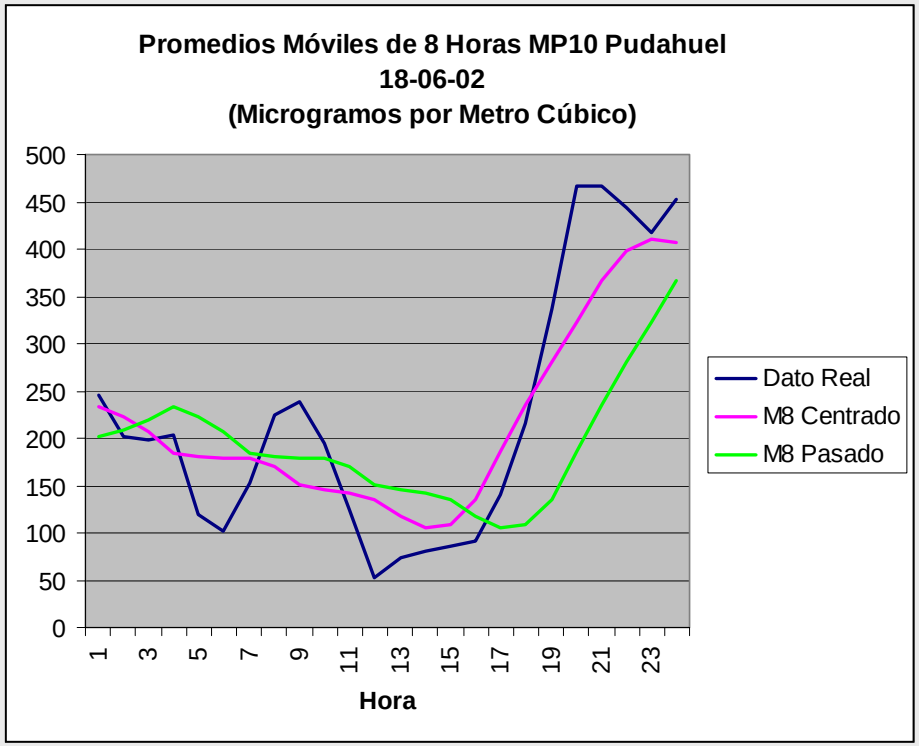
Contaminants:

- CO
- SO_2
- NO/NO_2
- O_3
- MP_{10}
- Others

Available data

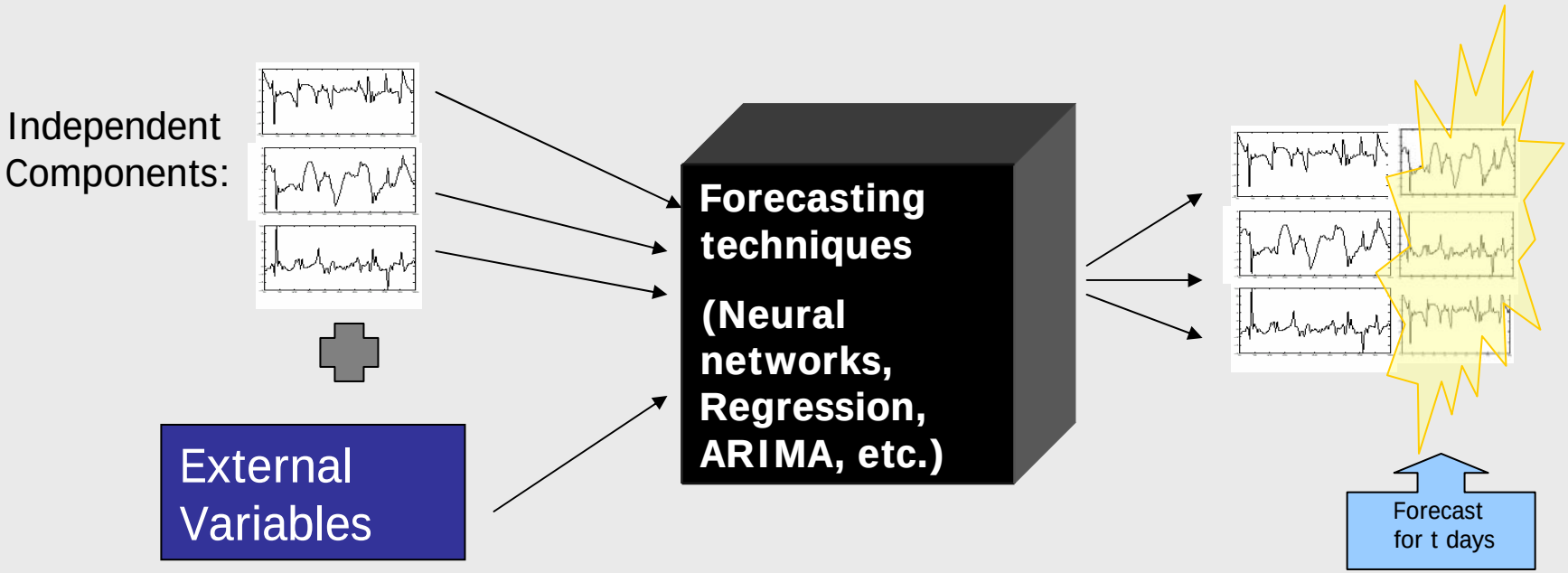
AÑO	CO	MP10	O3	SO2	MP25	NO2
1997						
1998						
1999						
2000						
2001						
2002						

Preprocessing: Moving average



Forecasting

- Independent components + external variables (weather, emergency measures, holidays, etc) as input to forecast each component for t days.



ICA Model

We have a system of equations with variables s_i ("latent variables"):
$x_j = a_{j1}s_1 + a_{j2}s_2 + .. + a_{jn}s_n$, for each sensor j

$$\mathbf{x} = \mathbf{As}$$

Where:

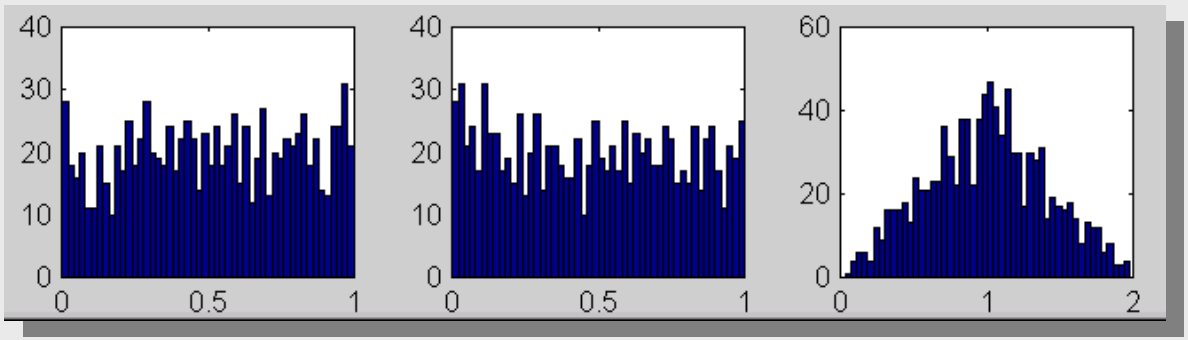
- x = Measurements
- A = (unknown) Matrix
- s = (unknown) Real Sources

Determine A and s using x
Having A we can determine $W=A^{-1}$ in order to calculate:

$$\mathbf{s = Wx = A^{-1}x}$$

How?

- Central Limit Theorem:
  "The sum of independent random variables converges to a Gaussian Distribution."

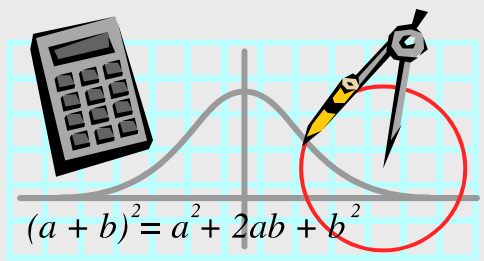


- **Definition 1** (General definition) ICA of the random vector x consists of finding a linear transform A so that the components si are **as independent as possible**, in the sense of maximizing some function ***F***(***s***1,...,***sm***) that measures independence.

ICA Model

How to measure independence?

- Kurtosis
- Entropy
- Neg-Entropy
- Minimizing mutual information



ICA Model

Assumptions:

- Sources are Independent
- At most one source is gaussian
  - Cannot distinguish two gaussian sources

ICA

- Ejemplo de ICA funcionando:

http://www.cis.hut.fi/projects/ica/cocktail/cocktail_en.cgi