Big Data y Estadística

Daniel Peña
Universidad Carlos III de Madrid

Workshop on Big Data and Statistics UDC, junio 2015

Índice

1. Introducción: el mundo de Big Data
2. Big Data y Estadística
3. Estadística y otras disciplinas: Data Mining y Knowledge Discovery in DataBases (KDD)
4. Tres ejemplos de técnicas para análisis de Big Data
5. Conclusiones

1. El mundo de Big Data

Desde hace pocos años el crecimiento de los bancos de datos es exponencial.

- Cualquier aparato digital genera **gratis** mucha información sobre su uso.(Medidores, aparatos TIC, redes, etc)
- La web genera mucha información valiosa sobre los usuarios y los contenidos
- La presión por mayor transparencia pondrá cada vez más datos a disposición de todos

Del texto a los videos (Tamaño digital de objetos)

Nombre	Tamaño	Interpretación
bit (b)	1 , 0	Nombre dado por Tukey a un dígito binario (2 posibilidades).
Byte (B)	8 bits	una letra o un número en lenguaje informático (2^8 = 256 posibilidades)
Kilobyte (KB)	1000 B	TEXTO Una página de texto escrito son entre 2 y 20 KB según programa. Un libro medio en epub 400 KB (un diskette 1972: 150 KB, 1984: 720 KB)
Megabyte (MB)	1.000 KB	MUSICA e IMAGEN Una canción moderna en mp3 son 4 MB (10 libros), una sinfonía 80 MB (200 libros), una foto de 40 KB a 4 MB (diskette 1997: 240 MB, CD, 1984 700 MB).
Gigabyte (GB)	1.000 MB	VIDEOS y PELICULAS Una película de hora y media ocupa 1 Gigabyte (250 canciones y 2500 libros) (DVD 4,7/8,7 GB)
Terabyte (TB)	1.000 GB	2,5 millones de libros, (Todos los libros de la Biblioteca del Congreso de EE.UU. en 15 TB) , todas las películas en 300 TB
Petabyte (PB)	1.000 TB	Google procesa 25 PB al dia
Exabyte (EB)	1.000 PB	Toda la información existente hasta 2013 600EB
Zettabyte (ZB)	1.000 EB	
Yottabyte (YB)	1.000 ZB	

El fenómeno Big Data

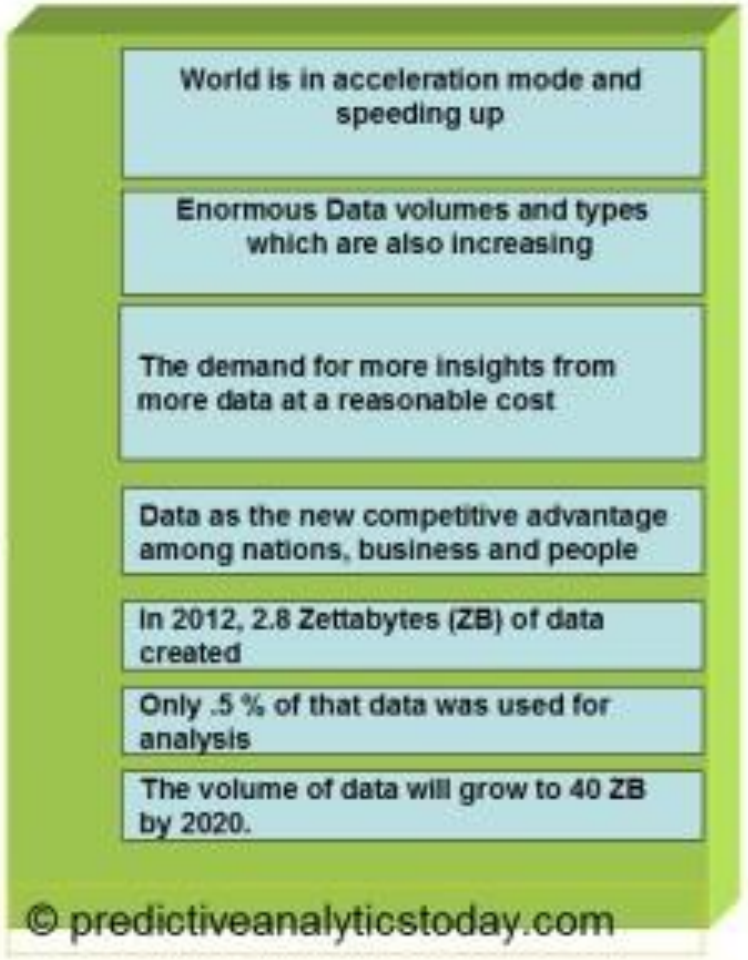
En Google (mayo 2015):

Computer	2.330 M
Statistics:	1.560 M
Big Data	737 M
Mathematics	250M

No es un fenómeno pasajero

Su interés crece con

gran velocidad



El fenómeno Big Data

- 1. World Data Centre for Climate El WDCC (Centro Mundial de datos para el clima), base de datos más grande del mundo. Almacena unos 400 terabytes de información sobre el clima en todo el mundo.

- 2. National Energy Research Scientific Computing Center El NERSC investiga distintos tipos de energía. Su base de datos tiene 2.8 Petabytes.

- 3. AT&T. compañía de telecomunicaciones. almacena 350 terabytes de información.

  4. Google Recibe más de 100 millones de consultas al día. Se supone que almacena cientos de terabytes de información.

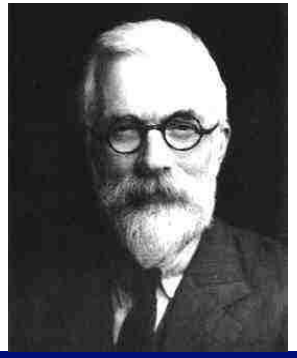
El fenómeno Big Data

- El efecto cuantitativo de aumentar la dimensión de los datos disponibles, Big Data, va a producir cambios cualitativos en el análisis de datos.
- En física al aumentar la velocidad con la que un objeto se mueve respecto a otro y acercarse a la de la luz necesitamos la relatividad y no la física clásica.
- Algo parecido puede ocurrir en el análisis de datos: necesitamos nuevos métodos para datos masivos.

2. Estadística ¿Qué cambiar?

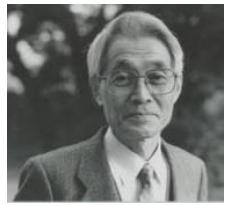
La Estadística fue creado por K. Pearson y RA Fisher para aprender de los datos hace un siglo. La gran mayoría de la estadística que se enseña hoy sigue estando bajo esta influencia.

- Muestras pequeñas (n<200)
- Descripción univariante de cada variable
- Modelos paramétricos simples y escuetos
- Estimación óptima (suficiencia, eficiencia) y contraste de hipótesis
- Inferencia: homogeneidad de los datos y utilización óptima de la información
- Contrastes de ajuste



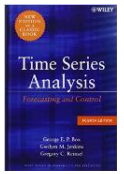
5 Cambios clave en 70’s y 80’s por los nuevos datos y métodos de cálculo

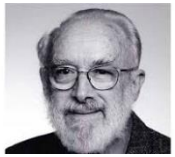
- La importancia de criterios automáticos de selección de modelos. Akaike, 1973.

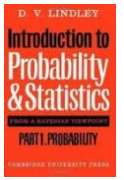


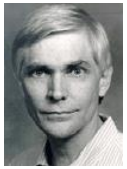
- El valor del análisis de datos como herramienta exploratoria. Tukey, 1977.

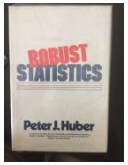
- Series temporales para predicción y modelos ARIMA. Box y Jenkins 1970.
- Métodos Bayesianos (Box and Tiao,Lindley, 70’s)
- Métodos Robustos (Huber, 1981)







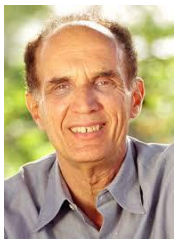


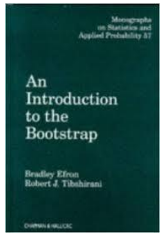


Cambios en los últimos 30 años

- El poder de la informática: Bootstrap

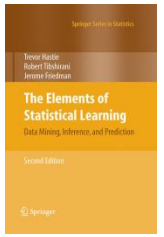
(Efron,1991), MCMC (Gelfand, Smith,1990)





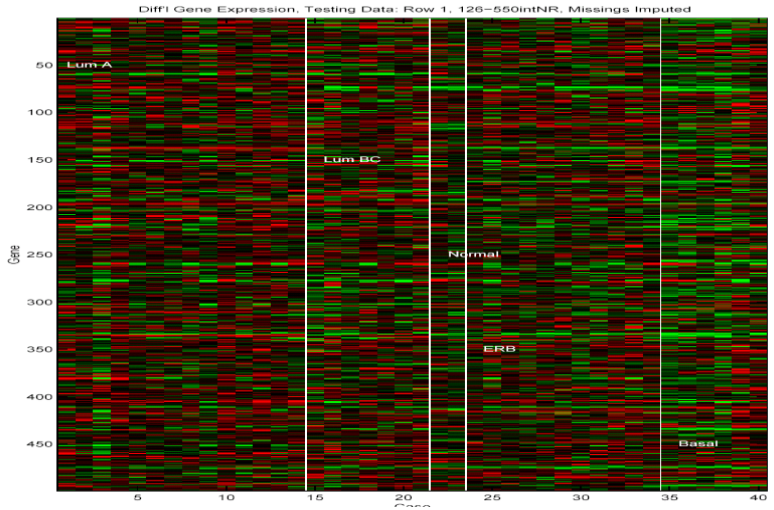
- Estimación con ajuste+penalización (splines

Ridge regression, efectos aleatorios, cross validation, lasso, etc), Statistical Learning: Hastie, Tibshirani

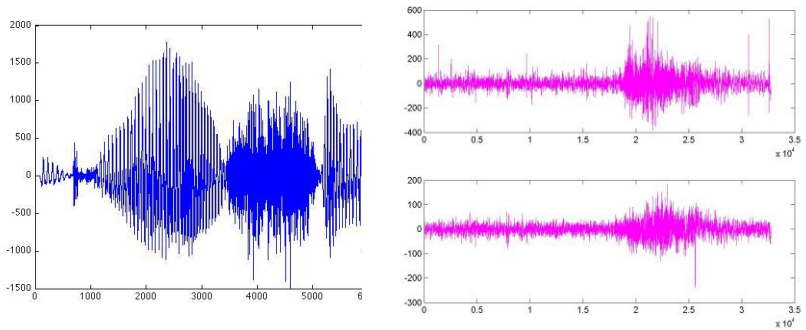


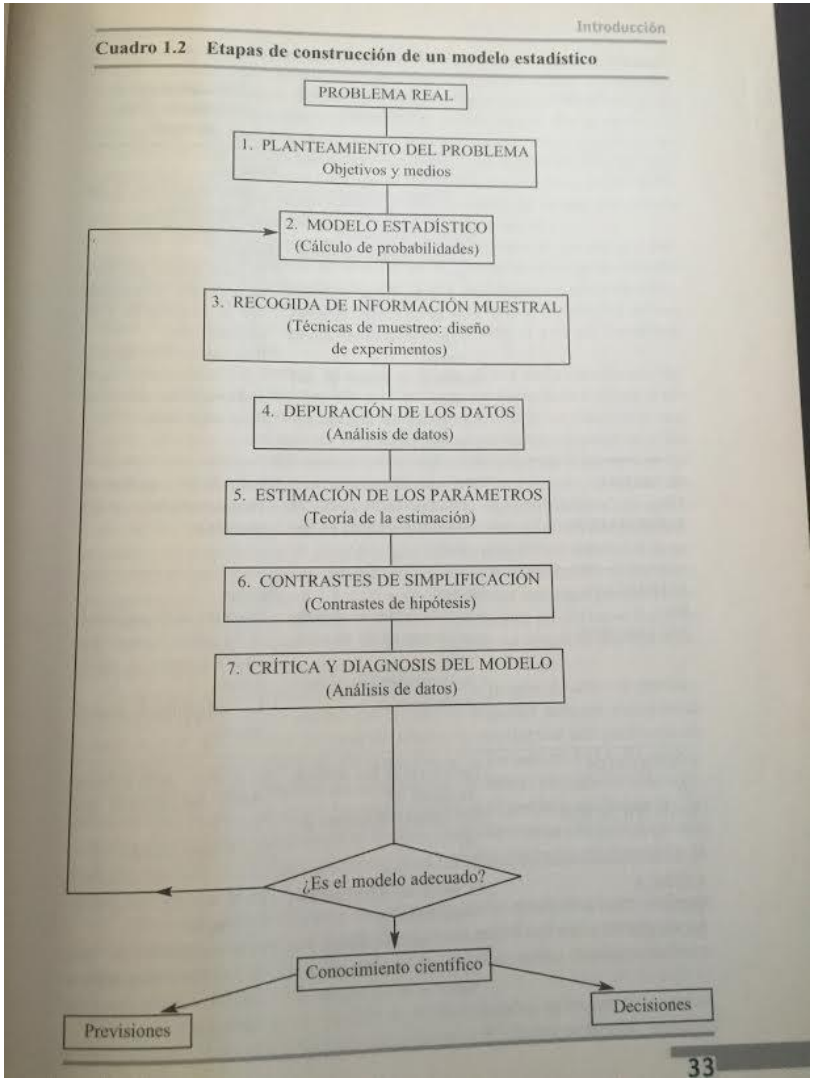
Cambios en Datos

- Los datos pueden ser objetos y no solo mediciones (Digital information, matrices/images, functions, surfaces in many dimensions, texts, social network messages,…)



- Estos datos requieren modelos complejos y heterogéneos (Dynamic, Multivariate, Non parametric/semiparametric).





Exploracion de datos como etapa inicial

Busar relaciones y estructuras

Utilizarlas para prever y comprender

Fuente: Peña ,D. (2005). Fundamentos de Estadística, Alianza Editorial

Algunos cambios en la metodología estadística

- Definimos un objetivo y tomamos una muestra homogénea de una población definida. (no tenemos los datos inicialmente)

(Tenemos los datos y el primer paso es descubrir su estructura)

- Queremos hacer inferencia respecto a unos pocos parámetros de un modelo establecido.

(No sabemos como parametrizar y cuantos grupos tenemos)

- Es importante Eficiencia y parsimonia: aprovechar bien los datos.

(eficiencia no es importante tenemos casi la población)

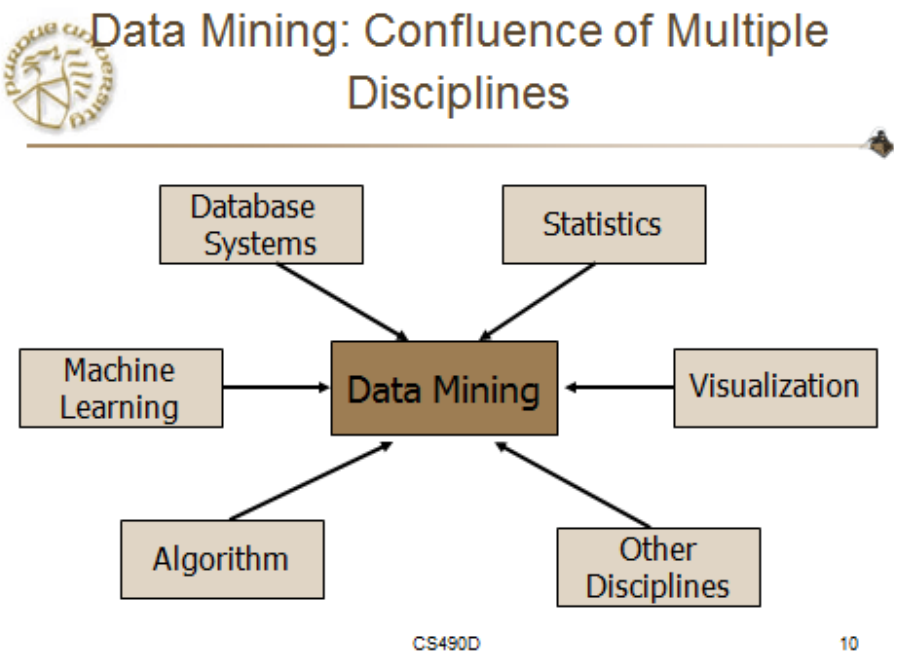
Cambios de enfoque con grandes bases de datos

- Nuevos métodos gráficos para resumir las variables y su estructura
- Esperar Heterogeneidad: diferentes modelos en diferentes zonas del espacio
- Eficiencia menos importante que robustez.
- Existencia de grupos y heterogeneidad. Necesidad de clasificar y segmentar
- Ajuste del modelo: Se rechaza cualquier contraste de ajuste a un modelo paramétrico simple

From Big Data to Big Statistics
John Sall, SAS

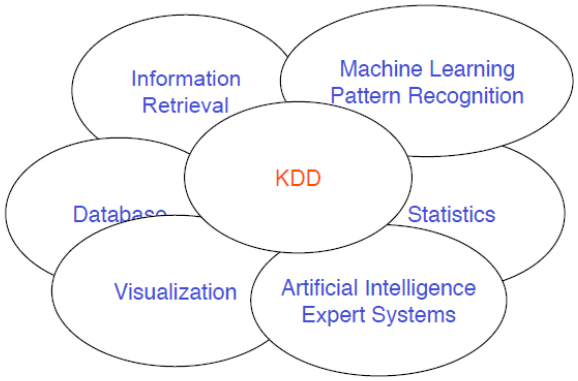
Now that we have lots of data and can process it amazingly fast, we still need ways to look at it without being overwhelmed. We don't want to look at 10,000 graphs--we want one graph that shows the bright spots among 10,000 graphs. We need volcano plots and false-discovery-rate plots. We want the computer and software to do the work of finding what is most interesting and bringing it to our attention. We want our results sorted and summarized, but with access to the detail we need to understand it. Also, when we look at the most significant of thousands of statistical tests, we want to know if we are seeing random coincidence selected out of thousands, or if we are seeing real effects

3. Estadística y otras disciplinas

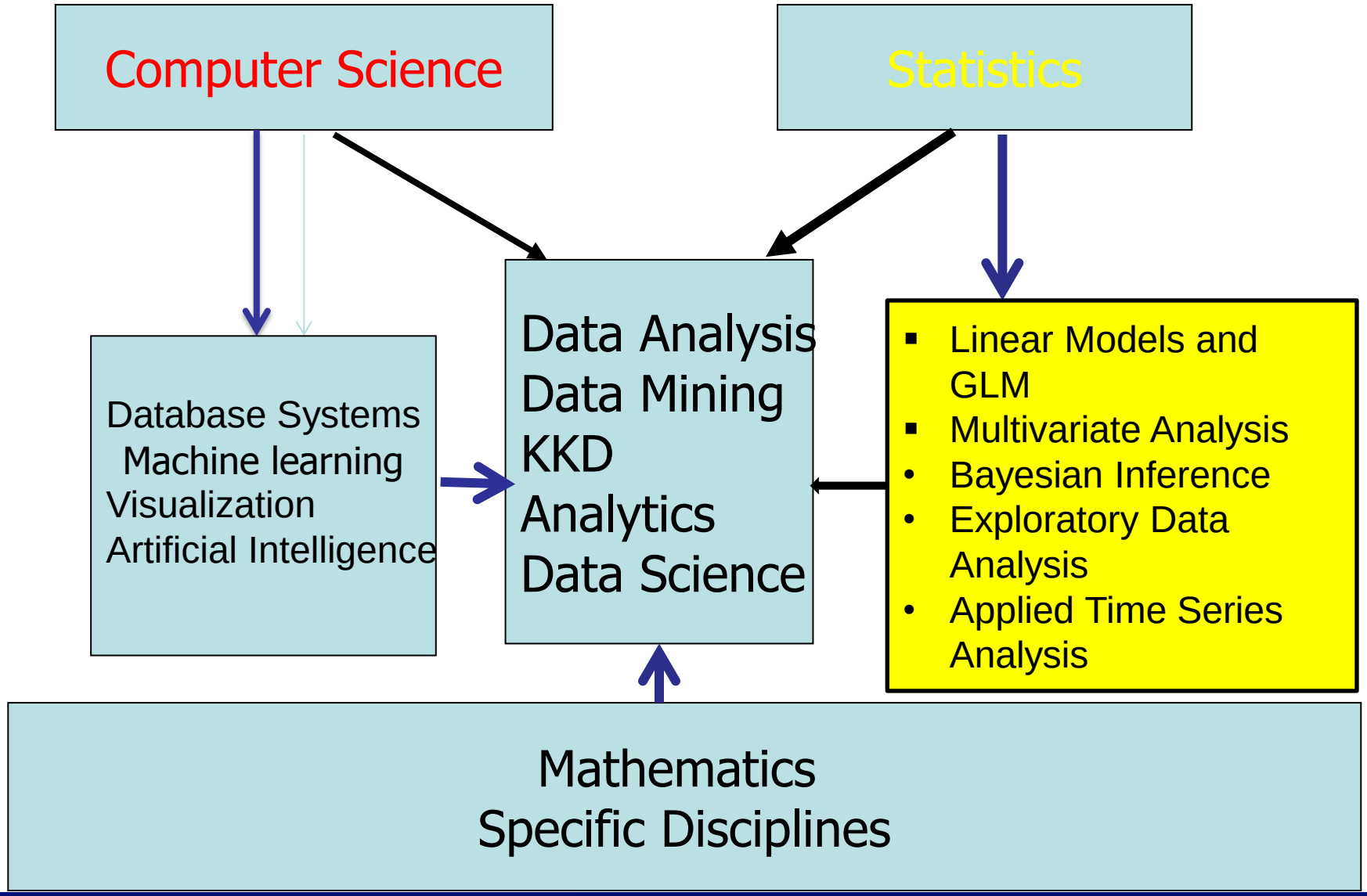


Knowledge Discovery in Databases

KDD is a multidisciplinary field



Fuente: Introduction to Data Mining
Prof. Chris Clifton, Pardue U.


Computer Science
Statistics
Database Systems
Machine learning
Visualization
Artificial Intelligence
Data Analysis
Data Mining
KKD
Analytics
Data Science
Linear Models and GLM
Multivariate Analysis
Bayesian Inference
Exploratory Data Analysis
Applied Time Series Analysis
Mathematics
Specific Disciplines

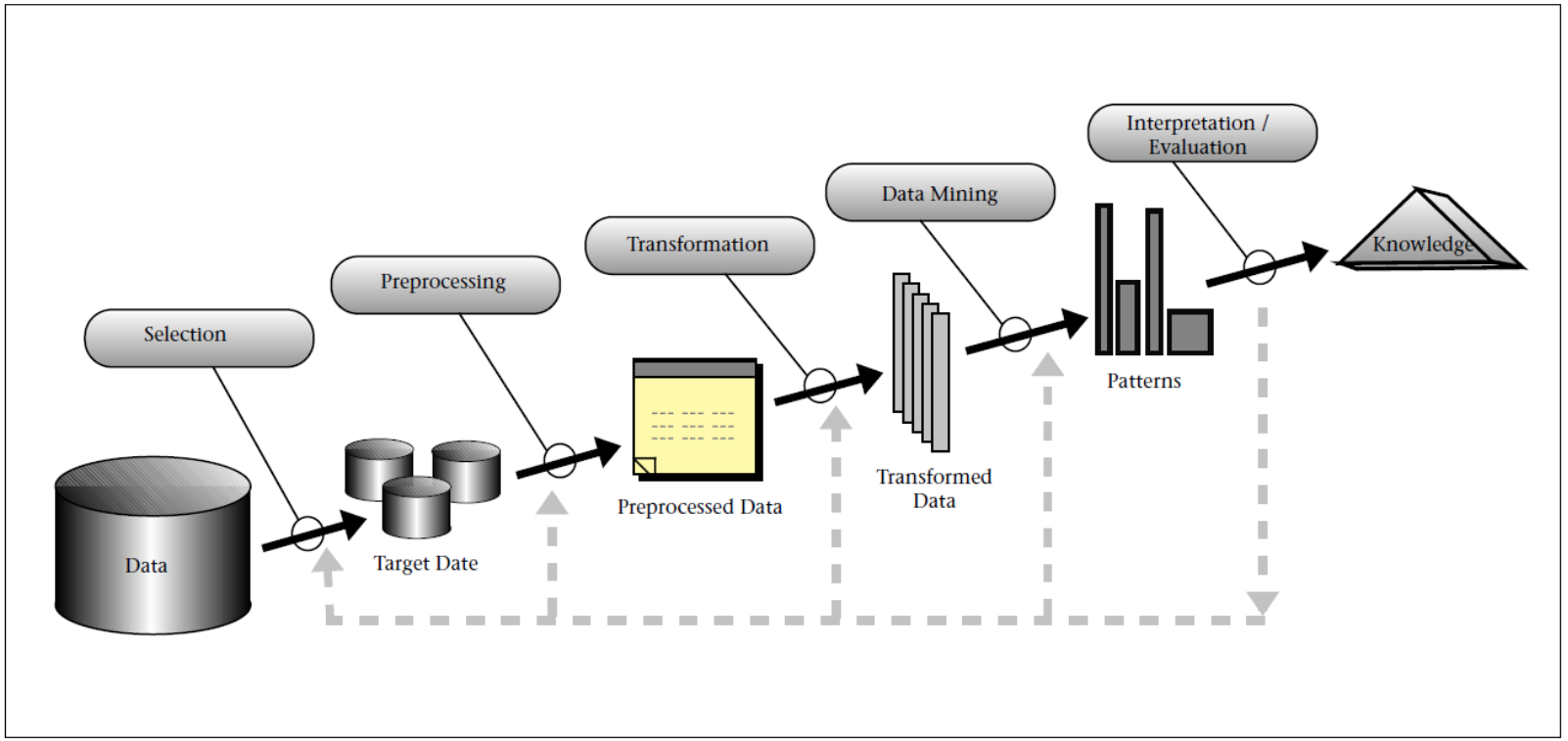


Figure 1. An Overview of the Steps That Compose the KDD Process.

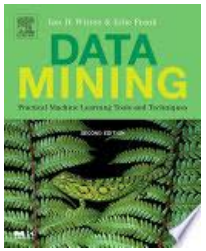
Fuente:

AI Magazine Volume 17 Number 3 (1996) (© AAAI)

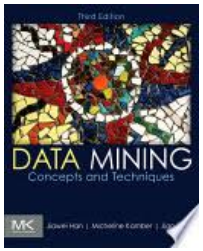
From Data Mining to Knowledge Discovery in Databases

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth

Data Mining: buscar automáticamente regularidades en los datos



Data Mining: Practical Machine Learning Tools and Techniques, Second Edition
Ian H. Witten,Eibe Frank
(25,000 ref Google)



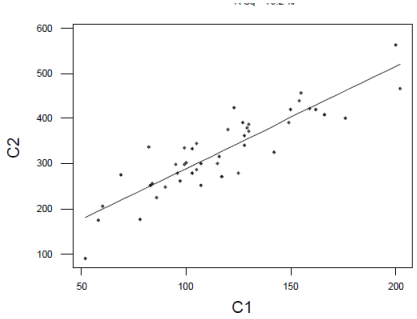
Data Mining: Concepts and Techniques
Jiawei Han,Micheline Kamber,Jian Pe
(25,000 ref Google)

- Decision trees
- Linear models
- Clustering
- Neural Networks
- Bayesian networks
- Bootstrap
- Cross-Validation
- Association rules

- Data (chap2-5)
- Association and Correlation (6/7)
- Clasification (8/9)
- Clustering (10/11)
- Outliers (12)

Riesgos de explorar sin modelo…

1. Asociación no es relación causal



Alta correlación encontrada entre:

(1) Número de matrimonios en un mes y temperatura en ese mes

(2) Cigueñas observadas en Alemania en un mes y Nacimientos en dicho mes

(3) Disminución del número de burros en España y aumento del presupuesto de educación

La simple asociación no es suficiente, la ciencia requiere una relación causal comprensible y replicable.

Riesgos de explorar sin modelo…

2. Si buscamos suficiente siempre encontramos algo que no existe.

P=probabilidad de encontrar algo falso (rechazar Ho si es cierta)

Prob que no ocurra en N casos = $(1-p)^N \simeq$*1-Np*

p=0,05 y $(1-p)^N = 0{,}95^{100} = 0{,}005$ *con seguridad nos equivocaremos varias veces al hacer contrastes múltiples*

- *False Discovery rate*= Expected value of the proportion of rejected hypothesis that are true.

¿Lo encontrado es una asociación real o una particularidad espuria de la muestra que analizamos?

Riesgos de explorar sin modelo…

3. Ojo a los sesgos en los datos

Los clientes que pagan comidas con tarjetas de crédito pueden no ser representativos de todos los clientes del Banco

Los clientes que escriben en twitter o en facebook pueden diferir en aspectos de consumo importantes respecto a los que no lo hacen

Es más probable que un cliente descontento lo diga que uno satisfecho

Es necesario comparar los clientes observados con el conjunto de referencia con una muestra representativa **(técnicas de muestreo)** o su comportamiento con un estudio controlado **(diseño de experimentos).**

Riesgos de explorar sin modelo…

4. Ojo con las conclusiones extraídas de **poblaciones heterogéneas** (paradoja de Simpson)

TABLA 12.4a. *Admisiones a una Universidad por sexo*

	Solicitudes	*Admisiones*	*Proporción*
Mujeres	2,000	1,136	56,80 %
Hombres	2,000	955	47,75 %

Desglosamos por facultades, humanidades, ingenieria y economía

		Solicitudes	*Admisiones*	*Proporción*
Hum	*Mujeres*	800	560	70 %
-------	*Hombres*	300	225	75 %
Ing	*Mujeres*	200	36	18 %
-------	*Hombres*	700	140	20 %
Econ	*Mujeres*	1,000	540	54 %
	Hombres	1,000	590	59 %

Riesgos de explorar sin modelo…

5. Con datos temporales es imprescindible tener en cuenta su dinámica.

(1) Todo lo que crece esta correlacionado

(2) Atípicos son los valores impredecibles

(3) Ante datos ausentes hay que interpolar sin alterar la estructura de correlación o introduciremos datos falsos no ausentes

(4) Las relaciones también son dinámicas con posibles retardos que hay que modelar (Econometría)

Sin embargo…ventajas de colaboración entre disciplinas distintas

- Cuando algo funciona generalmente hay una razón: neural network y modelos factoriales no lineales.
- Más énfasis en heterogeneidad y no linearidad.
- Nuevas formas de visualización mediante proyecciones y giros.
- Equilibrio entre Rigurosidad (Ciencia) y Pragmatismo (Aplicaciones)
- Métodos mejores y de mayor alcance.

4. Tres ejemplos de nuevas técnicas para Big Data

- Encontrar atípicos y grupos (cluster) dos problemas clásicos en Multivariate Analysis, Statistical Learning, Data Mining and KKD.
- Analizar datos dinámicos. Nuevos métodos de series temporales multivariantes.

Ejemplo 4.1: Identificar lo atípico

- En el firmamento la aparición de nuevos objetos para explicar la masa del universo.
- En una red social los elementos claves o virales.
- En un proceso de fabricación la aparición de defectos.

Buscar lo imprevisto

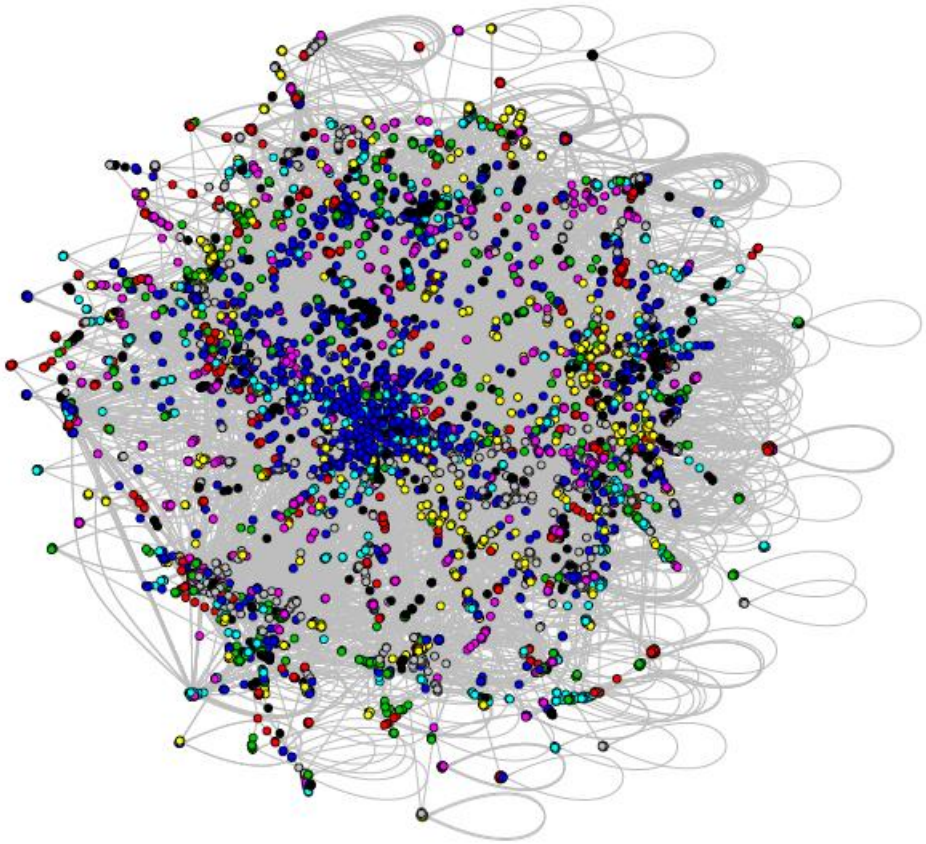
Identificar objetos en el espacio

Potentes telescopios hacen fotos continuas del cielo para identificar nuevos objetos que ayuden a explicar la masa del universo. Desde 2019 se almacenarán 30 TB cada noche

Fuente: Foto propia

Fuente: tomada de la Web

Elementos clave en redes de clientes



Fuente: Agradezco esta imagen a Pedro Galeano

También atípicos respecto a un modelo

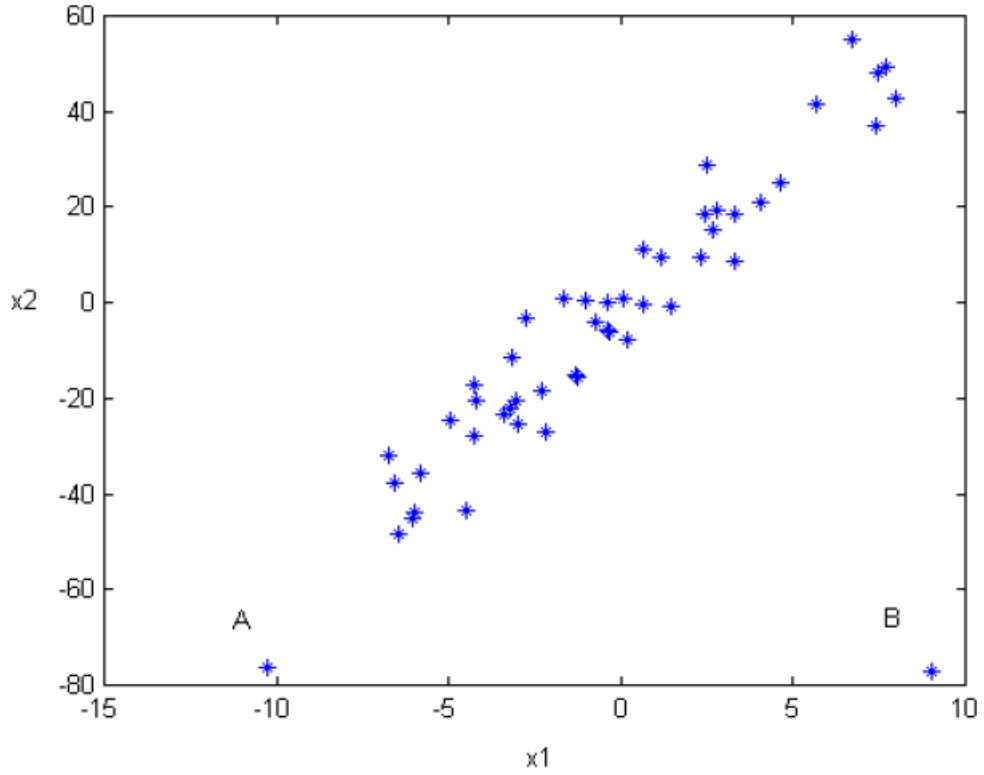


Figura 9. 16 Aunque la distancia euclidea al centro de los datos es la misma para los puntos A y B es claro que el punto B es más extremo que el A.

Fuente: Peña, D. (2010). Regresión y Diseño de Experimentos. Alianza Editorial

Ejemplo: Control de calidad por imágenes

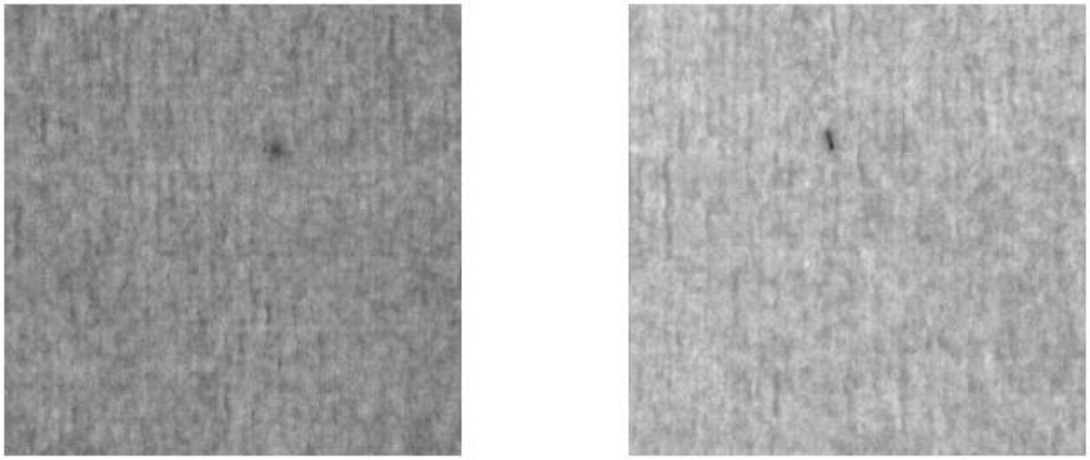


Fig. 1. Textured paper surface with a class of defect

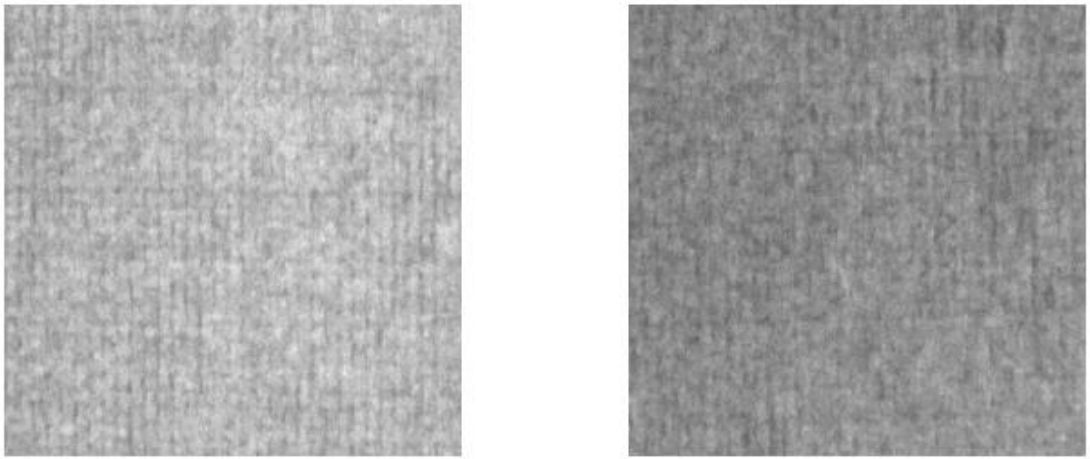


Fig. 2. Textured paper surface

Fuente: (Benito y Peña ((2007). Detecting Defects with image data. CSDA

Buscar pixels atípicos no funciona

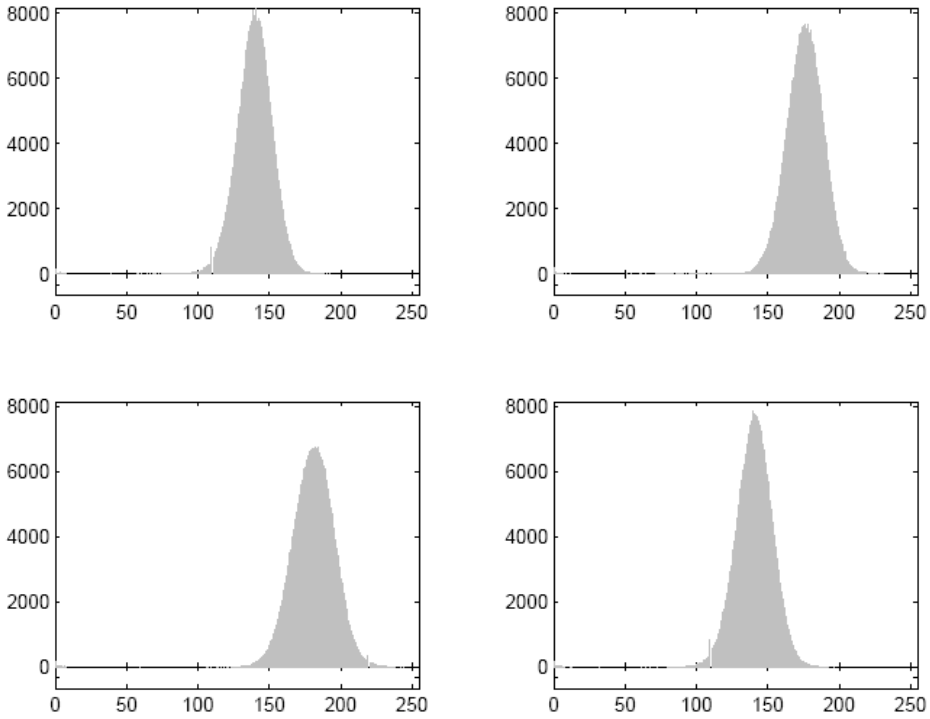


Fig. 3. Histogram for the paper surfaces of Figure 1 (top panels) and Figure 2 (bottom panels)

Fuente: Benito y Peña (2007). Detecting Defects with image data. CSDA

Propuesta: (Benito y Peña (2007). Detecting Defects with image data. CSDA

M1: Hacer bloques en la imagen y calcular la varianza por filas. Los bloques con defectos tienen más variabilidad interna

Estadístico

$$\frac{\max(S_r^2(k))}{\text{med}(S_r^2(k))}$$

M2: construir un modelo espacial CAR y prever los pixel con el modelo, los defectos serán atípicos.

Ejemplo 4.2: Formar grupos

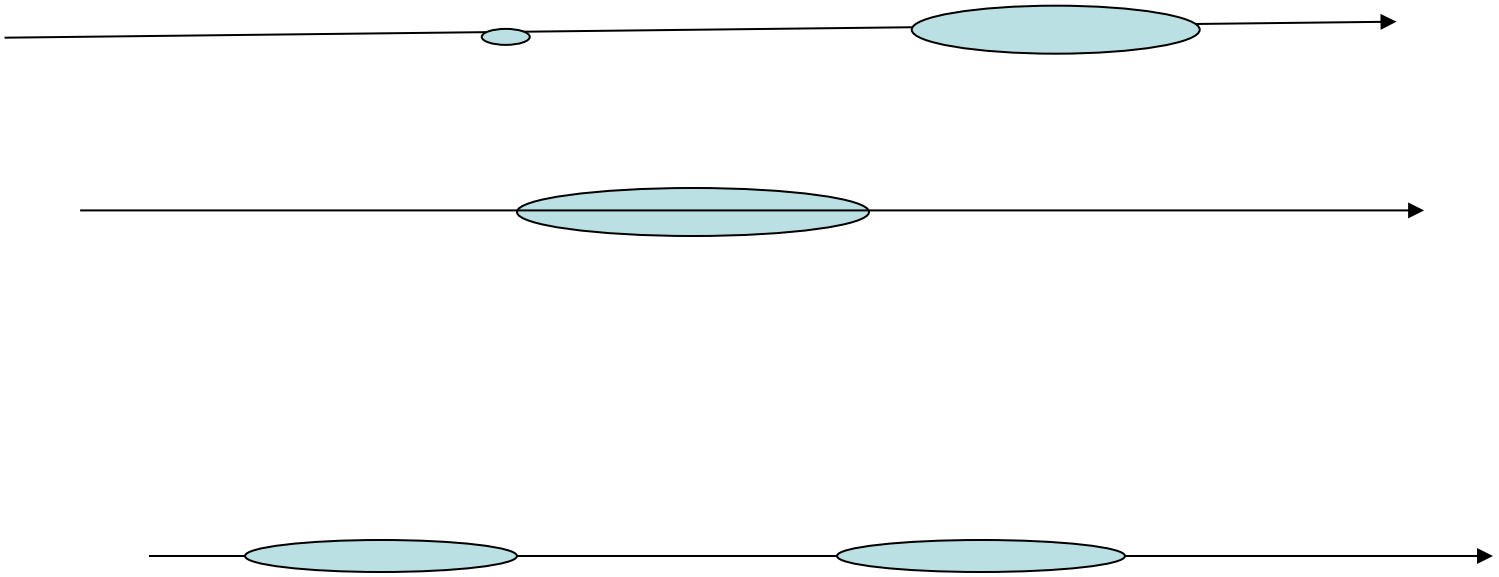
K-medias y Cluster jerárquico tienen muchas limitaciones como métodos de cluster.

Cuando p, número de variables, y n, número de observaciones son grandes conviene acudir a métodos de proyección

- Idea central: buscar direcciones de proyección que muestren la heterogeneidad de una muestra.
- Proyectar los datos y buscar grupos sobre las proyecciones
-

Heterogeneidad

¿Cómo encontrar direcciones que muestren la heterogeneidad?



Heterogeneidad univariante

- Llamemos

$$d_{ij} = (x_{ij} - \overline{x}_j)^2$$

A la variabilidad de una variable (la j) respecto a su media

se define la *desviación típica* por:

$$s_j = \sqrt{\frac{\sum_{i=1}^{n} d_{ij}}{n}} = \sqrt{\frac{\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2}{n}}$$

$$H_j = \frac{\frac{1}{n}\sum_{i=1}^{n}(d_{ij} - s_j^2)^2}{s_j^4}.$$

Este coeficiente es siempre mayor o igual a cero. Desarrollando el cuadrado del numerador como $\sum_{i=1}^{n}(d_{ij}-s_j^2)^2 = \sum_{i=1}^{n} d_{ij}^2 + n s_j^4 - 2 s_j^2 \sum_{i=1}^{n} d_{ij}$ este coeficiente puede escribirse también como:

$$H_j = \frac{\frac{1}{n}\sum (x_{ij} - \bar{x}_j)^4}{s_j^4} - 1 = K_j - 1.$$

El primer miembro de esta expresión, K_j, es una forma alternativa de medir la homogeneidad y se conoce como *coeficiente de kurtosis*. Como $H_j \geq 0$, el coeficiente de kurtosis será igual o mayor que uno. Ambos coeficientes miden la relación entre la variabilidad de las desviaciones y la desviación media. Es fácil comprobar que :

Kurtosis, para la normal =3

1. Si hay unos pocos datos atípicos muy alejados del resto, la variabilidad de las desviaciones será grande, debido a estos valores y los coeficientes de kurtosis o de homogeneidad serán altos.

Coef. Kurtosis =12

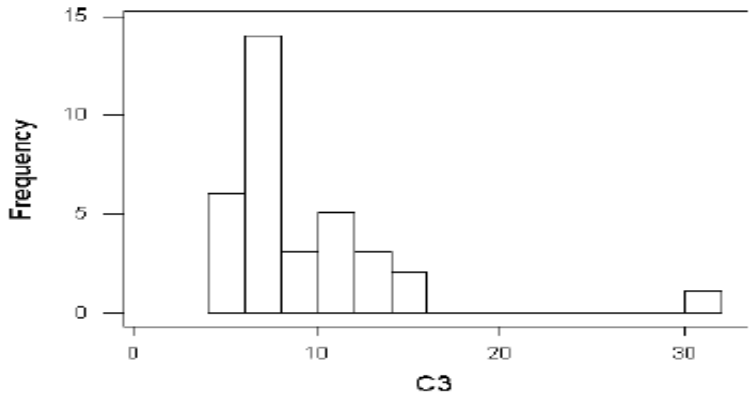
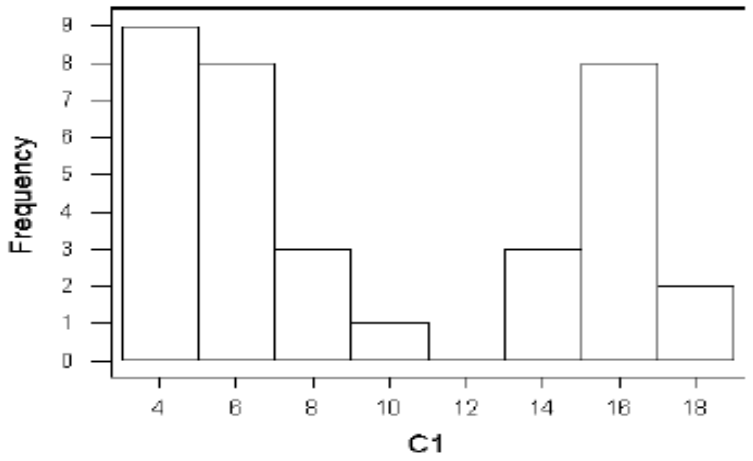


Figura 3.3: Histograma del precio por acción con relación a los beneficios (per)

2. Si los datos se separan en dos mitades correspondientes a dos distribuciones muy alejadas entre sí, es decir, tenemos dos conjuntos separados de datos distintos, la media de los datos estará equidistante de los dos grupos de datos y las desviaciones de todos los datos serán similares, con lo que el coeficiente H_1 será muy pequeño (cero en el caso extremo en que la mitad de los datos son iguales a cualquier número, $-a$, y la otra mitad igual a a).

Coef. Kurtosis= 1.38



Histograma de la rentabilidad por dividendos.

Resultado principal

- Si los datos han sido generados por dos normales multivariantes con la misma matriz de varianzas, la dirección que minimiza la kurtosis es la dirección óptima de Fisher para la discriminación cuando sabemos que hay dos poblaciones normales.

Peña, D. y Prieto, J. (2001). *"Cluster Identification using Projections" The Journal of American Statistical Association, 96, 456, 1433-1445, 2001*

Búsqueda de clusters sobre las direcciones

Se proyectan los datos en 2p direcciones y se utilizan los espacios, (spacing) o distancia entre los estadísticos de orden para encontrar grupos univariantes

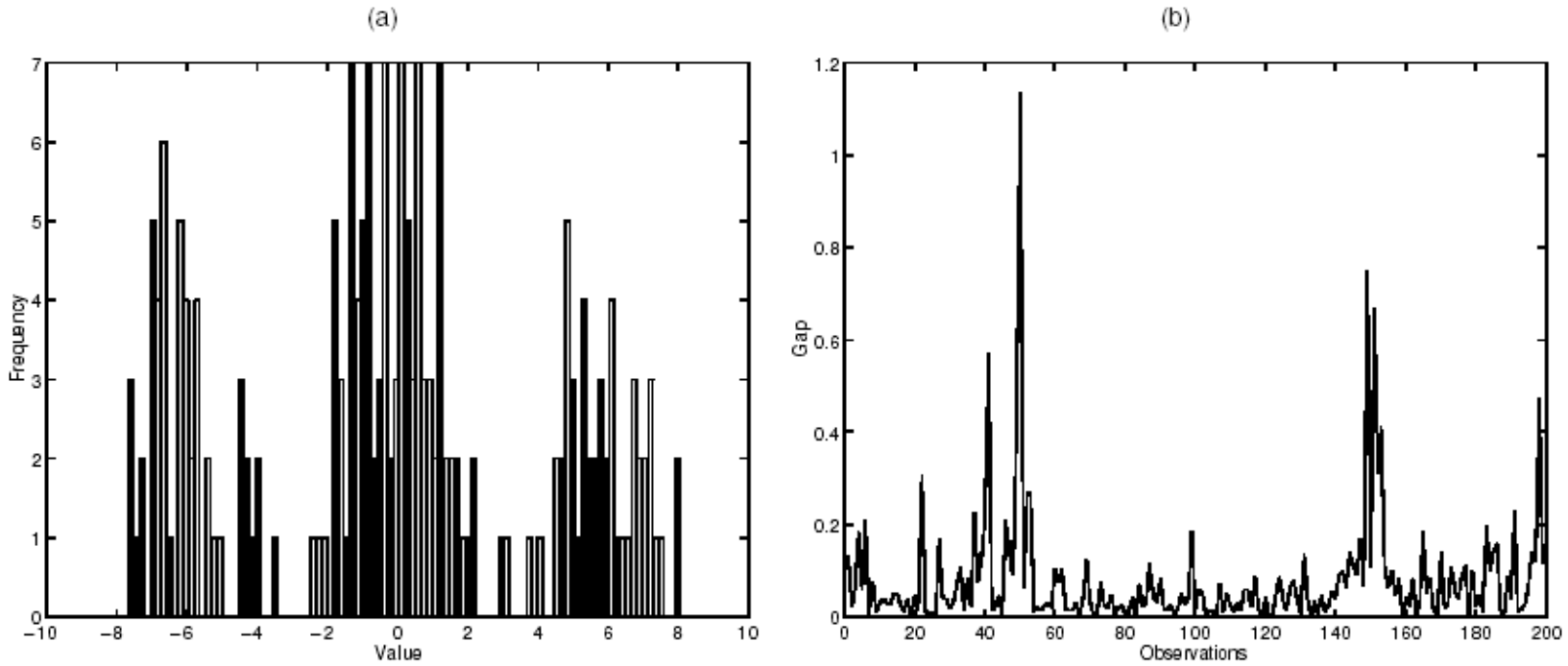


Figure 1. (a) Histogram for a Set of 200 Observations From Three Normal Univariate Distributions. (b) Gaps for the Set of 200 observations.

Método: buscar 2k (k<p) direcciones ortogonales de máxima y mínima kurtosis y buscar cluster en las proyecciones.

1. Es muy eficiente en dimensión alta
2. Es óptimo para mezclas de normales con la misma varianza
3. Asegura alta capacidad de separación lineal para cualquier distribución

Resultados

Table 4. Percentages of Mislabeled Observations for the Suggested Procedure, the k-means and Mclust Algorithms, and the Jones and Sibson Procedure (different overlaps between clusters)

	Kurtosis	*k means*	*Mclust*	*J&S*
Normal				
1% overlap	.09	.15	.17	.29
8% overlap	.15	.17	.22	.36
Uniform				
1% overlap	.05	.19	.12	.23
8% overlap	.07	.19	.13	.27
Student-t				
1% overlap	.14	.16	.19	.32
8% overlap	.19	.21	.23	.37

Fuente: Peña, D. y Prieto, J. (2001). *"Cluster Identification using Projections" The Journal of American Statistical Association, 96, 456, 1433-1445, 2001*

Generalización

- Buscar espacios óptimos para encontrar grupos.

Peña, D., Prieto, J. and Viladomat, J. (2010) “Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure” , *Journal of Multivariate Analysis* 9, 1995 -2007, 2010.

- Encontrar grupos en espacios de dimensión reducida por nuevo métodos

Peña, D., Viladomat J. and Zamar, R. (2012) “Nearest-neighbours medians clustering” *Statistical Analysis and Data Mining,* 5*,* 4*,* 349–362, 2012.

Alvarez, A. y Peña, D. (2015). “The SAGRA method for clustering: Splitting and Recombining By Bayes Factors”. Manuscript.

Encontrar Espacios de dimension reducida para cluster

- Encontrar k vectores propios de la matriz de kurtosis

Let X be a multivariate $p \times 1$ random vector, μ its mean vector, Σ its covariance matrix and $Z = \Sigma^{-1/2}(X - \mu)$ the corresponding standardized vector. For $p = 1$ the univariate kurtosis coefficient is $E(z^4)$, where $z = (x - \mu)/\sigma$, and

to Cardoso [2] and Móri et al. [16], who define

$$K = I_p * M_4 = E(Z^{\mathrm{T}} Z Z Z^{\mathrm{T}}),$$

It can be shown that with mixtures of k elliptical distributions the optimal Fisher space is formed by a the set of k eigenvalues of the kurtosis matrix associated to Different eigenvalues. There will be p-k eigenvectors with the same eigenvalue.

- If n/p is large, the Kurtosis matrix is very efficient
- If n/p is small, it is better to compute the directions one by one with a modified Newton algorithm.

Método de Medianas Locales

- Mover los datos hacia su centro, (mediana)
- el número de centros determina el número de grupos.

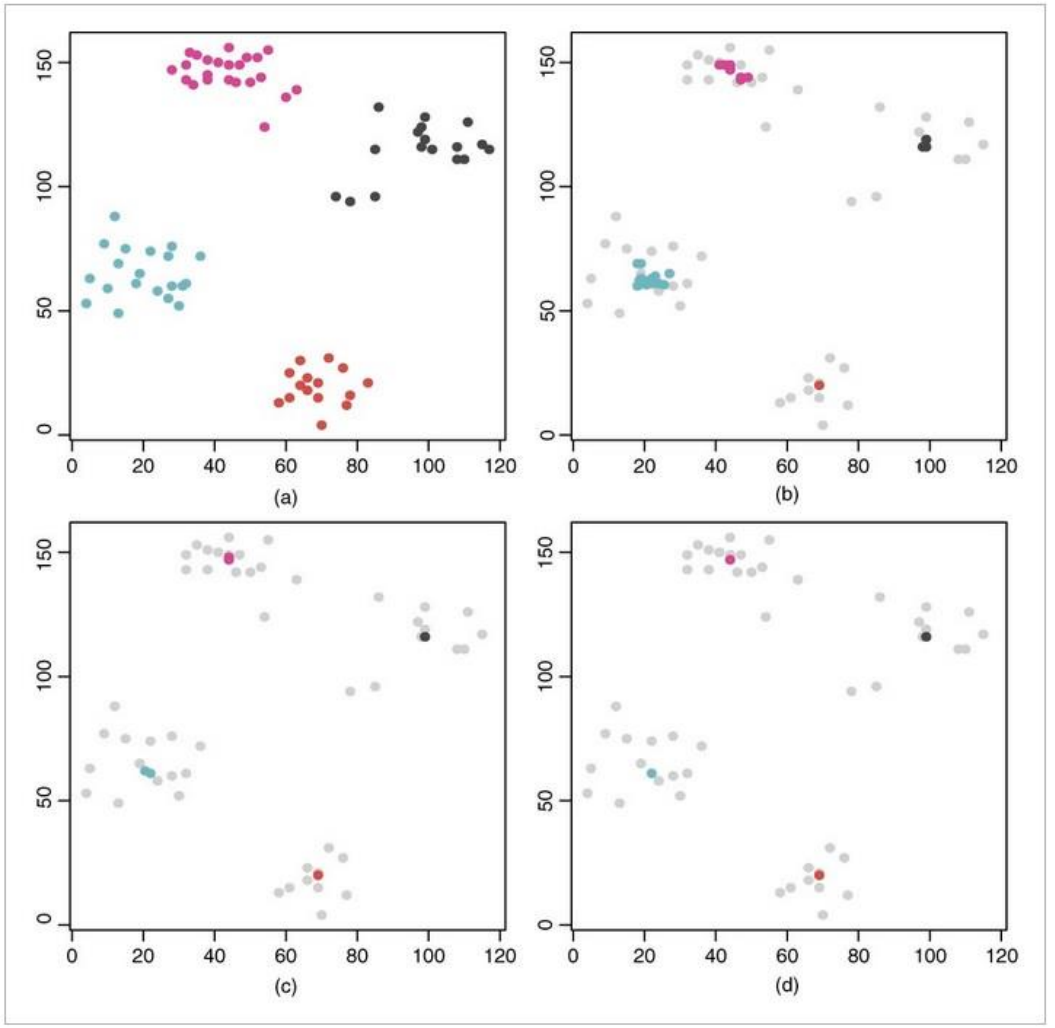


Table 3. Percentage of times the estimated and true number of cluster coincide.

	$p = 4$			$p = 8$			$p = 15$		
Student's-t mixtures	$g = 2$	$g = 4$	$g = 8$	$g = 2$	$g = 4$	$g = 8$	$g = 2$	$g = 4$	$g = 8$
ATTRACTORS	100	92	60	99	91	60	100	82	57
MCLUST	0	0	0	0	0	1	0	0	2
KMEANS	85	18	9	77	37	8	69	28	10
HC_{average}	42	15	15	9	5	1	80	1	0
HC_{complete}	53	22	14	20	8	5	20	6	0
HC_{single}	29	11	9	19	2	2	76	1	2
MEANSHIFT	15	21	14	2	9	2	1	7	1
CLUES CH	67	24	15	70	29	15	62	27	15

Ejemplo 4.3: reducir la dimensión en series temporales

- Con muchas series es imprescindible reducir la dimensión
- Modelos factoriales y componentes principales dinámicos
- Nuevo campo con gran futuro.

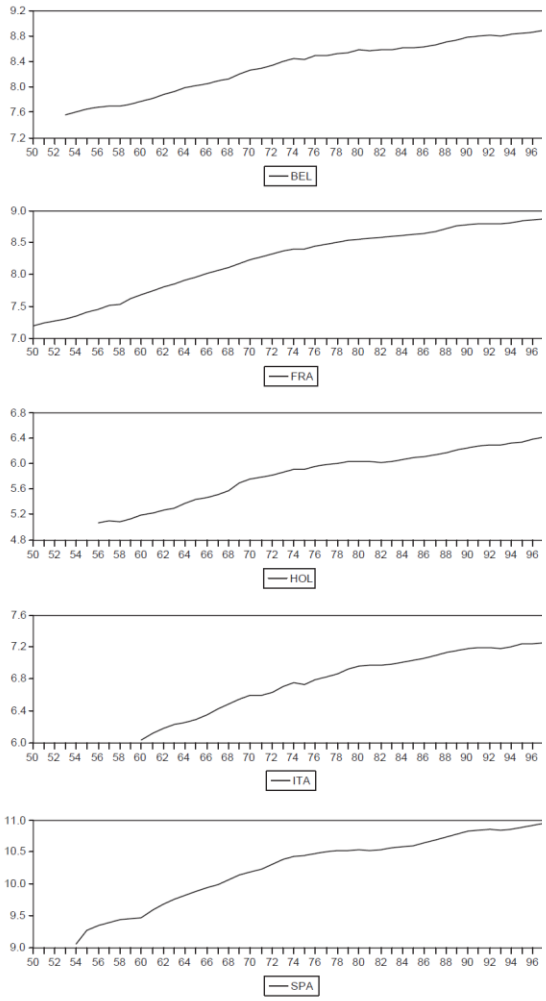


Fig. 1. Logs of real GNP of Belgium, France, Holland, Italy and Spain

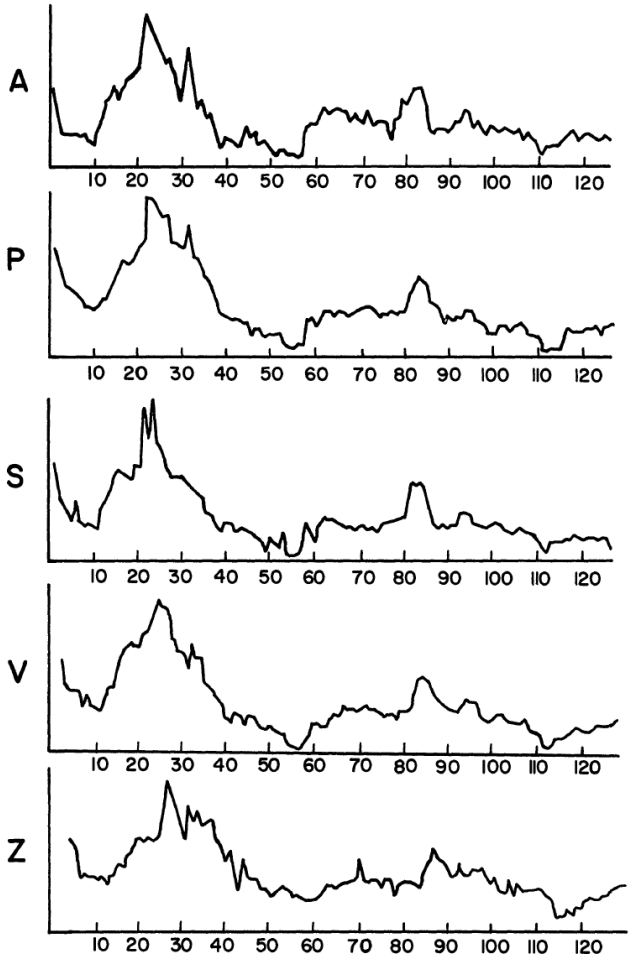


Figure 1. Wheat-Price Series (July 1880–December 1890) of Avila (A), Palencia (P), Segovia (S), Valladolid (V), and Zamora (Z).

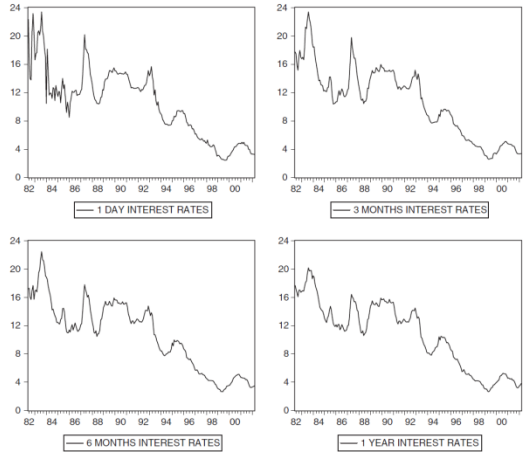


Fig. 1. Graphs of the four series of Spanish interbank interest rates.

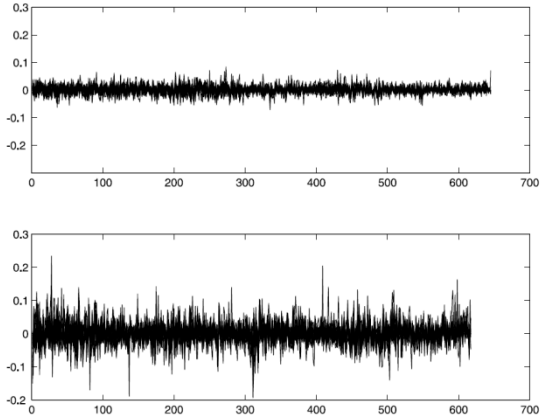


Figure 3. Observations in each of the two groups: A (top) and B (bottom).*

Dimension reduction is very important in vector time series because the number of parameters in a model grows with the square of the dimension m of the vector of time series.

Simplifying structures or *factors* reduce the number of parameters to model the series.

Linear combinations with interesting properties:

Optimal reconstructions of the series Quenouille (1968) PC; Brillinger (1981), DPC; maximun/minimun predictabiliy-canonical analysis (Box and Tiao, 1977); Stationary combinations in nonstationary time series (Granger and Engle, 1987), the scalar component models, SCM, Tiao and Tsay (1989); white noise with stationary data (Ash and Reinsel, 1990), among others.

An alternative approach is Dynamic Factor Models:

$$Z_t = P f_t + e_t$$
$$\varphi(B) f_t = a_t$$

Geweke (1977), Engle and Watson (1981), Peña and Box (1987), Stock and Watson (1988), Peña and Poncela (2006), Forni, Hallin, Lippi, and Reichlin (2000,2005), Forni, Giannone, Lippi, and Reichlin (2009), Lam and Yao (2012).

Note $PF_t = P(B) f_t$;.

- Principal Components for data reduction

Suppose the time series vector $\mathbf{z}_t = (z_{1,t}, ..., z_{m,t})'$

we assume for simplicity that $\overline{\mathbf{z}} = T^{-1} \sum_{t=1}^{T} \mathbf{z}_t$,

the mean if the process is stationary, is zero.

the first principal component, $p_{1,t}, 1 \leq t \leq T$, minimizes the mean squared prediction error of the reconstruction of the vector time series, given by

$$\sum_{j=1}^{m} \sum_{t=1}^{T} (z_{j,t} - \alpha_j p_{1,t})^2$$

and, in general, the first k principal components, $k \leq m$, $p_{1t}, ..., p_{kt}$, $1 \leq t \leq T$,minimize the mean squared prediction error

$$\sum_{j=1}^{m} \sum_{t=1}^{T} (z_{j,t} - \sum_{i=1}^{k} \alpha_{j,i} p_{i,t})^2$$

to reconstruct the vector of time series. Let C be the sample covariance matrix, that is,

$$C = \frac{1}{T} \sum_{t=1}^{T} \mathbf{z}_t \mathbf{z}_t',$$

and let $\lambda_1 \geq \lambda_1 \geq \lambda_m$ be the eigenvalues of C. Then $\alpha_i = (\alpha_{1,i}, ..., \alpha_{m,i})'$, $1 \leq i \leq m$, is the eigenvectors of C corresponding to the eigenvalue λ_i.

Brillinger (1981) addressed this problem as follows. Suppose now the zero mean m dimensional stationary process $\{\mathbf{z}_t\}$, $-\infty < t < \infty$. Then, the dynamic principal components are defined by searching for $m \times 1$ vectors $\mathbf{c}_k$, $-\infty < k < \infty$ and β_j, $-\infty < j < \infty$, so that if we consider as first principal component the linear combination

$$f_t = \sum_{k=-\infty}^{\infty} \mathbf{c}_k' \mathbf{z}_{t-k},$$

then

$$E\left[(\mathbf{z}_t - \sum_{i=-\infty}^{\infty} \beta_j f_{t+j})'(\mathbf{z}_t - \sum_{i=-\infty}^{\infty} \beta_j f_{t+j})\right].$$

is minimum. Brillinger elegantly solved this problem by showing that $\mathbf{c}_k$ is the inverse Fourier transform of the principal components of the spectral matrices for each frequency, and β_j is the inverse Fourier transform of the conjugates of the same principal components. See Brillinger (1981) and Shumway and Stoffer (2000) for the details of the method. Although this result solves the theoretical problem it has the following shortcomings:

Brillinger DPC approach:
It is not appropriate for large m or when T/m is small

1. It can be applied only to stationary series
2. The optimal solution requires the unrealistic assumption that an infinite series is observed, and it is not clear how should be modified when the observed series is finite.
3. It is not easy to see how to robustify these principal components using a reconstruction criterion

Our procedure gives an optimal reconstruction of the vector of time series from a finite number of lags:

(1) the solution can be easily computed even if m is large.

(2) it does not require stationarity.

(3) It does not assume that the DPC is a linear combination of the series.

(4) it can be easily made robust by changing in the minimization criterion of the squared function by a bounded function.

Generalized Dynamic Principal Components (DPC)

Suppose that we observe $z_{j,t}, 1 \leq j \leq m,\ 1 \leq t \leq T,$ and consider two integer numbers $k_1 \geq 0$ and $k_2 \geq 0$. We can define the first dynamic principal component with $k = k_1 + k_2$ lags (first DPC_k) as a vector

$$\mathbf{f} = (f_{1-k_1}, f_{-k_1}, \ldots, f_0, f_1, \ldots, f_t, f_{t+1}, \ldots, f_T, f_{T+1}, \ldots, f_{T+k_2-1}, f_{T+k_2})$$

so that the reconstruction of the series from $\mathbf{f}$ is optimal

Given $\mathbf{f}$, the $m \times (k_1 + k_2)$ matrix of coefficients.

$$\beta = (\beta_{j,i})_{1 \leq j \leq m, -k_1+1 \leq i \leq k_2},$$

and

$$\alpha = (\alpha_1, \ldots, \alpha_m)$$

are used to reconstruct the values $z_{j,t}$ as

$$\widehat{z}_{j,t}(\mathbf{f}, \beta_j, \alpha_j) = \sum_{i=-k_1}^{k_2} \beta_{j,i} f_{t+i} + \alpha_j,$$

where β_j is the j-th row of β.

We can always assume one side lags

$$\widehat{z}_{j,t} = \sum_{i=-k_1}^{k_2} \beta_{j,i} f_{t+i} + \alpha_j.$$

Let $k = k_1 + k_2$ and put

$$f_t^* = f_{t-k_1}, 1 \leq t \leq T + k, \ \beta_{j.,h}^* = \beta_{j,h-k_1}, 0 \leq h \leq k$$

Also define

$$f_t^{**} = f_{t+k}^*, 1 - k \leq t \leq T, \beta_{j.,h}^{**} = \beta_{j,k-h}, 0 \leq h \leq k \tag{3}$$

then, the reconstructed series can also be obtained as

$$\widehat{z}_{j,t} = \sum_{i=-k_1}^{k} \beta_{j,i} f_{t+i+k_1} + \alpha_j = \sum_{h=0}^{k} \beta_{j,h,}^* f_{t+h}^* + \alpha_j = \sum_{h=0}^{k} \beta_{j,h,}^{**} f_{t-h}^{**} + \alpha_j$$

For this reason in the remaining of the paper we will assume without loss of generality $k_1 = 0$ and we will use k to denote the number of forward leads. Once obtained

Consider the loss function

$$MSE(\mathbf{f},\beta,\alpha) = \frac{1}{T}\sum_{j=1}^{m}\sum_{t=1}^{T}(z_{j,t} - \widehat{z}_{j,t}(\mathbf{f},\beta_j,\alpha_j))^2$$

$$= \frac{1}{T}\sum_{j=1}^{m}\sum_{t=1}^{T}(z_{j,t} - \sum_{i=0}^{k}\beta_{j,i+1}f_{t+i} - \alpha_j)^2. \qquad (1)$$

The values of $\mathbf{f} = (f_1, ..., f_{T+k})'$, $\beta = (\beta_{j,i})$ and $\alpha = (\alpha_1, ...\alpha_m)$ which minimize the mean square error, are

$$(\widehat{\mathbf{f}},\widehat{\beta},\widehat{\alpha}) = \arg\min_{\mathbf{f},\beta,\alpha} MSE(\mathbf{f},\beta,\alpha).$$

Clearly if $\mathbf{f}$ is optimal, $\gamma\mathbf{f}+\delta$ is optimal too. Then we can choose $\mathbf{f}$ so that

$$\frac{1}{T+k}\sum_{t=1}^{T+k} f_t^2 = 1.$$

and

$$\frac{1}{T+k}\sum_{t=1}^{T+k} f_t = 0.$$

Then, we call $\widehat{\mathbf{f}}$ the first DPC of order k of the observed series $\mathbf{z}_1, ..., \mathbf{z}_t$.

Note that the first DPC of order 0 corresponds to the first regular principal component of the data.

Moreover the matrix $\widehat{\beta}$ contains the coefficients to be used to reconstruct the m series from $\widehat{\mathbf{f}}$ in an optimal way.

Given $\widehat{\mathbf{f}}$

$$\begin{pmatrix} \widehat{\beta}_j \\ \widehat{\alpha}_j \end{pmatrix} = \left(\mathbf{F}(\widehat{\mathbf{f}})'\mathbf{F}(\widehat{\mathbf{f}})\right)^{-1} \mathbf{F}(\widehat{\mathbf{f}})' \, \mathbf{z}^{(j)},$$

and given $\widehat{\beta}_j$ and $\widehat{\alpha}_j$ we have

$$\mathbf{f} = \mathbf{D}(\mathbf{f},\widehat{\beta})^{-1} \sum_{j=1}^{m} \mathbf{C}_j(\mathbf{f},\widehat{\alpha})\widehat{\beta}_j.$$

The coefficients β_j and α_j, $1 \leq j \leq m$ can be obtained using the least squares estimator, where $\mathbf{z}^{(j)} = (z_{j,1}, ..., z_{j,T})'$ and $\mathbf{F}(\mathbf{f})$ is the $T \times (k+2)$ matrix with t-th row $(f_t, f_{t+1}, ..., f_{t+k}, 1)$.

Iterative algorithm:

step 1 Define $\beta_j^{(h)}$ and $\alpha_j^{(h)}$ by

$$\begin{pmatrix} \beta_j^{(h)} \\ \alpha_j^{(h)} \end{pmatrix} = \left(\mathbf{F}(\mathbf{f}^{(h)})'\ \mathbf{F}(\mathbf{f}^{(h)})\right)^{-1} \mathbf{F}(\mathbf{f}^{(h)})'\mathbf{z}^{(j)}$$

step 2 Then $\mathbf{f}^{(h+1)}$ can be defined by

$$\mathbf{f}^* = \mathbf{D}(\mathbf{f}^{(h)}, \beta^{(h)}, \alpha^{(h)})^{-1} C(\mathbf{f}^{(h)}, \beta^{(h)}, \alpha)\beta^{(h)}$$

and

$$\mathbf{f}^{(h+1)} = (\mathbf{T}+k)^{1/2}(\mathbf{f}^* - \bar{\mathbf{f}}^*)/||\mathbf{f}^* - \bar{\mathbf{f}}^*||.$$

The initial value $\mathbf{f}^{(0)}$ can be chosen equal to the standard (non dynamic) first principal component, completed with k zeros.

The second S -DPC is defined as the first S-DPC of the residuals $r_{j,t}(\mathbf{f}, \beta)$.

Higher order S-DPC are defined in a similar manner.

Note:

D$j(\beta j)$ is $(T + k)$x$(T + k)$, and **C**$j(\alpha j)$ is $(T + k)$x$(k + 1)$

F(**f**) is the Tx$(k + 2)$

Remark 1. Note that the dimension of the matrices to be inverted to compute $\mathbf{f}^{(h)}, \beta^{(h)}, \alpha^{(h)}$ are independent of the number of time series and therefore we can deal with large number of variables.

Remark 2. Note also that there are no restrictions on the values $\mathbf{f}$ and in particular we do not assume, as in Brillinger, that they must be linear combinations of the series. In this way the values of $\mathbf{f}$ can be adapted to the nonstationarity character of the time series.

3 Dynamic Principal Components when $k = 1$

To illustrate the computation of the first DPC, let us consider the simplest case of $k = 1$. Then, we search for $\widehat{\beta}$ and $\widehat{\mathbf{f}} = (\widehat{f}_1, ..., \widehat{f}_{T+1})'$ such that

$$(\widehat{\mathbf{f}}, \widehat{\beta}) = \arg\min_{1} \sum_{t=1}^{T} \sum_{j=1}^{m} (z_{j,t} - \beta_{j,1} f_t - \beta_{j,2} f_{t+1})^2.$$

It can be shown that with $0 \leq c < 1$

$$\widehat{f}_t = \frac{1}{\delta} \left[\sum_{j=1}^{m} \widehat{\beta}_{j,1} \sum_{q=1}^{T} c^{|t-q|} z_{j,q} + \sum_{j=1}^{m} \widehat{\beta}_{j,2} \sum_{q=2}^{T+1} c^{|t-q|} z_{j,q-1} \right] + R_t,$$

where $R_t \to 0$ except for t close to 1 or close to T.

The DPC is computed as a weighted average of the observations by a two side moving average. For $\widehat{f}_t$ the maximum weight is given to $z_{j,t}$ and $z_{j,t-1}$.

Suppose now that $\mathbf{z}_t$ is stationary, then except in both ends $\widehat{f}_t$ can be approximated by the stationary process

$$\widehat{f}_t^* = \frac{1}{\alpha}\left[\sum_{j=1}^{m}\widehat{\beta}_{j,1}\sum_{q=-\infty}^{\infty}c^{|t-q|}z_{j,q} + \sum_{j=1}^{m}\widehat{\beta}_{j,2}\sum_{q=-\infty}^{\infty}c^{|t-q|}z_{j,q-1}\right].$$

The DPC is approximated as linear combinations of the geometrically and symmetrically filtered series

$$z_{j,t} + \sum_{i=1}^{\infty}c^i(z_{j,t+i} + z_{j,t-i}), 1 \leq j \leq m$$

and

$$z_{j,t-1} + \sum_{i=1}^{\infty}c^i(z_{j,t-1+i} + z_{j,t-1-i}), 1 \leq j \leq m$$

This series give the largest weight to the periods t and $t-1$ respectively and the weights decrease geometrically when we move away of these values.

We conjecture that in the case of the first DPC of order k, a similar approximation outside both ends of $\widehat{f}_t$ by an stationary process can be obtained.

Monte Carlo results

We perform a Monte Carlo study using as vector series $\mathbf{z}_t = (z_{1,t}, z_{2,t}, ..., z_{m,t})'$, $1 \leq t \leq T$ generated as follows:

$$z_{i,t} = 10\sin(2pi(i/m))f_t + 10\cos(2pi(i/m))f_{t-1} + 10(i/m)f_{t-2} + u_{i,t}, 1 \leq i \leq m, 1 \leq t \leq T, \quad (12)$$

where $f_t, -2 \leq t \leq T$ and $u_{i,t}, \leq t \leq T$, $1 \leq i \leq m$ are i.i.d. random variables with distribution $\mathrm{N}(0,1)$. We compute three different principal components: (i) The ordinary principal component used in a dynamic way with k lags to reconstruct the original series (OPC_k) (ii) the dynamic principal component (DPC_k) proposed here, (iii) Brillinger dynamic principal components (BDPC_k) adapted for finite samples as follows:

m	T	OPC_2	DPC_2	BDPC_{10}
20	100	52.53	0.91	0.94
	200	55.86	0.92	0.95
100	100	54.89	0.95	0.99
	200	57.65	0.97	0.99
500	100	53.56	0.96	1.00
	200	57.14	0.98	1.00
1000	100	54.88	0.96	-
	200	59.09	1.00	-

Table 1: MSE of the Reconstructed Series for the Stationary Model with one Factor

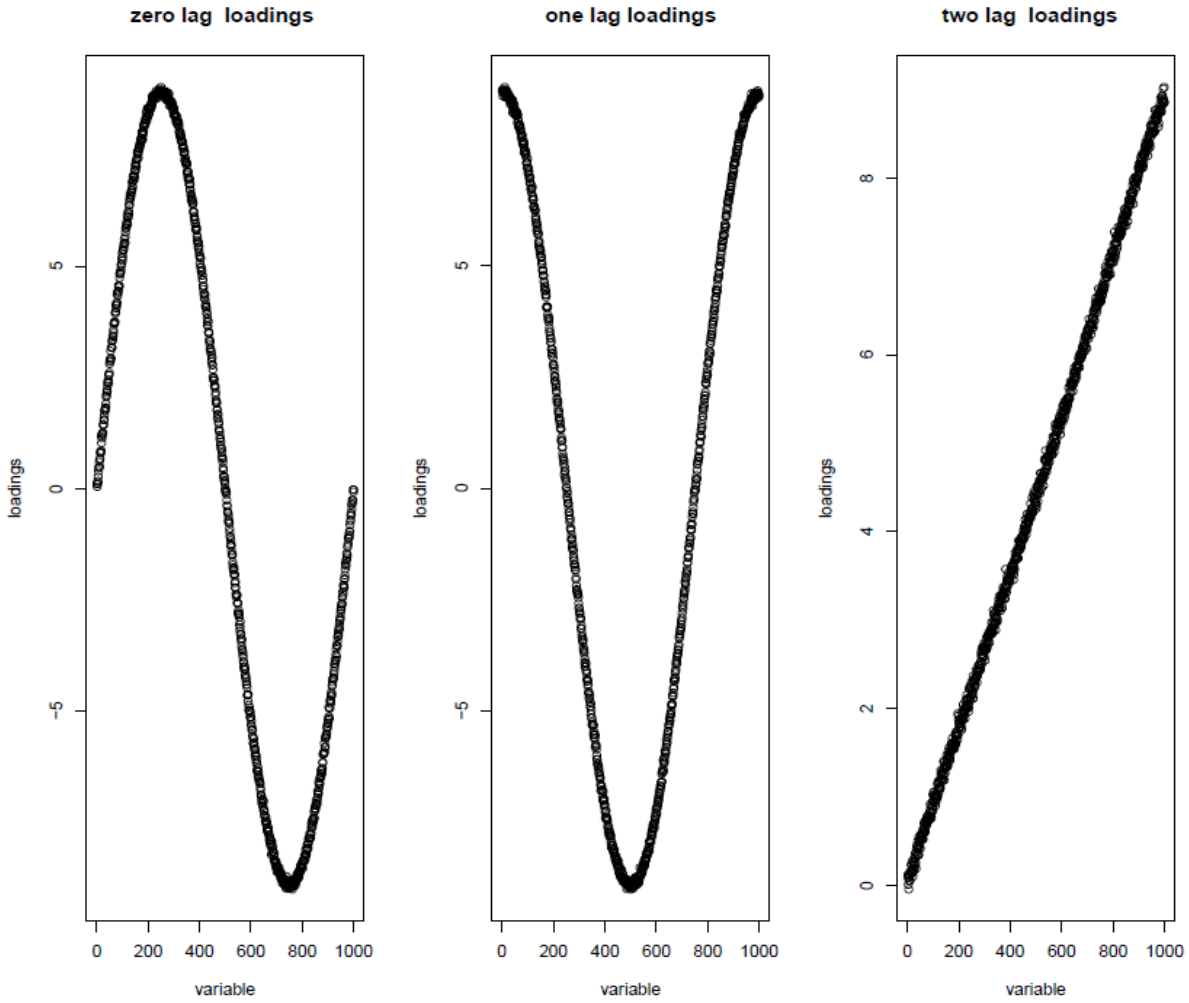


Figure 1: Loadings for one Replication of the Stationary Model with T=200 and m=1000

In this case we consider a VARI$(1,1)$ $m-$dimensional vector series $\mathbf{z}_t$ generated as follows. Consider an stationary VAR(1) model $\mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{u}_t, 1 \leq t \leq T$, where the $\mathbf{u}_t$s are i.i.d. m-dimensional vectors with distribution $\mathrm{N}_m(\mathbf{0}, \mathbf{I})$ and let $\mathbf{z}_t = \mathbf{z}_{t-1} + \mathbf{x}_t$. We consider 1000 replications and in each replication we generate a new matrix A of the form $A = V\Lambda V'$, where V is an orthogonal matrix generated at random with uniform distribution and Λ is a diagonal matrix, where the diagonal elements are independent with uniform distribution in the interval $[0, 0.9]$.

m	$\mathbf{OPC}_{10}$	$\mathbf{DPC}_{10}$	BDPC_{10}
20	67	83	55
100	67	86	62
200	69	86	62

Table 2: Percentage of Explained Variance in the VARI(1,1) Model

Conclusión

- Big Data va a transformar la Estadística y hacerla más convergente con las ciencias de la computación.
- Se necesitan profesionales con buena base estadística e interés en computación
- La formación conjunta Stat/CompS será clave en Big Data y muy demandada.

Referencias

- Delicado, P. (2014) A course in Big Data. UPC
- Fan, J., F. Han, and H. Liu (2014). Challenges of big data analysis. National Science Review, nwt032
- Hand, D. J. (2013). Data, not dogma: Big data, open data, and the opportunities ahead. In Advances in Intelligent Data Analysis XII, pp. 1{12. Springer
- Han, J. Kamber,M. and Pe, J. (2012) Data Mining: Concepts and Techniques
- Hastie, T., Tibshirani, R. and Friedman, J. (2011) *The Elements of Statistical Learning:Data Mining, Inference, and Prediction,* Scnd Edition. Springer Series in Statistics.
- Hilbert, M. and López, P. (2011) The world's technological capacity to store, communicate, and compute information, *Science,* February 10.
- Jordan, M. I. (2013). On statistics, computation and scalability. Bernoulli 19(4), 1378{1390.
- Mayer-Schönberger, V. and Cukier, K.(2013). Big Data: A Revolution That Will Transform How We Live, Work and Think, John Murray (Publishers).
- O'Neil, C. and Schutt, R. (2013) Doing Data Science: Straight Talk from the Frontline, O'Really Media Inc
- Peña D. and Prieto, J. (2001). Cluster identification using Projections. *Journal of American Statistical Association*, 96, 1433-1445.
- Peña D., Prieto, J. and Viladomat, J. (2010). Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure. *Journal of Multivariate Analysis* 101, 9, 1995 -2007.
- Peña D., Yohai, V. (2016). Generalized Dynamic Principal Components. *Journal of American Statistical Association (in press)*
- Provost, F. and Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking, O'Really Media Inc.
- Witten, H. And Frank, E. (2012). Data Mining: Practical Machine Learning Tools and Techniques, Second Edition
- .