DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

Clustering en subespacios

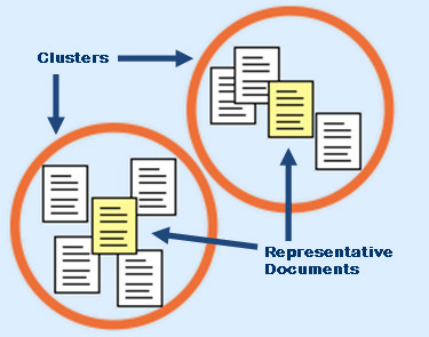
© Fernando Berzal, berzal@acm.org

Clustering en subespacios

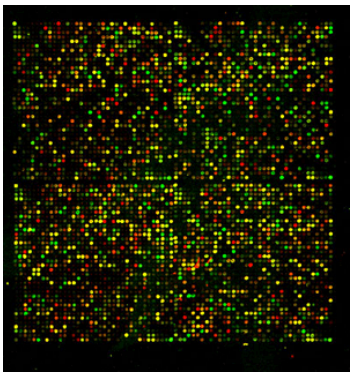
- El problema de la dimensionalidad
- Subspace Clustering
- CLIQUE [CLustering In QUEst]
- Projected Clustering: PROCLUS & ORCLUS
- SUBCLU
- Robust Subspace Clustering
- Apéndice
  Reducción de la dimensionalidad (PCA & SVD)

La dimensionalidad de los datos

En muchas aplicaciones,
la dimensionalidad de los datos es elevada.



Clustering de documentos

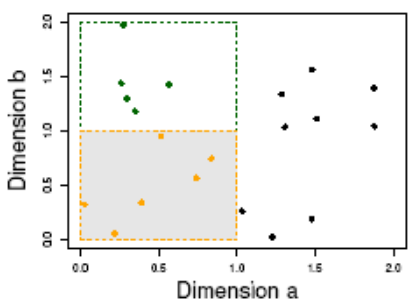


DNA Microarrays

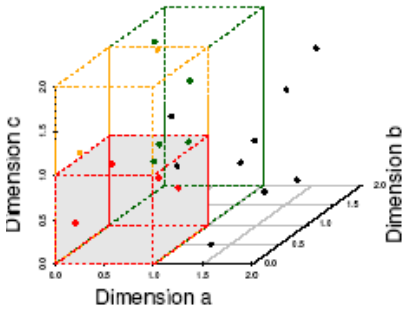
La dimensionalidad de los datos

¿Por qué es un problema?

- Los datos en una dimensión están relativamente cerca
- Al añadir una nueva dimensión, los datos se alejan.
- Cuando tenemos muchas dimensiones, las medidas de distancia dejan de ser útiles ("equidistancia").



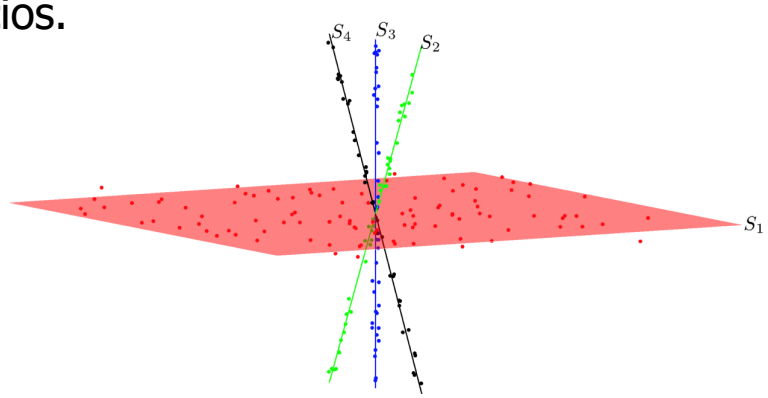
(b) 6 Objects in One Unit Bin



(c) 4 Objects in One Unit Bin

La dimensionalidad de los datos

- La existencia de dimensiones irrelevantes puede enmascarar la presencia de clusters en el conjunto de datos.
- Los clusters puede que existan sólo en algunos subespacios.



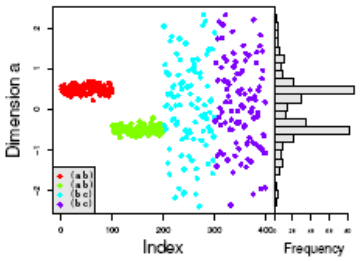
La dimensionalidad de los datos

Posibles soluciones

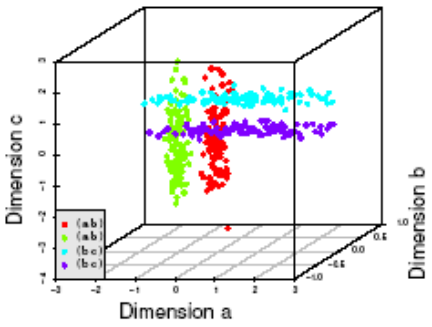
- **Transformación de características** (PCA, SVD) para reducir la dimensionalidad de los datos, útil sólo si existe correlación/redundancia.
- **Selección de características** (wrapper/filter) útil si se pueden encontrar clusters en subespacios.
- **"Subspace clustering"** Buscar clusters usando distintas combinaciones de atributos, vg. CLIQUE o PROCLUS.

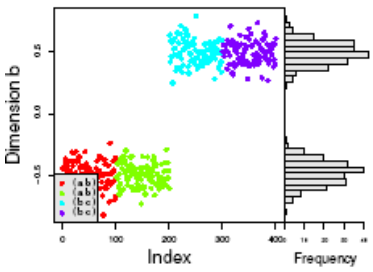
Subspace clustering

Proyecciones 1D

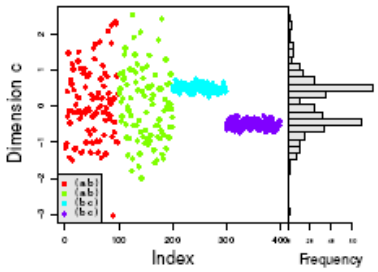


(a) Dimension a





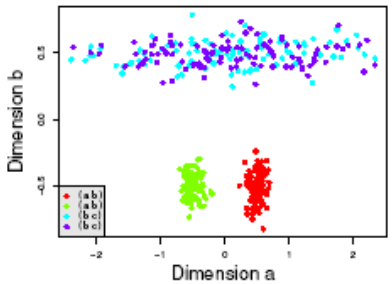
(b) Dimension b



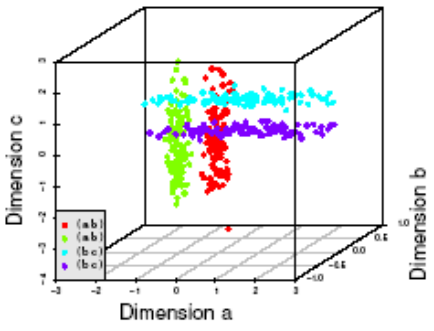
(c) Dimension c

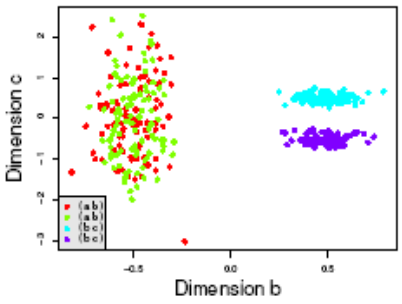
Subspace clustering

Proyecciones 2D

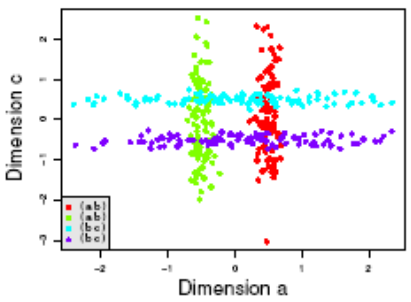


(a) Dims a & b





(b) Dims b & c



(c) Dims a & c

CLIQUE

CLIQUE = CLustering In QUEst

Agrawal, Gehrke, Gunopulos & Raghavan (SIGMOD'98)

QUEST:

Proyecto original de minería de datos en IBM (1996-)

IBM Data Mining Research

The Quest data mining project at the IBM Almaden Research Center has developed innovative technologies to discover useful patterns in large databases. Our technolgies include mining for association rules, sequential patterns, classification, and time-series clustering. IBM is making these technologies available through its data mining product, IBM Intelligent Mine

CLIQUE

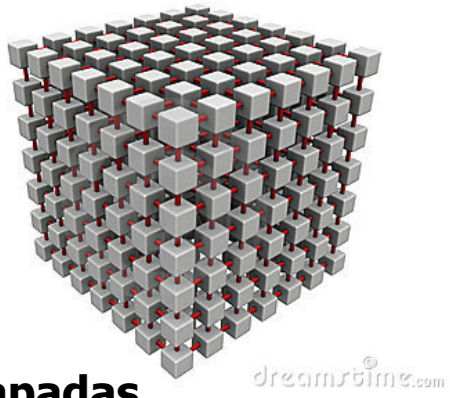
Objetivos

- Identificar automáticamente subspacios en un espacio de alta dimensionalidad de forma que se puedan obtener mejores clusters que en el espacio original.
- Mejorar la interpretabilidad de los resultados proporcionando una descripción comprensible de los resultados del algoritmo de clustering
- Obtener un método escalable (con respecto al tamaño del conjunto de datos y al número de dimensiones del problema).

CLIQUE

CLIQUE puede interpretarse como un método de agrupamiento basado en grids:

- Se divide cada dimensión en el mismo número de intervalos (de la misma longitud, i.e. "equiwidth partitioning").
- Se particiona un espacio m-dimensional en un conjunto de unidades rectangulares **no solapadas.**



CLIQUE

CLIQUE puede considerarse un método basado en densidad:

- Una unidad se considera densa si la fracción de puntos contenida en ella excede un parámetro del modelo.
- Un cluster es un conjunto maximal de unidades densas conectadas en un subespacio determinado.

CLIQUE

Etapas del algoritmo

- Particionar el espacio de datos y calcular el número de puntos que quedan dentro de cada celda de la partición.
- Identificar los subespacios que contienen clusters utilizando el principio Apriori.
- Identificar los clusters.
- Generar una descripción mínima de los clusters.

CLIQUE

Identificación de subespacios con clusters

Propiedad (monotonicidad): Si una unidad k-dimensional es densa, también lo son todas sus proyecciones en un espacio (k-1)-dimensional.

Algoritmo ascendente (tipo Apriori):

- Se determinan las unidades densas unidimensionales.
- Se obtienen las unidades densas k-dimensionales combinando unidades densas (k-1)-dimensionales.
  - Podemos descartar las unidades k-dimensionales candidatas que tienen proyecciones no densas en (k-1) dimensiones.

CLIQUE

Identificación de los clusters

Se determinan los conjuntos de unidades densas conectadas en todos los subespacios de interés:

- A partir del conjunto D de unidades densas de cada subespacio k-dimensional, se crea una partición con los conjuntos conexos de unidades densas
- Algoritmo: DFS (búsqueda en profundidad sobre el grafo de unidades densas)

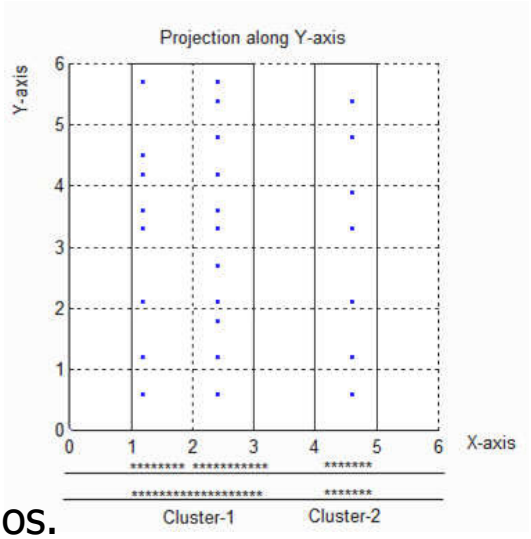
CLIQUE

Ejemplo

Grid 6x6

Umbral de soporte = 3

- Ninguna celda es densa.
- Si proyectamos sobre el eje X, hay 3 unidades densas.
- Dos de estas unidades están conectadas, por lo que las unimos.
- Se obtienen dos clusters



CLIQUE

Descripción de los clusters

- Se determinan las regiones maximales que cubren cada cluster de unidades densas conexas.
- Se determina un recubrimiento mínimo para los conjuntos de regiones asociados a cada cluster.
- Se construye una expresión DNF para cada cluster identificado.

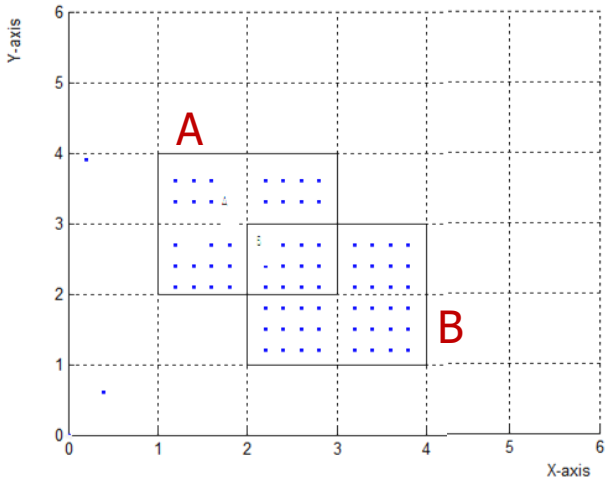
CLIQUE

Ejemplo

Grid 6x6

Umbral de soporte = 3

- A U B es un cluster.
- A o B son las regiones maximales del cluster ($A \cap B$ no lo es).



- La descripción mínima de este cluster en DNF es $((2 \le x \le 4) \wedge (1 \le y \le 3)) \vee ((1 \le x \le 3) \wedge (2 \le y \le 4))$

CLIQUE

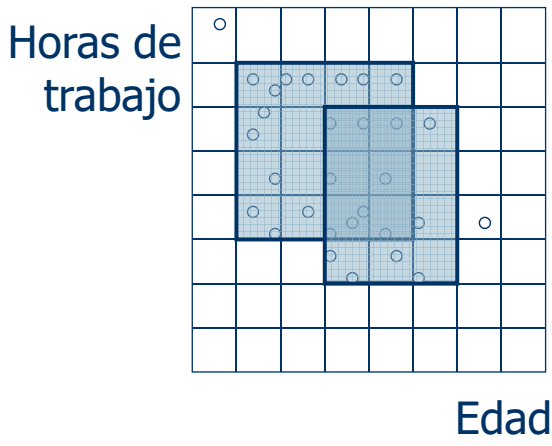
Ventajas

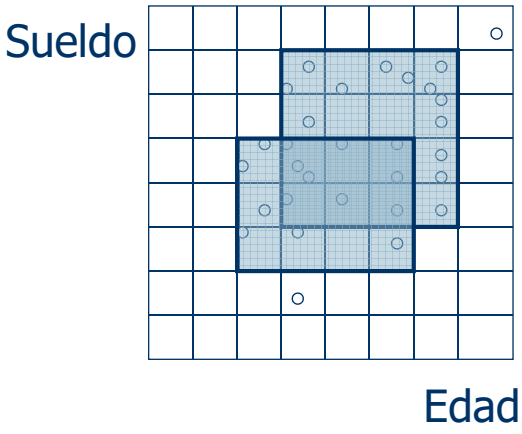
- Encuenta automáticamente los subespacios de máxima dimensionalidad en los que existen clusters de alta densidad.
- No depende del orden de presentación de los datos ni presupone ninguna distribución de datos.
- Escala linealmente con el tamaño de la entrada.

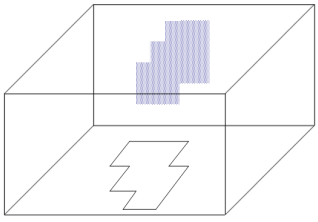
Limitación

- La exactitud de los resultados del clustering puede verse degradada por la simplicidad del método.

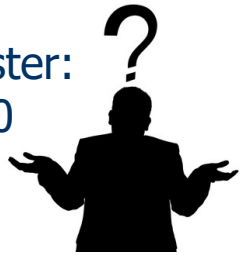
CLIQUE





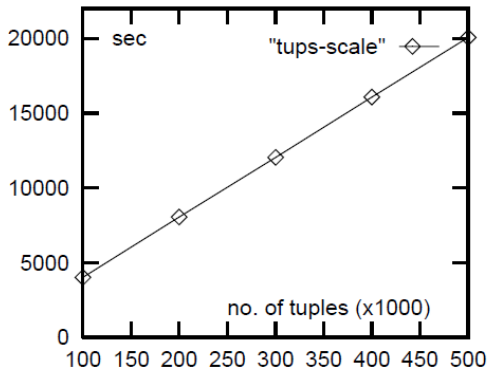


Descripción del cluster:
20 <= Edad <= 60



CLIQUE

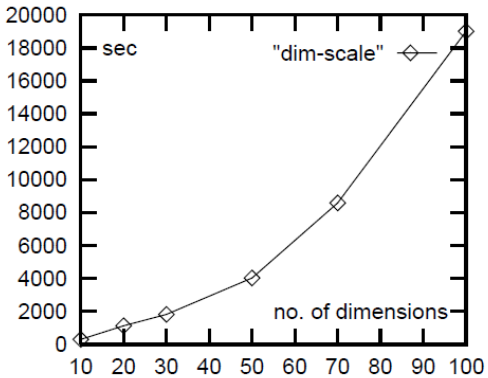
CLIQUE escala linealmente con respecto al tamaño del conjunto de datos (en cuanto a su número de tuplas)



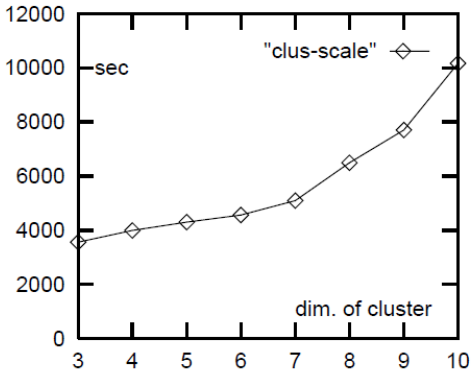
Escalabilidad con el número de tuplas

CLIQUE

CLIQUE tiene una "buena" escalabilidad conforme aumenta el número de dimensiones (del conjunto de datos y de los clusters ocultos).



Dimensiones del conjunto de datos



Dimensiones de los clusters ocultos

CLIQUE

DEMO

http://www.cs.ualberta.ca/~yaling/Cluster/Applet/Code/Cluster.html

Projected Clustering

PROCLUS [PROjected CLUStering]

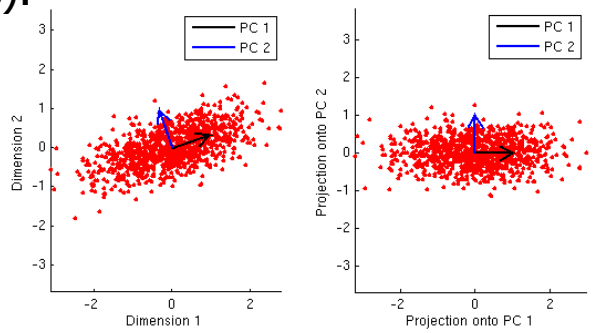
Aggarwal, Wolf, Yu, Procopiuc & Park (SIGMOD'1999)

- Algoritmo iterativo (tipo K-means o K-medoids)
- Enfoque descendente:
  Elimina las dimensiones
  menos relevantes de cada uno de los clusters.
- **No** es un algoritmo de agrupamiento basado en densidad y, de forma similar al algoritmo de las k medias, sólo es capaz de encontrar clusters globulares (esto es, más o menos esféricos)

Projected Clustering

ORCLUS: Generalized Projected Clustering
Aggarwal & Yu (SIGMOD'2000)

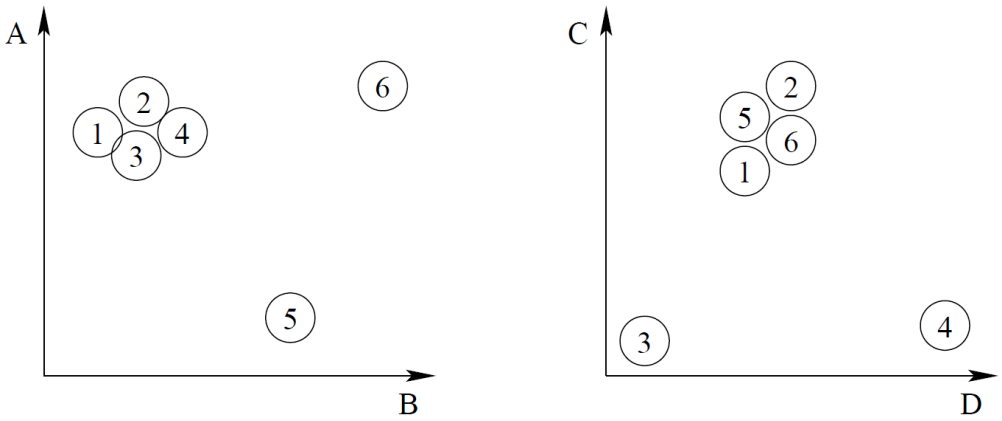
- Añade un proceso de mezcla de clusters.
- Selecciona componentes principales (en lugar de atributos).



Projected Clustering

Limitaciones de PROCLUS

En distintos subespacios,
los objetos se agrupan de diferentes formas:



SUBCLU

Density-connected SUBspace CLUstering

Kailing, Kriegel & Kröger (SDM'2004)

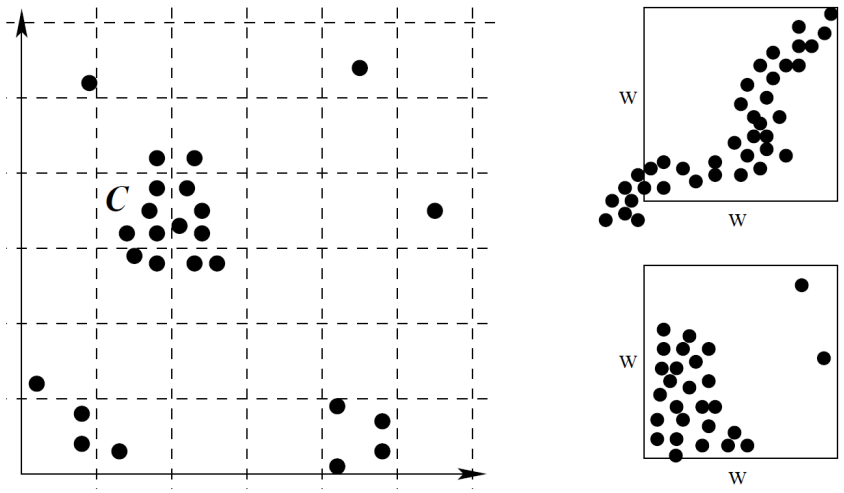
- En vez de utilizar un grid predefinico, como CLIQUE o PROCLUS, es capaz de detectar clusters de formas de formas arbitrarias en cualquier posición del espacio.

- Utiliza un enfoque ascendente (como CLIQUE) y emplea la monotonicidad de las regiones densas conectadas para podar subespacios en el proceso de generar todos los clusters.

SUBCLU

Limitaciones de CLIQUE, ENCLUS, MAFIA, DOC…

Los resultados dependen de la posición del grid



SUBCLU

Características

- SUBCLU es capaz de detectar clusters de formas arbitrarias en cualquier posición de un subespacio.
- A diferencia de CLIQUE y sus sucesores, no depende del uso de un grid ("la noción de cluster está bien definida").
- Dado que no utiliza técnicas heurísticas de poda de subespacios (CLIQUE lo hace por cuestiones de eficiencia), SUBCLU obtiene los mismos resultados que DBSCAN para cada subespacio.

SUBCLU

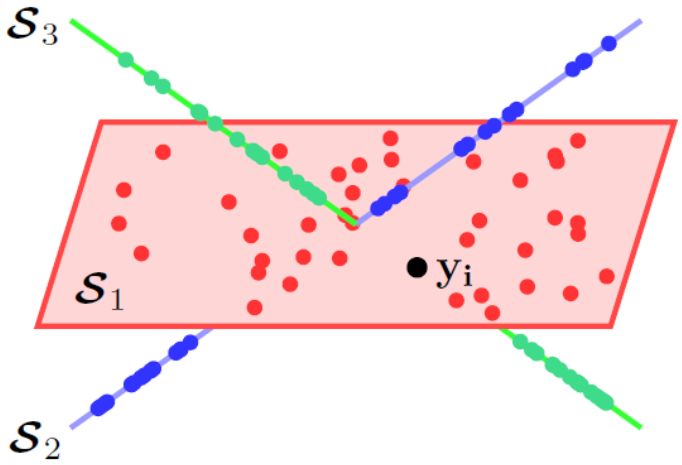
SUBCLU(SetOfObjects DB, Real ε, Integer m)

/* **STEP 1 Generate all 1-D clusters** */
$\mathcal{S}_1 := \emptyset$ // set of 1-D subspaces containing clusters
$\mathcal{C}_1 := \emptyset$ // set of all sets of clusters in 1-D subspaces
FOR each $a_i \in \mathcal{A}$ DO
  $\mathcal{C}^{\{a_i\}} := DBSCAN(DB, \{a_i\}, \varepsilon, m)$ // set of all clusters in subspace a_i;
  IF $\mathcal{C}^{\{a_i\}} \neq \emptyset$ THEN // at least one cluster in subspace $\{a_i\}$ found
    $\mathcal{S}_1 := \mathcal{S}_1 \cup \{a_i\}$;
    $\mathcal{C}_1 := \mathcal{C}_1 \cup \mathcal{C}^{\{a_i\}}$;
  END IF
END FOR

/* **STEP 2 Generate $(k+1)$-D clusters from k-D clusters** */
$k := 1$;
WHILE $\mathcal{C}_k \neq \emptyset$
  /* **STEP 2.1 Generate $(k+1)$-dimensional candidate subspaces** */
  $CandS_{k+1}$:= GenerateCandidateSubspaces(S_k);
  /* **STEP 2.2 Test candidates and generate $(k+1)$-dimensional clusters** */
  FOR EACH $cand \in CandS_{k+1}$ DO
    // Search k-dim subspace of $cand$ with minimal number of objects in the clusters
    $bestSubspace := \min_{s \in \mathcal{S}_k \wedge s \subseteq cand} \sum_{C_i \in \mathcal{C}^s} |C_i|$
    $\mathcal{C}^{cand} := \emptyset$;
    FOR EACH cluster $cl \in \mathcal{C}^{bestSubspace}$ DO
      $\mathcal{C}^{cand} = \mathcal{C}^{cand} \cup DBSCAN(cl, cand, \varepsilon, m)$;
      IF $\mathcal{C}^{cand} \neq \emptyset$ THEN
        $\mathcal{S}_{k+1} := \mathcal{S}_{k+1} \cup cand$;
        $\mathcal{C}_{k+1} := \mathcal{C}_{k+1} \cup \mathcal{C}^{cand}$;
      END IF
    END FOR
  END FOR
  $k := k + 1$
END WHILE

Robust Subspace Clustering

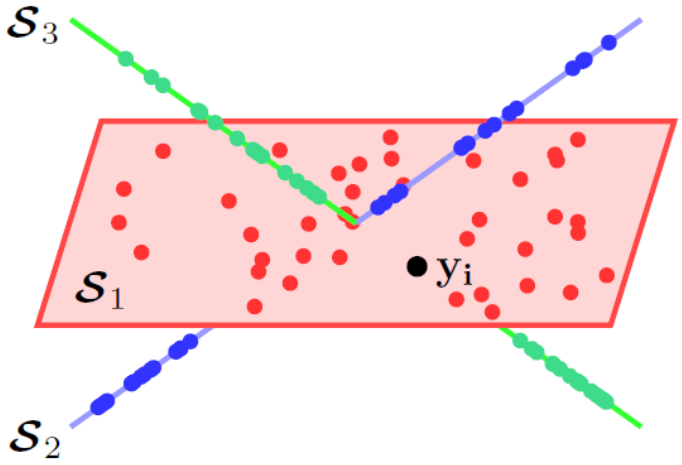
Sparse Subspace Clustering



OBSERVACIÓN: Cada punto en una unión de subespacios tiene una representación "sparse" con respecto a un diccionario formado por todos los demás puntos.

Robust Subspace Clustering

Sparse Subspace Clustering

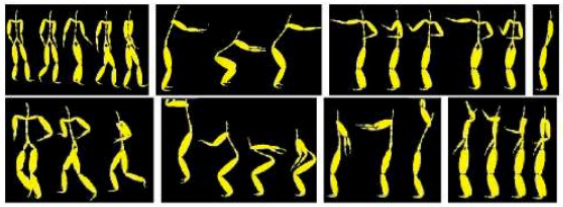


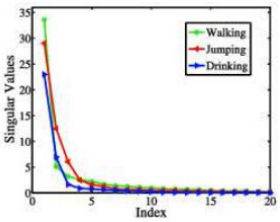
Dicha representación puede obtenerse resolviendo un problema de optimización…

Robust Subspace Clustering

SSC utiliza técnicas de **clustering espectral** (utilizando el espectro de la matriz de similitud para reducir la dimensionalidad de los datos antes de agrupar en un menor número de dimensiones) y **análisis de componentes principales** [PCA] en cada uno de los clusters obtenidos.

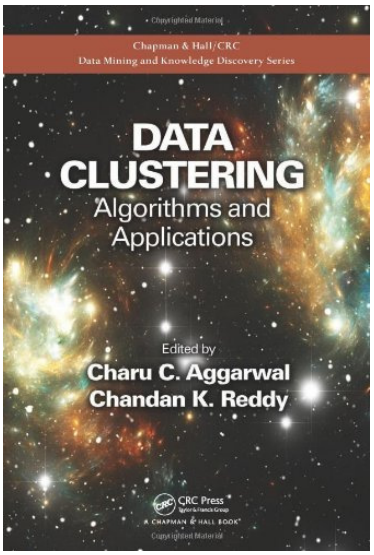
RSC está inspirado en SSC y propone un modelo estadístico para datos con ruido (medidas de similitud "confiables").





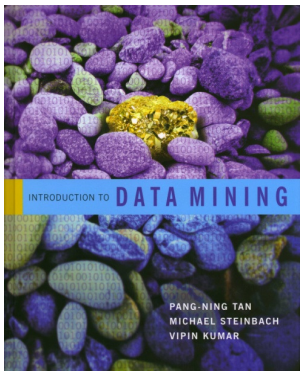
Bibliografía

- Charu C. Aggarwal & Chandan K. Reddy (editors): **Data Clustering: Algorithms and Applications**. Chapman & Hall / CRC Press, 2014. ISBN 1466558210.

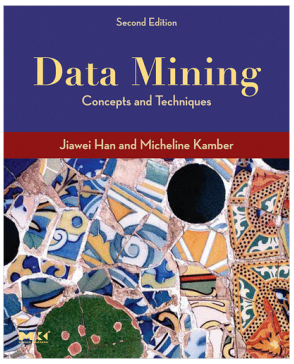


Bibliografía

- Pang-Ning Tan, Michael Steinbach & Vipin Kumar: **Introduction to Data Mining** Addison-Wesley, 2006. ISBN 0321321367 [capítulos 8&9]



- Jiawei Han & Micheline Kamber: **Data Mining: Concepts and Techniques** Morgan Kaufmann, 2006. ISBN 1558609016 [capítulo 7]



Bibliografía - Algoritmos

CLIQUE

- Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos & Prabhakar Raghavan: **Automatic subspace clustering of high dimensional data for data mining applications**. In Proceedings of the 1998 ACM SIGMOD international conference on Management of Data (SIGMOD '98), pp. 94-105. DOI 10.1145/276304.276314
- Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos & Prabhakar Raghavan: **Automatic Subspace Clustering of High Dimensional Data**. Data Mining and Knowledge Discovery 11(1):5-33, July 2005. DOI 10.1007/s10618-005-1396-1

PROCLUS

- Charu C. Aggarwal, Joel L. Wolf, Philip S. Yu, Cecilia Procopiuc & Jong Soo Park: **Fast algorithms for projected clustering**. In Proceedings of the 1999 ACM SIGMOD international conference on Management of Data (SIGMOD '99), pp. 61-72. DOI 10.1145/304182.304188

ORCLUS

- Charu C. Aggarwal & Philip S. Yu: **Finding generalized projected clusters in high dimensional spaces**. In Proceedings of the 2000 ACM SIGMOD international conference on Management of Data (SIGMOD '00), pp. 70-81. DOI 10.1145/342009.335383
- Charu C. Aggarwal & Philip S. Yu: **Redefining Clustering for High-Dimensional Applications**. IEEE Transactions on Knowledge and Data Engineering 14(2):210-225, March 2002. DOI 10.1109/69.991713

Bibliografía - Algoritmos

ENCLUS [ENtropy-based CLUStering]

- Chun-Hung Cheng, Ada Waichee Fu & Yi Zhang: **Entropy-based subspace clustering for mining numerical data**. In Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '99), pages 84-93. DOI 10.1145/312129.312199

MAFIA [Merging of Adaptive Finite IntervAls]

- Harsha S. Nagesh, Sanjay Goil & Alok S. Choudhary: **Adaptive grids for clustering massive data sets**. In Proceedings of the 1st SIAM International Conference on Data Mining (SDM'2001), pages 1-17. DOI 10.1137/1.9781611972719.7

DOC [Density-based Optimal projected Clustering]

- Cecilia M. Procopiuc, Michael Jones, Pankaj K. Agarwal & T. M. Murali: **A Monte Carlo algorithm for fast projective clustering**. In Proceedings of the 2002 ACM SIGMOD international conference on Management of data (SIGMOD '02), pp. 418-427. DOI 10.1145/564691.564739

SUBCLU [density-connected SUBspace CLUstering]

- Karin Kailing, Hans-Peter Kriegel & Peer Kröge: Density-Connected Subspace Clustering for High-Dimensional Data. In Proceedings of the 4th SIAM International Conference on Data Mining (SDM'2004), pages 246-256. DOI 10.1137/1.9781611972740.23

Bibliografía

Robust Subspace Clustering

- Mahdi Soltanolkotabi & Emmanuel J.A. Candés: **A geometric analysis of subspace clustering with outliers**. The Annals of Statistics 40(4):2195-2238, 2012. DOI 10.1214/12-AOS1034
- Mahdi Soltanolkotabi, Ehsan Elhamifar & Emmanuel J.A. Candés: **Robust subspace clustering**. The Annals of Statistics 42(2):669-699, 2014. DOI 10.1214/13-AOS1199

Otros métodos relacionados

- Wei Wang, Jiong Yang & Richard R. Muntz: **STING: A Statistical Information grid Approach to Spatial Data Mining**, Proceedings of the 23rd International Conference on Very Large Data Bases (VLDB'97), pp. 186-195. ISBN 1-55860-470-7
- Gholamhosein Sheikholeslami, Surojit Chatterjee & Aidong Zhang: **WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases**. In Proceedings of the 24rd International Conference on Very Large Data Bases (VLDB '98), pp. 428-439. ISBN 1-55860-566-5

Bibliografía

Surveys

- Lance Parsons, Ehtesham Haque & Huan Liu: **Subspace clustering for high dimensional data: a review**. SIGKDD Explorations Newsletter 6(1):90-105, June 2004. DOI 10.1145/1007730.1007731L
- Hans-Peter Kriegel, Peer Kröger & Arthur Zimek: **Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering**. ACM Transactions on Knowledge Discovery from Data 3, 1, Article 1, March 2009, 58 pages. DOI 10.1145/1497577.1497578
- Kelvin Sim, Vivekanand Gopalkrishnan, Arthur Zimek & Gao Cong: **A survey on enhanced subspace clustering**. Data Mining and Knowledge Discovery 26(2):332-397, March 2013. DOI 10.1007/s10618-012-0258-x

Apéndice
Notación O

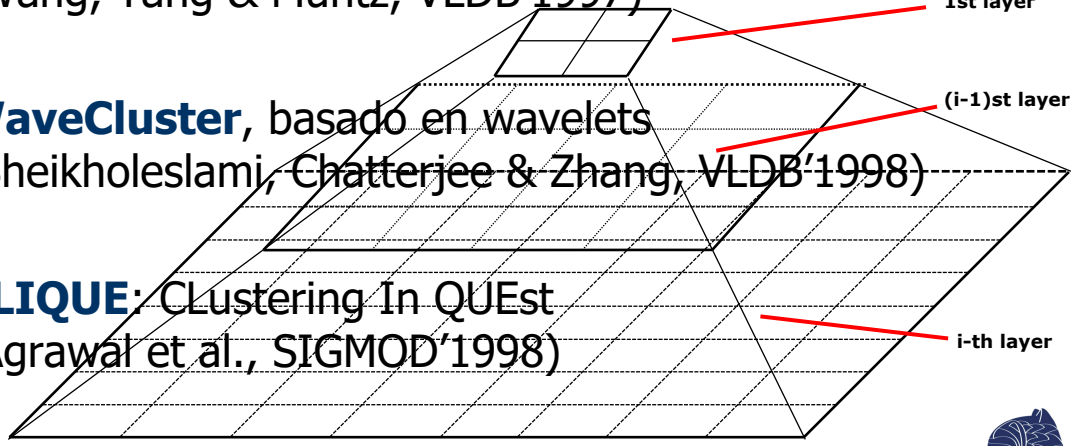
El impacto de la eficiencia de un algoritmo…

n	10	100	1000	10000	100000
O(n)	10ms	0.1s	1s	10s	100s
O(n·log$_2$ n)	33ms	0.7s	10s	2 min	28 min
O(n^2)	100ms	10s	17 min	28 horas	115 días
O(n^3)	1s	17min	12 días	31 años	32 milenios

Apéndice
Otros métodos de clustering

Grids multiresolución

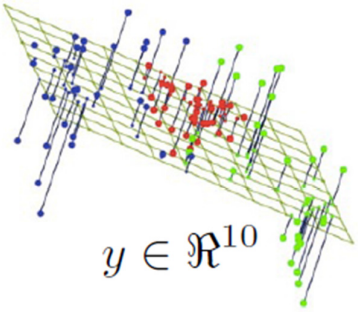
- **STING**, a STatistical INformation Grid approach (Wang, Yang & Muntz, VLDB'1997)
- **WaveCluster**, basado en wavelets (Sheikholeslami, Chatterjee & Zhang, VLDB'1998)
- **CLIQUE**: CLustering In QUEst (Agrawal et al., SIGMOD'1998)



Apéndice
Reducción de dimensionalidad

Proyección en un espacio de menos dimensiones

$x \in \Re^{64 \times 64} = \Re^{4096}$



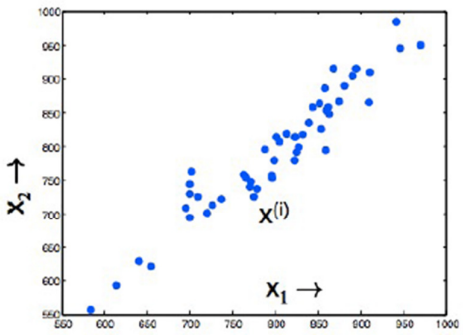
$y \in \Re^{10}$

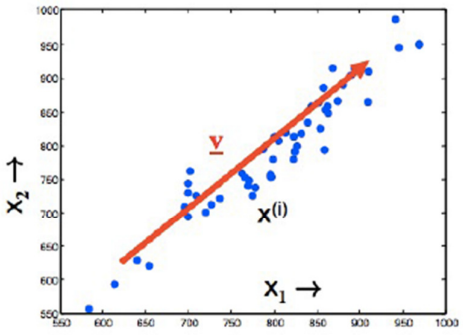
$$y = Ux$$

http://www.bigdataexaminer.com/understanding-dimensionality-reduction-principal-component-analysis-and-singular-value-decomposition/

Apéndice
Reducción de dimensionalidad

Análisis de componentes principales





$$\vec{x} = [x_1, x_2]$$

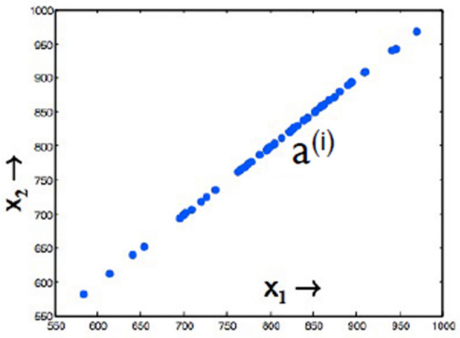
$$\vec{x} \approx s\vec{v} = s[v_1, v_2]$$

http://www.bigdataexaminer.com/understanding-dimensionality-reduction-principal-component-analysis-and-singular-value-decomposition/

Apéndice
Reducción de dimensionalidad

Análisis de componentes principales

Se escogen los vectores que minimizan la varianza de los residuos



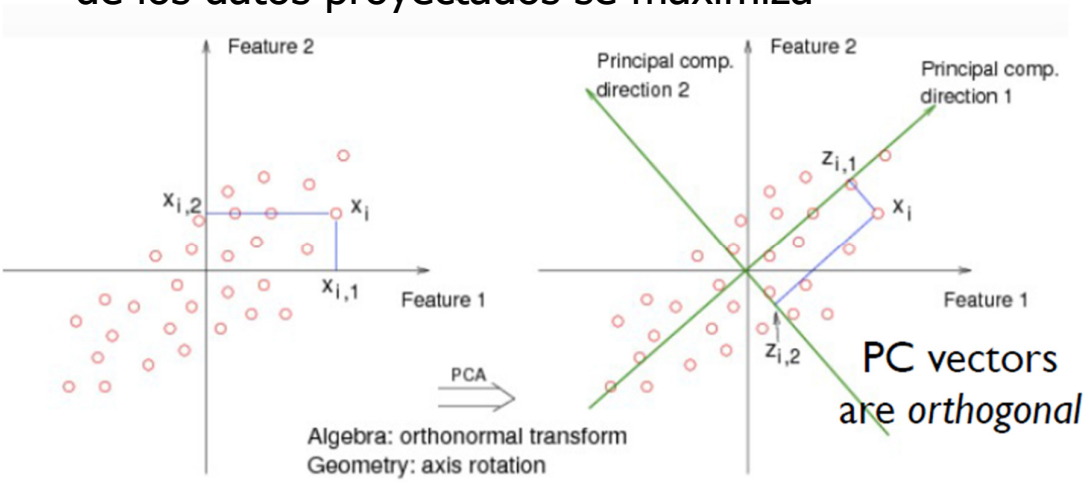
$$\min_{a,v} \sum_i (x^{(i)} - a^{(i)} v)^2$$

http://www.bigdataexaminer.com/understanding-dimensionality-reduction-principal-component-analysis-and-singular-value-decomposition/

Apéndice
Reducción de dimensionalidad

Análisis de componentes principales

Se proyecta en subespacios en los que la varianza de los datos proyectados se maximiza



http://www.bigdataexaminer.com/understanding-dimensionality-reduction-principal-component-analysis-and-singular-value-decomposition/

Apéndice
Reducción de dimensionalidad

Análisis de componentes principales

Cálculo de los componentes principales:

1. Se calcula la matriz de covarianza Σ de los datos.
2. Se calculan los k mayores eigenvectores de la matriz de covarianza Σ (los componentes principales).

Pero este cálculo puede ser demasiado costoso cuando se hace iterativamente sobre la matriz de covarianza…

… por lo que se suele utilizar la descomposición de valores singulares [SVD] para calcular los componentes principales.

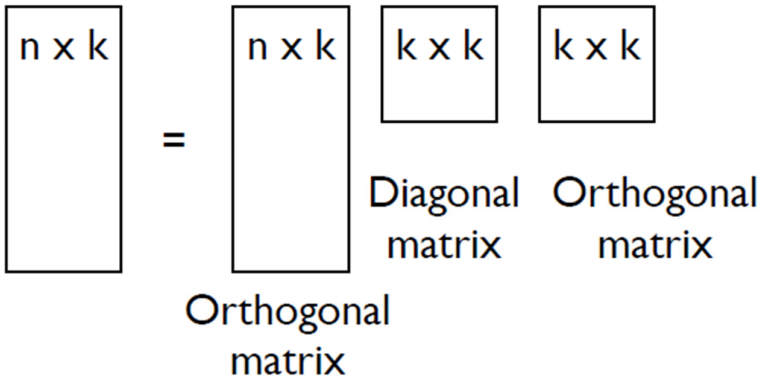
http://www.bigdataexaminer.com/understanding-dimensionality-reduction-principal-component-analysis-and-singular-value-decomposition/

Apéndice
Reducción de dimensionalidad

Descomposición de valores singulares [SVD]

Los eigenvectores son las columnas de la matriz U.

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathsf{T}}$$

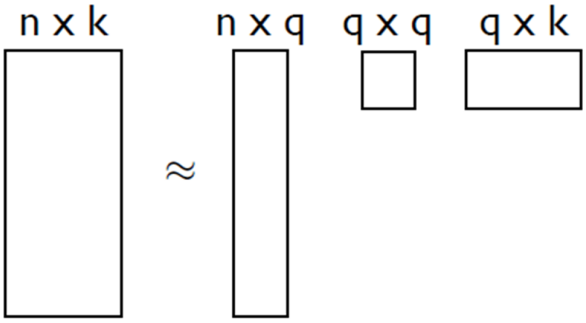


http://www.bigdataexaminer.com/understanding-dimensionality-reduction-principal-component-analysis-and-singular-value-decomposition/

Apéndice
Reducción de dimensionalidad

Se reduce la dimensionalidad quedándonos sólo con los q primeros componentes principales (q<k):

$$\mathbf{X} \approx \tilde{\mathbf{U}}\tilde{\mathbf{D}}\tilde{\mathbf{V}}^{\mathsf{T}}$$



http://www.bigdataexaminer.com/understanding-dimensionality-reduction-principal-component-analysis-and-singular-value-decomposition/

Apéndice
Reducción de dimensionalidad

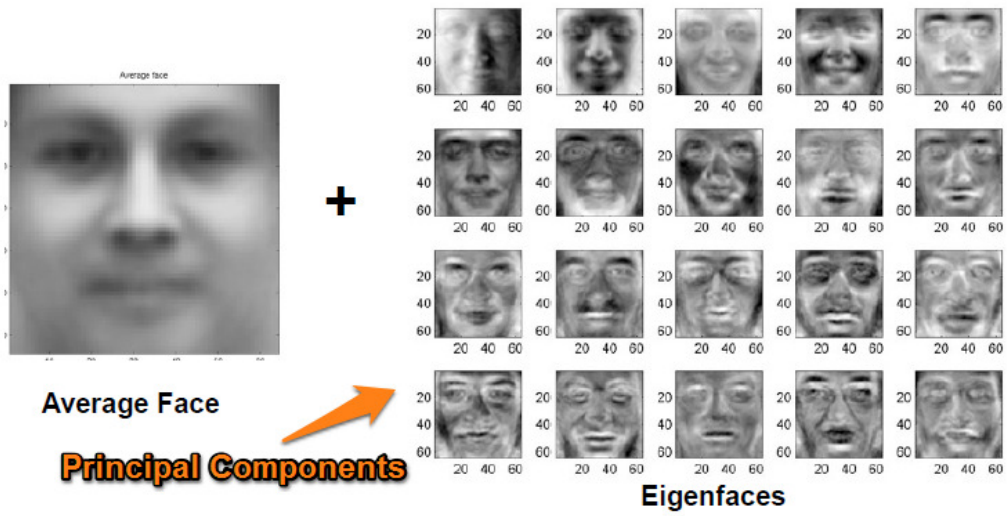
Ejemplo: Reconocimiento de caras

64x64 images of faces = 4096 dimensional data

http://www.bigdataexaminer.com/understanding-dimensionality-reduction-principal-component-analysis-and-singular-value-decomposition/

Apéndice
Reducción de dimensionalidad

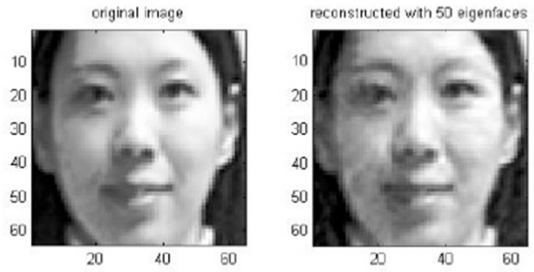
Ejemplo: Reconocimiento de caras



http://www.bigdataexaminer.com/understanding-dimensionality-reduction-principal-component-analysis-and-singular-value-decomposition/

Apéndice
Reducción de dimensionalidad

Ejemplo: Reconocimiento de caras



Los 50 primeros eigenvectores describen el 90% de la varianza, por lo que podemos quedarnos sólo con esas 90 dimensiones para reconstruir los datos sin perder demasiada calidad (de 4096 a 50 dimensiones ;-).

http://www.bigdataexaminer.com/understanding-dimensionality-reduction-principal-component-analysis-and-singular-value-decomposition/