Recuperación de Información Multimedia

Análisis de Componentes Principales

CC5213 – Recuperación de Información Multimedia
Departamento de Ciencias de la Computación
Universidad de Chile
Juan Manuel Barrios – https://juan.cl/mir/ – 2019

Reducción de la dimensión

- Transformar vectores de *d* dimensiones en vectores de *d'* dimensiones (*d'* << *d*) perdiendo la menor cantidad de información posible
- Ventajas
  - Eliminar ruido y/o dimensiones redundantes
  - Mejorar eficiencia de la búsqueda: Menos dimensiones → Distancias más rápidas de calcular, Menos requerimiento de memoria
- Método a estudiar: Análisis de Componentes Principales (PCA)

Conceptos

- Dados $x=\{x_1,\ldots,x_n\}$ e $y=\{y_1,\ldots,y_n\}$
  - Media:
    $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
  - Varianza:
    $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$
  - Covarianza:
    $$cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$
    - >0: valores aumentan o disminuyen juntos
    - <0: cuando un valor aumenta el otro disminuye

Matriz de Covarianza

- Para espacios *d*-dimensionales:

$$C^{d\times d} = \left(c_{ij} | c_{ij} = cov\left(Dim_i, Dim_j\right)\right)$$

  - Notar que es una matriz simétrica
  - Ejemplo: en 3 dimensiones

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$

Valores y Vectores Propios

- Para una matriz de *n* x *n* (transformación lineal) se buscan los vectores *v* tales que:

$$Av = \lambda v$$

  donde λ es un real, es decir, los vectores que no son modificados por la transformación (salvo escala)

- Resolver el polinomio característico igual a 0:

$$p(\lambda) = \det(A - \lambda I) = 0$$

- Para cada solución λ_i buscar el vector asociado (considerar multiplicidades):

$$(A - \lambda_i I)v = 0$$

Valores y Vectores Propios

- Si A es simétrica con valores reales:
  - Sus n valores propios son reales (multiplicidad ≥ 1)
  - Sus n vectores propios son ortogonales entre si
  - Sus n vectores propios son unitarios (norma 1)
  - A puede ser descompuesta en:

$$A = Q\, D\, Q^{\mathrm{T}}$$

    - Q es matriz ortogonal cuyas columnas son los vectores propios de A
    - D es matriz diagonal cuyos valores son los valores propios de A

https://en.wikipedia.org/wiki/Eigendecomposition_of_a_matrix

PCA

- Principal Component Analysis (PCA)
  - Aplicar una transformación lineal a los datos (multiplicar de cada vector por una matriz) que elimine las covarianzas
  - Descartar dimensiones de baja varianza (es decir, con “poca información”)
  - También conocida como la transformación Karhunen-Loève (KLT)

Algoritmo de PCA

- Entrada: vectores en R^d
- Centrar los datos (restar la media en cada dimensión)

$$\hat{x} = x - \bar{x}$$

- Calcular la matriz de covarianzas

$$\hat{C} = \begin{pmatrix} cov(\hat{x},\hat{x}) & cov(\hat{x},\hat{y}) & cov(\hat{x},\hat{z}) \\ cov(\hat{y},\hat{x}) & cov(\hat{y},\hat{y}) & cov(\hat{y},\hat{z}) \\ cov(\hat{z},\hat{x}) & cov(\hat{z},\hat{y}) & cov(\hat{z},\hat{z}) \end{pmatrix}$$

Algoritmo de PCA

- De la matriz de covarianzas calcular valores propios $\{\lambda_1, \ldots, \lambda_d\}$ con sus correspondientes vectores propios $\{w_1, \ldots, w_d\}$
- Reordenar valores propios de mayor a menor:
  - 1er componente principal es el vector propio asociado al mayor valor propio
  - 2º componente principal es el vector propio asociado al segundo mayor valor propio
  - El *d*-ésimo (y último) componente principal es el vector propio asociado al menor valor propio

Algoritmo de PCA

- Definir matriz de transformación:

$$W = (\ w_1\ ;\ w_2\ ;\ \dots\ ;\ w_d\ )$$

- Para disminuir la dimensión basta seleccionar las primeras *k* columnas (*k* ≤ *d*)
- Criterio para escoger *k*:
  - La suma de todos los valores propios representa la “cantidad de varianza” en el conjunto
  - Seleccionar los *k* valores propios que sumen una cierta fracción del total, es decir, dado un parámetro *p* cercano a 1 escoger *k* que obtenga:

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \sim p$$

Algoritmo de PCA

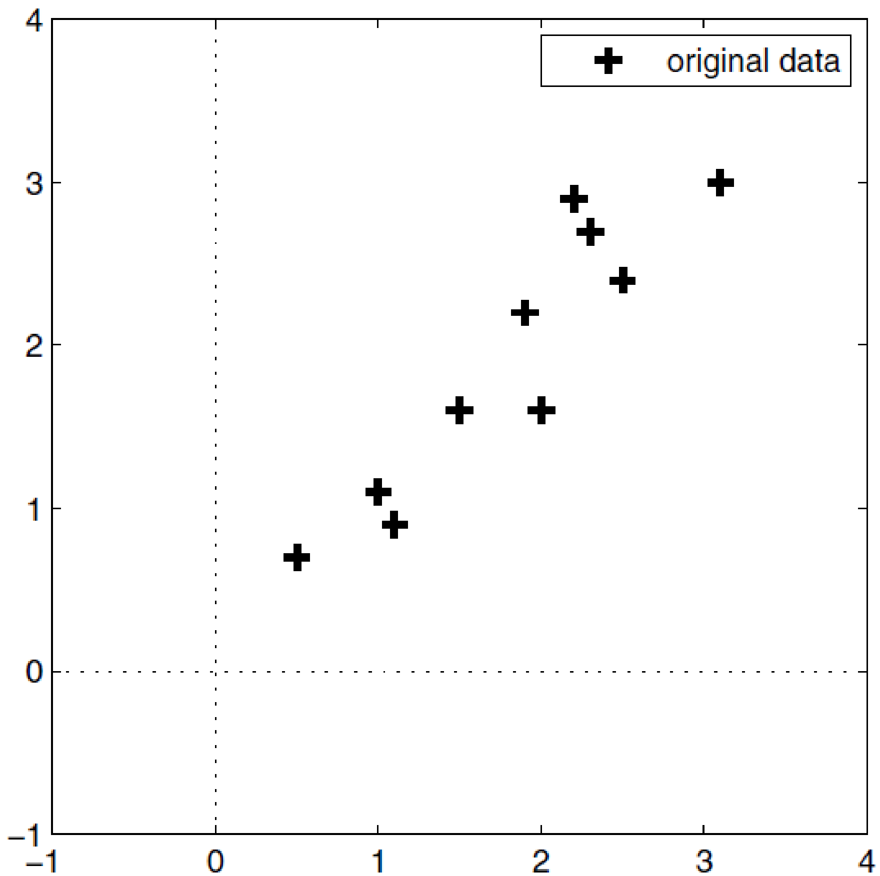
- Transformación de los datos:

$$y = W^T \cdot \hat{x}$$

- Reconstrucción de los datos:

$$x = \left((W^T)^{-1} \cdot y\right) + \hat{x} = \left((W^T)^T \cdot y\right) + \hat{x} = (W \cdot y) + \hat{x}$$

Ejemplo



- Centrar los datos (restar el promedio)
  - Opcional: escalar por el inverso de la desviación (estandarización)

- Calcular la matriz de covarianzas:

$$\hat{\Sigma} = \begin{pmatrix} 0.61656 & 0.61544 \\ 0.61544 & 0.71656 \end{pmatrix}$$

Ver Blanken et al., cap 3.C

Ejemplo

- Calcular valores y vectores propios
  - Valores propios:

$$\lambda_1 = 0.049083 \qquad \lambda_2 = 1.284$$

  - Vectores propios:

$$\begin{pmatrix} -0.73518 & -0.67787 \\ 0.67787 & -0.73518 \end{pmatrix}$$

$$w_1 \qquad\qquad w_2$$

Ejemplo

- Reordenar de mayor a menor
  - Valores propios (reordenados):

$$\lambda_1 = 1.284 \qquad \lambda_2 = 0.049083$$

  - Vectores propios (reordenados):

$$\begin{pmatrix} -0.67787 & -0.73518 \\ -0.73518 & 0.67787 \end{pmatrix}$$

$$\begin{array}{cc} \uparrow & \uparrow \\ w_1 & w_2 \end{array}$$

Ejemplo

- Transformación que descarta las covarianzas (mantiene el número de dimensiones):

$$W = (w_1; w_2) = \begin{pmatrix} -0.67787 & -0.73518 \\ -0.73518 & 0.67787 \end{pmatrix}$$

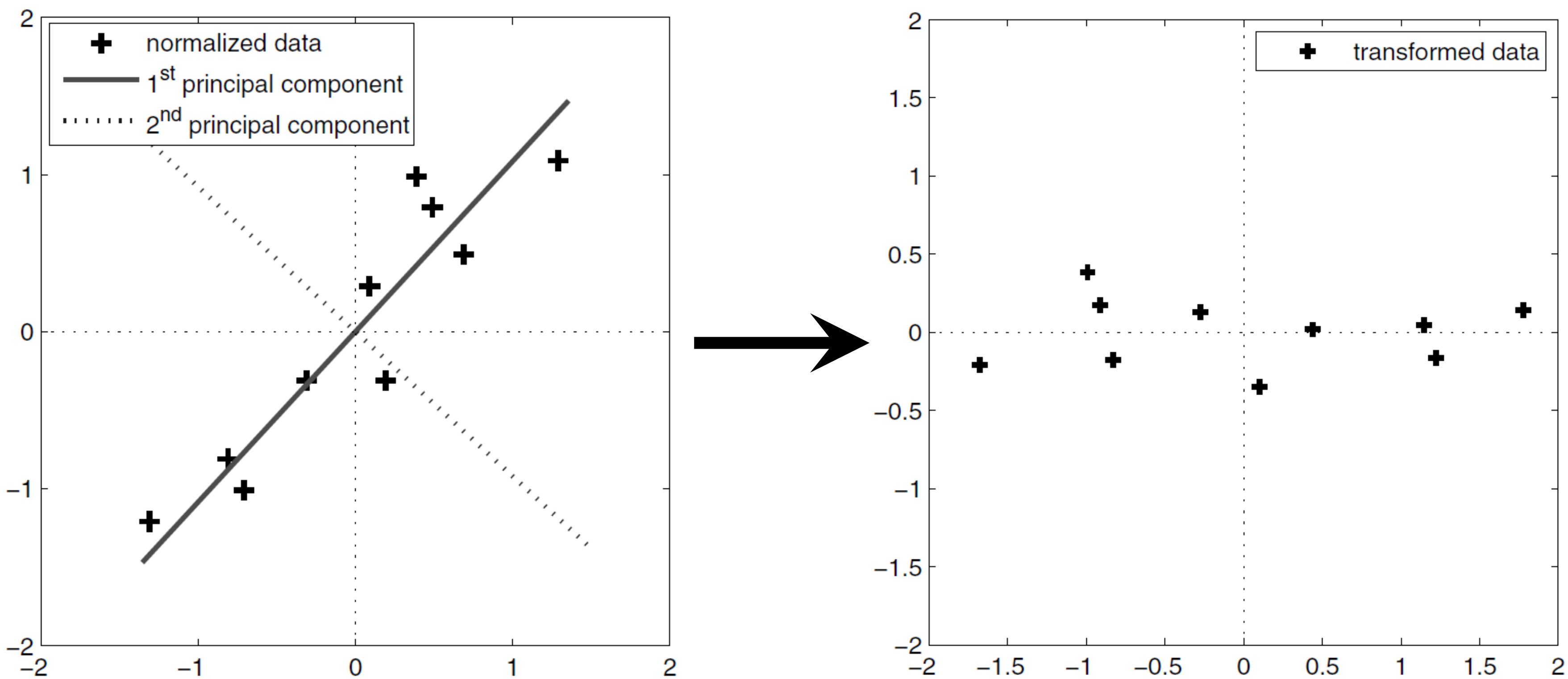
- Transformación que selecciona las k componentes principales:

$$W = (w_1) = \begin{pmatrix} -0.67787 \\ -0.73518 \end{pmatrix}$$

- Transformar los datos:

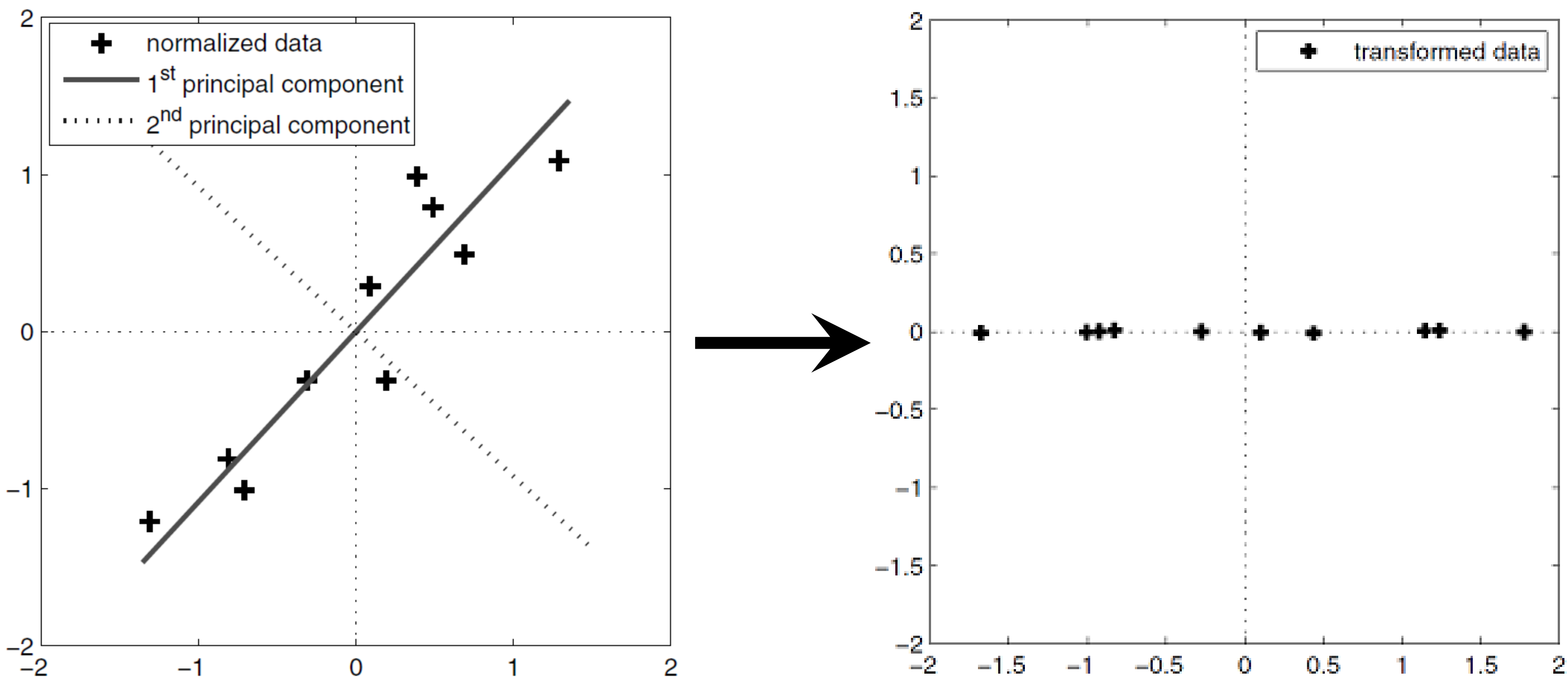
$$y = W^T \cdot \hat{x}$$

Transformación



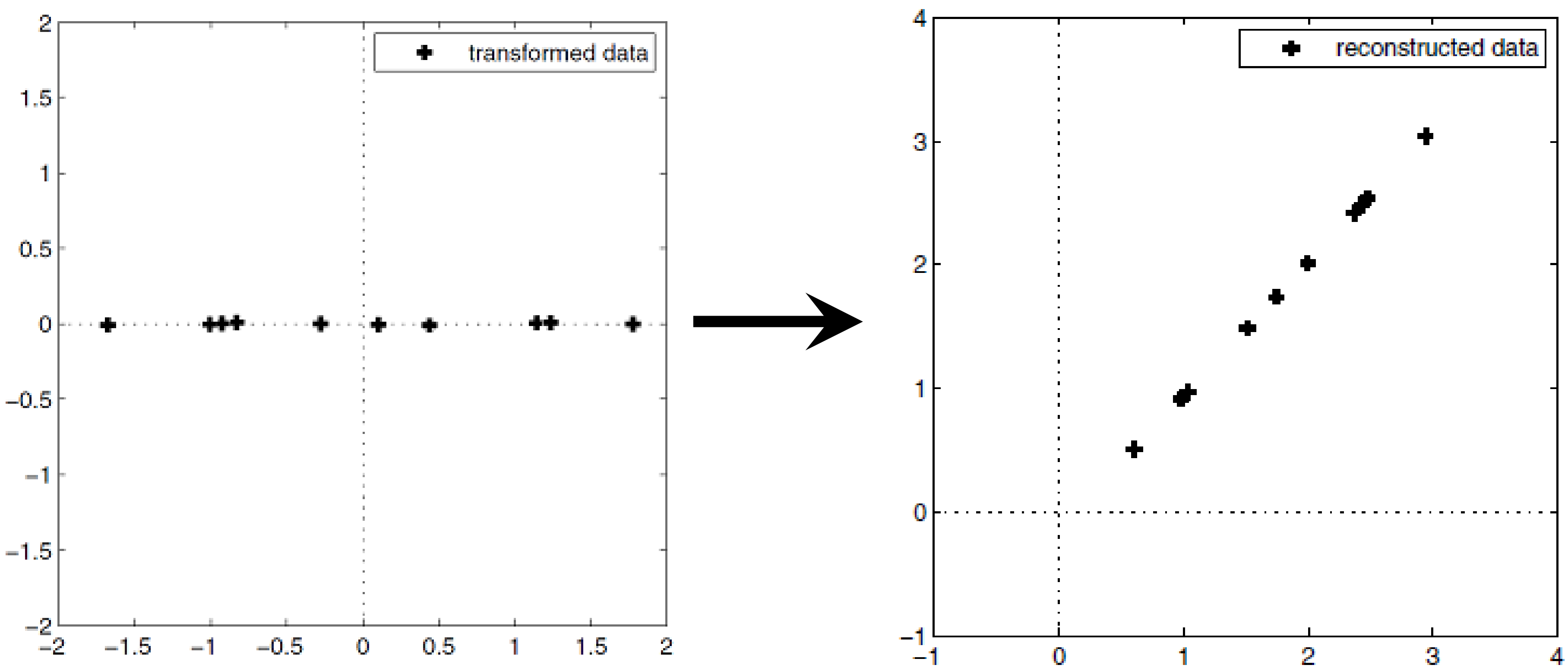
Transformación que mantiene el número de dimensiones, sólo elimina las covarianzas de los datos

Reducción



Mantiene un $\lambda_1/(\lambda_1+\lambda_2)$=96.3% del “total de la varianza” del conjunto

Reconstrucción



Con la transformación inversa y sumando la promedio se pueden devolver los datos al espacio original

USOS

Normalización de Pose

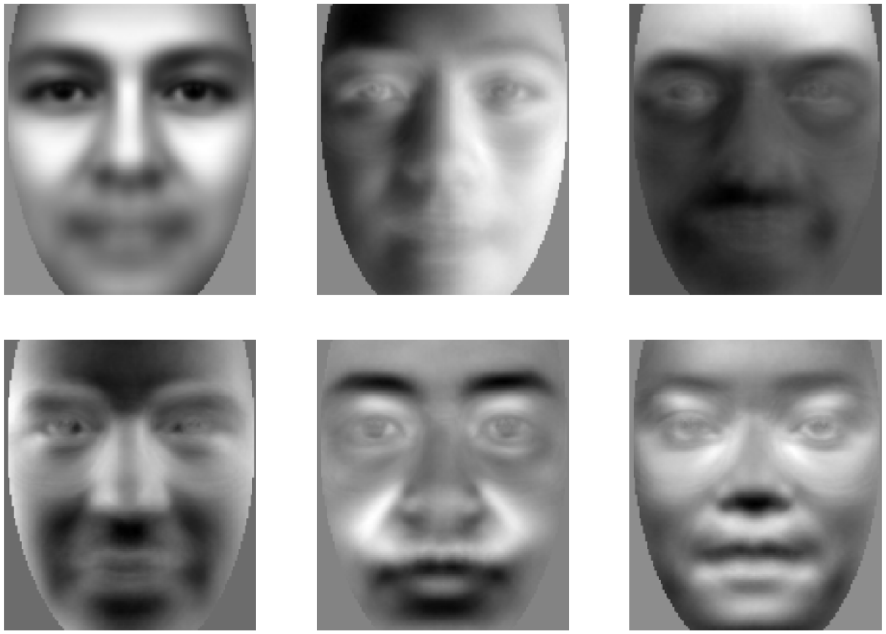
- PCA se puede usar para obtener ejes de rotación de un objeto
  - Permite normalizar la postura de un objeto
  - No funciona bien para objetos simétricos

PCA sobre imágenes

- Usar la imagen gris como vector de W*H dimensiones
- Usando PCA se reduce la dimensión, por ej., de varios miles a 256
- Se calcula distancia Euclidiana en el espacio de dimensión reducida

Eigenfaces

- Sistema reconocedor de rostros.
- Cara promedio y los primeros 5 vectores propios



Ver Blanken et al., cap 5.

PCA sobre imágenes

- Eigenobjects:
  - Si tenemos muchas imágenes parecidas, cada imagen puede ser vista como la imagen promedio más la suma ponderada de M vectores propios, llamados los “eigenobjects”
  - Cada imagen de la colección se representa por un conjunto de M escalares

Ver Blanken et al., cap 5.

Variantes de PCA

- PCA es un método lineal no supervisado
  - Transformaciones no lineales: Kernel PCA
  - Supervisado: Fisher's LDA
- Otros métodos: CCA, MDS, ICA, CFA, SOM, ISOMAP, LLE, NNCA.
- Específico a descriptores de texto: LSA

Bibliografía

- **Multimedia Retrieval**. Blanken, de Vries, Blok, Feng. 2007.
  - pág 83 (PCA)
  - pág 165 (eigenobjects)