# PCA, Principal Components Analysis

## Fundamentals

Rafael Pérez Torres
Selected Topics on Pattern Recognition
Profesor: Dr. Wilfrido Gómez Flores
*LTI Cinvestav*

Jun 19th, 2015

# AGENDA

# WHY TO REDUCE DIMENSIONALITY?

### The curse of dimensionality

- Coined by Bellman in 1961.
- The size of sample for estimating a multivariate function increases exponentially to the number of variables.
- **More variables, more computational cost.**

# WHY TO REDUCE DIMENSIONALITY? II

## Sparse space phenomenom

- Coined by Scott and Thompson.
- The actual guilty of *the curse of dimensionality*.
- High-dimensional spaces are inherently sparse.

## Intrinsic dimension

- The precursor for looking at dimensionality reduction.
- It refers to the amount of independent variables enough for describing a phenomenom.

# DIMENSIONALITY REDUCTION

### Definition

- Given a set of features, select the most important ones for reducing the set's size, keeping the maximum discriminatory information as possible.
- Typically, many of the features are less representative than noise in data, becoming irrelevant.
- Typically, many of the features are correlated between themselves.

# Dimensionality reduction types

## Types

- **Features' Extraction-Generation**: Given a features set $\mathbf{x}_i \in \mathbb{R}^M$, find a mapping $\mathbf{y} = f(x) : \mathbb{R}^M \to \mathbb{R}^m$, with $m < M$, such as the transformed vector $\mathbf{y}_i \in \mathbb{R}^m$ ***preserves information*** in $\mathbb{R}^M$.
- **Features selection**: Given a features set $\mathbf{x}_i = \{x_j | j = 1, \ldots, M\}$ find a subset $\mathbf{y}_i = \{x_{i1}, \ldots, x_{im}\}$, with $m < M$, such as it ***maximizes the performance*** of classification.

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix} \rightarrow \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = f\left( \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_3 \end{bmatrix} \right)$$

Features extraction

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix} \rightarrow \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iM} \end{bmatrix}$$

Features selection

# PRINCIPAL COMPONENT ANALYSIS PCA

## Basic idea

- Tries to identify most significant basis (perspectives) for re-expressing a data set, filtering the noise and revealing hidden structures.

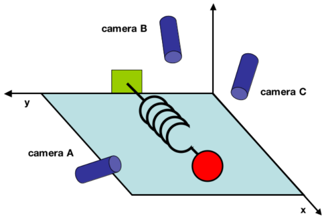


Figure: Different perspectives of problem and associated features

# PRINCIPAL COMPONENT ANALYSIS PCA II

## Basic idea

- We know $\overrightarrow{x}$ axis describes by itself the movement of the spring, but we don't know even what an axis is, we only have the raw perspectives given by data.

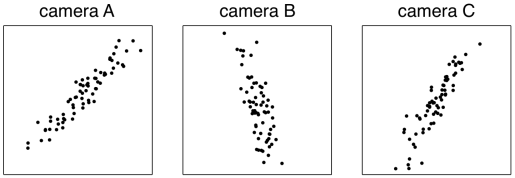


Figure: Data read from real world

- How to migrate from perspectives in Figure to the successful $\overrightarrow{x}$ perspective?

# Principal Component Analysis PCA III

## Basic idea

- Data can be expressed as a linear combination of its basis vectors.
- Let **X** a $m \times n$ matrix with the original data set, **Y** another $m \times n$ matrix for storing a new data representation built from matrix **P**, such as $\mathbf{PX} = \mathbf{Y}$.

$$\mathbf{PX} = \begin{bmatrix} p_1 \\ \vdots \\ p_m \end{bmatrix} \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \qquad Y = \begin{bmatrix} p_1 x_1 & \dots & p_1 x_n \\ \vdots & \ddots & \vdots \\ p_m x_1 & \cdots & p_m x_n \end{bmatrix}$$

New basis vectors

# PRINCIPAL COMPONENT ANALYSIS PCA IV

## Basic idea

- Geometrically, $P$ is a rotation and a stretch transforming $\mathbf{X}$ into $\mathbf{Y}$
- Rows of $\mathbf{P}$, $\{p_1, \ldots, p_m\}$ are the new set of basis vectors for representing $\mathbf{X}$.
- The $j$-th coefficient of $y_i$ is a projectionover the $j$-ith row of $\mathbf{P}$

# Elements for data description

## Noise and rotation

$$SNR = \frac{\sigma^2_{\text{signal}}}{\sigma^2_{\text{noise}}}$$

- A high SNR value means high accuracy, a low value indicates noise.

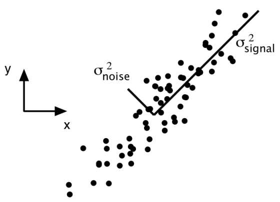


Figure: SNR and variance relation

# Elements for data description II

## Redundancy

- If you can explain attribute $r2$ from attribute $r1$, then they are correlated.
- The goal is to reduce the amount correlated variables.


Figure: Different degrees of data redundancy

# ELEMENTS FOR DATA DESCRIPTION III

## Covarianze matrix

- Obtains the degree of linear relation between two variables.
- High value means positive correlation, low value negative correlation.
- If each row of **X** represents all measurements of a type, then each column correponds to the set of measures in a particular time:

$$\mathbf{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}$$

and hence, covariance matrix $\mathbf{C_X}$ can be expressed as:

$$C_X \equiv \frac{1}{n}\mathbf{X}\mathbf{X}^T$$

# ELEMENTS FOR DATA DESCRIPTION III

## Desired features of covarianze matrix

- Its diagonal includes data variation, high values mean structural importance. **We look for high values**.
- Non diagonal items define covariance, high values mean high redundancy. **We look for 0 or low values (a uncorrelated matrix)**. This is the $\mathbf{P}$ in $\mathbf{Y} = \mathbf{PX}$

# Basic algorithm

## Basic algorithm

1: Select a direction in the $m$-dimensional dimensional such as variance of $\mathbf{X}$ is maximized and store it as $\mathbf{p_1}$.
2: Find another direction where variance is maximized, restricting search to orthogonal directions of those previously selected. Store it as $\mathbf{p_i}$.
3: Repeat procedure until $m$ vectors are selected.

# ALGEBRA OF SOLUTION

## Algebra

- The goal is to fin an ortonormal matrix $\mathbf{P}$ in $\mathbf{Y} = \mathbf{PX}$ such as $\mathbf{C_Y} \equiv \frac{1}{n}\mathbf{YY}^T$. Rows of $\mathbf{P}$ are the principal components of $\mathbf{X}$. Expressing $\mathbf{C_Y}$ in terms of unknown $\mathbf{P}$:

$$
\begin{aligned}
\mathbf{C_Y} &= \frac{1}{n}\mathbf{YY}^T && (1)\\
&= \frac{1}{n}(\mathbf{PX})(\mathbf{PX})^T && (2)\\
&= \frac{1}{n}\mathbf{PXX}^T\mathbf{P}^T && (3)\\
&= P(\frac{1}{n}\mathbf{XX}^T)\mathbf{P}^T && (4)\\
\mathbf{C_Y} &= \mathbf{PC_XP}^T && (5)
\end{aligned}
$$

# ALGEBRA OF SOLUTION II

## Algebra

- Any symmetric matrix $\mathbf{A}$ is diagonalized by an orthonormal matrix[1] built from its eigenvectores, that is $\mathbf{A} = \mathbf{EDE}^T$.
- When selecting $\mathbf{P}$ as a matrix where each of its rows $\mathbf{p_i}$ is an eigenvector of $\frac{1}{n}\mathbf{XX}^T$, it is achieved that $\mathbf{P} \equiv \mathbf{E^T}$.[2]
- Rewriting $\mathbf{C_Y}$:

$$
\begin{aligned}
\mathbf{C_Y} &= \mathbf{PC_XP}^T && (6)\\
&= \mathbf{P}(\mathbf{E^T DE})\mathbf{P^T} && (7)\\
&= \mathbf{P}(\mathbf{P^T DP})\mathbf{P^T} && (8)\\
&= (\mathbf{PP^T})\mathbf{D}(\mathbf{PP^T}) && (9)\\
&= (\mathbf{PP^{-1}})\mathbf{D}(\mathbf{PP^{-1}}) = \mathbb{I}\mathbf{D}\mathbb{I} && (10)\\
\mathbf{C_Y} &= \mathbf{D} && (11)
\end{aligned}
$$

[1] Matrix $A$ is ortogonal if $AA^T = \mathbb{I}$

[2] $\mathbf{P^{-1}} = \mathbf{P^T}$

# ALGORITHM OF EIGENVECTORS DECOMPOSITION SOLUTION

## Algorithm

1: Substract mean from data $\mu = \sum_{i=1}^{n} x_i - \overline{x}$
2: Calculate covariance matrix $\Sigma = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)}$
3: Calculate eigenvectors for $\Sigma$.
4: Select principal components (sort desc by eigenvalue).
5: Build new transformed data set $\mathbf{FinalData} = \mathbf{P}^T\mathbf{X}$

Thank you!