Principle Component Analysis

Jing Gao
SUNY Buffalo

Why Dimensionality Reduction?

- We have too many dimensions
  - To reason about or obtain insights from
  - To visualize
  - Too much noise in the data
  - Need to “reduce” them to a smaller set of factors
  - Better representation of data without losing much information
  - Can build more effective data analyses on the reduced-dimensional space: classification, clustering, pattern recognition

Component Analysis

- Discover a new set of factors/dimensions/axes against which to represent, describe or evaluate the data

- Factors are combinations of observed variables
  - May be more effective bases for insights
  - Observed data are described in terms of these factors rather than in terms of original variables/dimensions

Basic Concept

- Areas of variance in data are where items can be best discriminated and key underlying phenomena observed
  - Areas of greatest "signal" in the data
- If two items or dimensions are highly correlated or dependent
  - They are likely to represent highly related phenomena
  - If they tell us about the same underlying variance in the data, combining them to form a single measure is reasonable

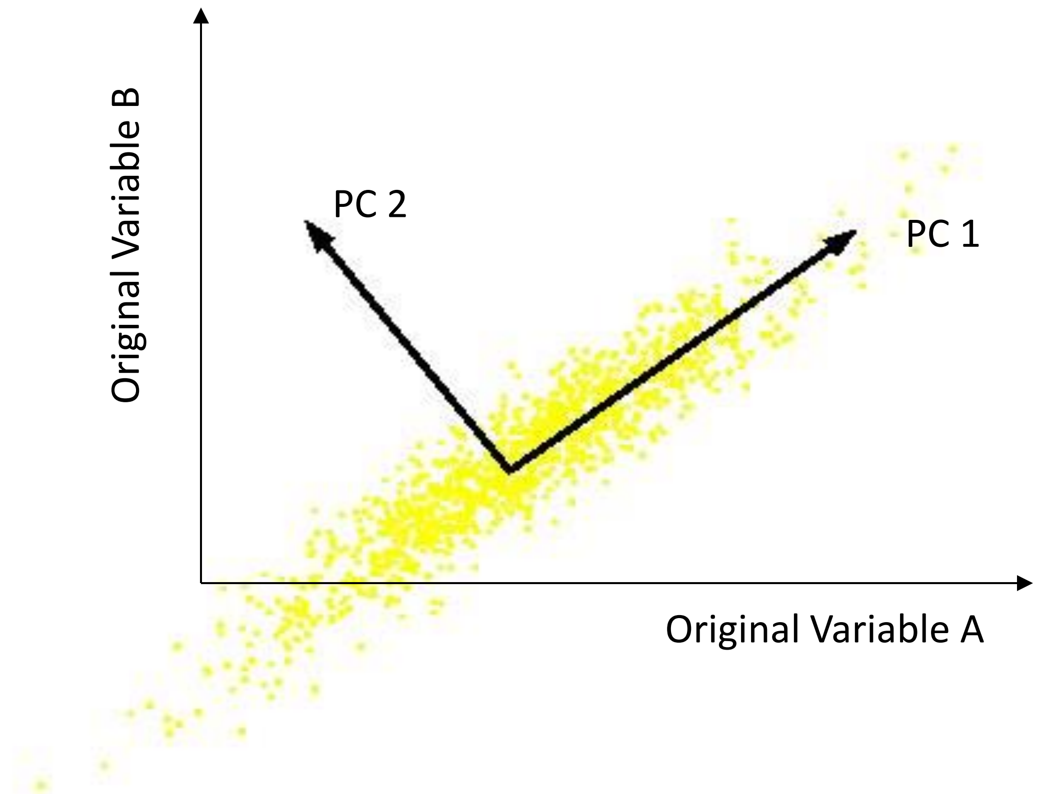
Basic Concept

- So we want to combine related variables, and focus on uncorrelated or independent ones, especially those along which the observations have high variance
- We want a smaller set of variables that explain most of the variance in the original data, in more compact and insightful form
- These variables are called “factors” or “principal components”

Principal Component Analysis

- Most common form of factor analysis
- The new variables/dimensions
  - Are linear combinations of the original ones
  - Are uncorrelated with one another
    - Orthogonal in dimension space
  - Capture as much of the original variance in the data as possible
  - Are called Principal Components

What are the new axes?



- Orthogonal directions of greatest variance in data
- Projections along PC1 discriminate the data most along any one axis

Principal Components

- First principal component is the direction of greatest variability (covariance) in the data
- Second is the next orthogonal (uncorrelated) direction of greatest variability
  - So first remove all the variability along the first component, and then find the next direction of greatest variability
- And so on …

Principal Components Analysis (PCA)

- Principle
  - Linear projection method to reduce the number of parameters
  - Transfer a set of correlated variables into a new set of uncorrelated variables
  - Map the data into a space of lower dimensionality

- Properties
  - It can be viewed as a rotation of the existing axes to new positions in the space defined by original variables
  - New axes are orthogonal and represent the directions with maximum variability

Algebraic definition of PCs

Given a sample of n observations on a vector of p variables

$$\left\{x_1, x_2, \cdots, x_n\right\} \in \Re^p$$

define the first principal component of the sample
by the linear transformation

$$z_1 = a_1^T x_j = \sum_{i=1}^{p} a_{i1} x_{ij}, \quad j = 1,2,\cdots,n.$$

where the vector

$$a_1 = (a_{11}, a_{21}, \cdots, a_{p1})$$

$$x_j = (x_{1j}, x_{2j}, \cdots, x_{pj})$$

is chosen such that $\mathrm{var}[z_1]$ is maximum.

Algebraic derivation of PCs

To find a_1 first note that

$$\mathrm{var}[z_1] = E((z_1 - \overline{z_1})^2) = \frac{1}{n}\sum_{i=1}^{n}\left(a_1^T x_i - a_1^T \overline{x}\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} a_1^T \left(x_i - \overline{x}\right)\left(x_i - \overline{x}\right)^T a_1 = a_1^T S a_1$$

where $$S = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(x_i - \overline{x}\right)^T$$

is the covariance matrix. $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the mean.

In the following, we assume the Data is centered. → $\overline{x} = 0$

Algebraic derivation of PCs

Assume $\bar{x} = 0$

Form the matrix: $X = [x_1, x_2, \cdots, x_n] \in \Re^{p \times n}$

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

Algebraic derivation of PCs

To find a_1 that maximizes $\text{var}[z_1]$ subject to $a_1^T a_1 = 1$

Let λ be a Lagrange multiplier

$$L = a_1^T S a_1 - \lambda(a_1^T a_1 - 1)$$

$$\frac{\partial}{\partial a_1} L = S a_1 - \lambda a_1 = 0$$

$$\Rightarrow S a_1 = \lambda a_1$$

$$\Rightarrow a_1^T S a_1 = \lambda$$

therefore a_1 is an eigenvector of S

corresponding to the largest eigenvalue $\lambda = \lambda_1$.

Algebraic derivation of PCs

To find the next coefficient vector a_2 maximizing $\text{var}[z_2]$

subject to $\text{cov}[z_2, z_1] = 0$ **uncorrelated**

and to $a_2^T a_2 = 1$

$$\text{cov}[z_2, z_1] = a_1^T S a_2 = \lambda_1 a_1^T a_2$$

then let λ and φ be Lagrange multipliers, and maximize

$$L = a_2^T S a_2 - \lambda(a_2^T a_2 - 1) - \phi a_2^T a_1$$

Algebraic derivation of PCs

We find that a_2 is also an eigenvector of S

whose eigenvalue $\lambda = \lambda_2$ is the second largest.

In general

$$\mathrm{var}[z_k] = a_k^T S a_k = \lambda_k$$

- The k^{th} largest eigenvalue of S is the variance of the k^{th} PC.
- The k^{th} PC z_k retains the k^{th} greatest fraction of the variation in the sample.

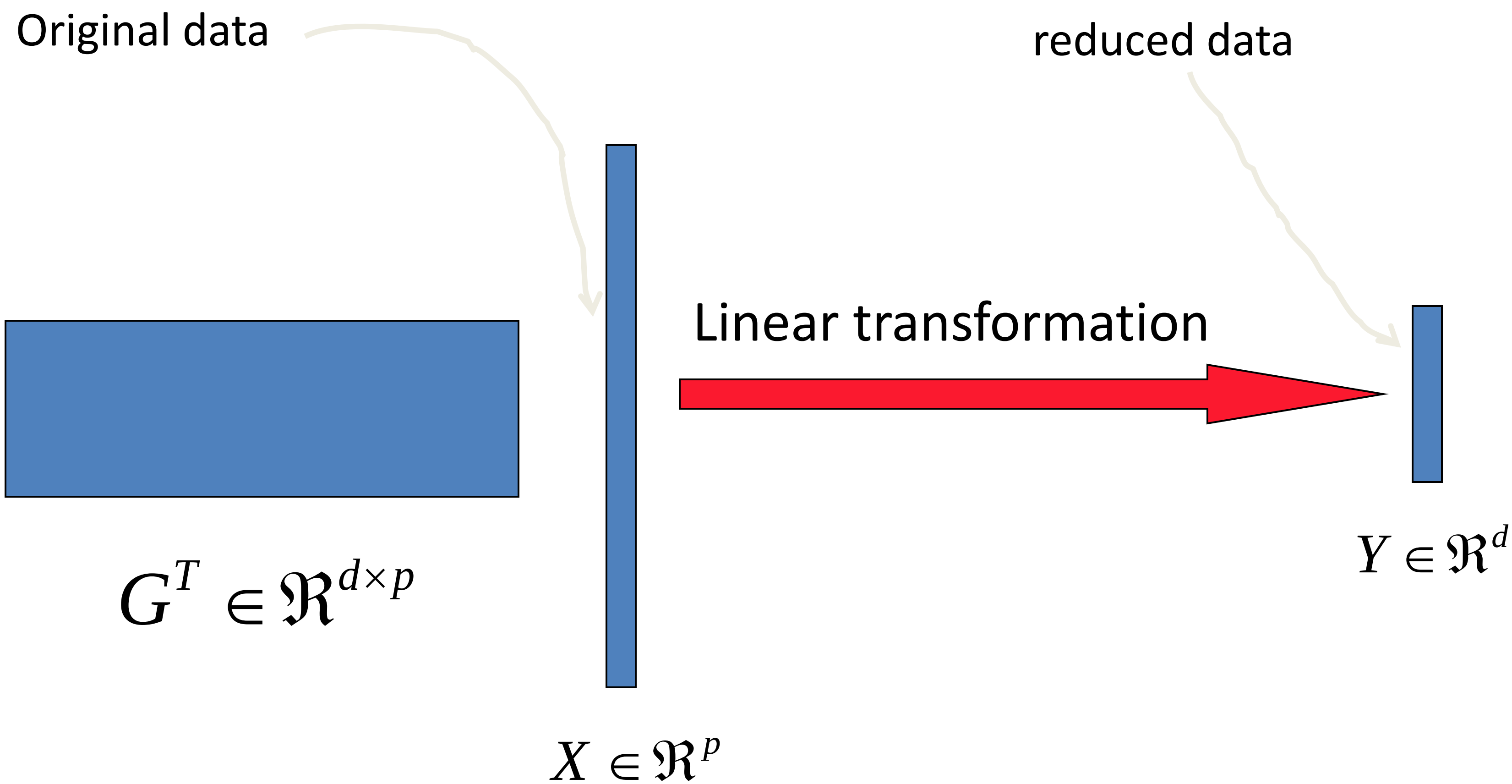
Algebraic derivation of PCs

- Main steps for computing PCs
  - Form the covariance matrix S.
  - Compute its eigenvectors: $\{a_i\}_{i=1}^{p}$
  - Use the first d eigenvectors $\{a_i\}_{i=1}^{d}$ to form the d PCs.
  - The transformation G is given by

$$G \leftarrow [a_1, a_2, \cdots, a_d]$$

A test point $x \in \Re^p \rightarrow G^T x \in \Re^d$.

Dimensionality Reduction

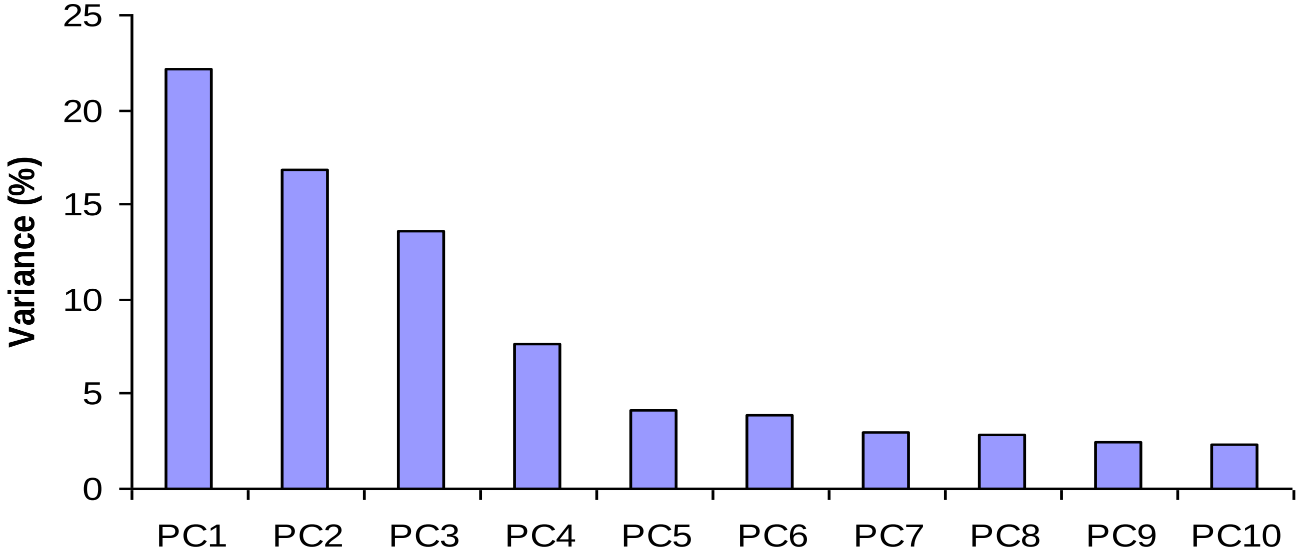


$$G \in \Re^{p \times d} : X \rightarrow Y = G^T X \in \Re^{d}$$

Steps of PCA

- Let $\overline{X}$ be the mean vector (taking the mean of all rows)
- Adjust the original data by the mean
  $X' = X - \overline{X}$
- Compute the covariance matrix S of adjusted X
- Find the eigenvectors and eigenvalues of S.

Principal components - Variance



Transformed Data

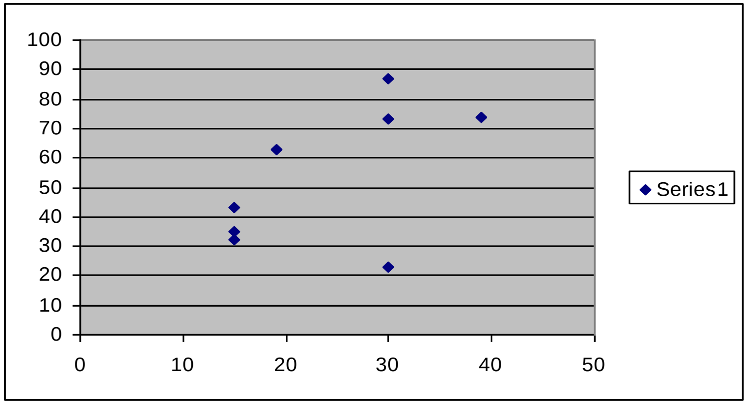
- Eigenvalues λ_j corresponds to variance on each component *j*
- *Thus, sort by* λ_j
- Take the first *d* eigenvectors $\mathbf{a_i}$; where d is the number of top eigenvalues
- These are the directions with the largest variances

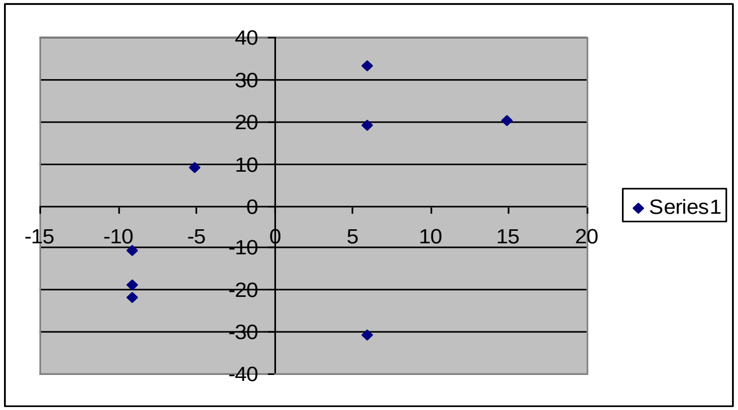
$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ \dots \\ y_{id} \end{pmatrix} = \begin{pmatrix} \vec{a}_1 \\ \vec{a}_2 \\ \dots \\ \vec{a}_d \end{pmatrix} \begin{pmatrix} x_{i1} - \overline{x_1} \\ x_{i2} - \overline{x_2} \\ \dots \\ x_{in} - \overline{x_n} \end{pmatrix}$$

An Example

X1	X2	X1'	X2'
19	63	-5.1	9.25
39	74	14.9	20.25
30	87	5.9	33.25
30	23	5.9	-30.75
15	35	-9.1	-18.75
15	43	-9.1	-10.75
15	32	-9.1	-21.75
30	73	5.9	19.25

Mean1=24.1
Mean2=53.8





Covariance Matrix

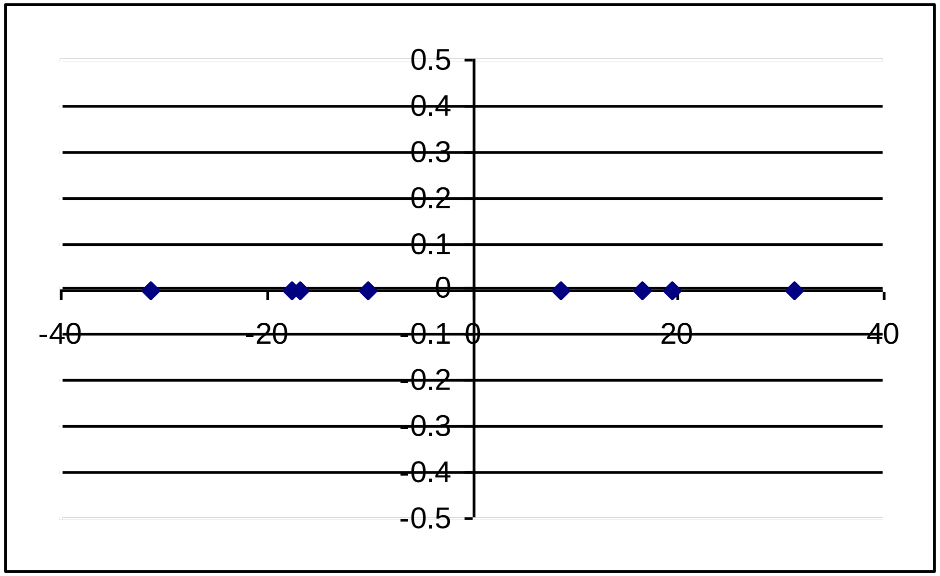
- C=

75	106
106	482

- We find out:
  - Eigenvectors:
  - a2=(-0.98,-0.21), λ2=51.8
  - a1=(0.21,-0.98), λ1=560.2

Transform to One-dimension

- We keep the dimension of a1=(0.21,-0.98)
- We can obtain the final data as



yi
-10.14
-16.72
-31.35
31.374
16.464
8.624
19.404
-17.63

$$y_i = \begin{pmatrix} 0.21 & -0.98 \end{pmatrix} \begin{pmatrix} x_{i1} \\ x_{i2} \end{pmatrix} = 0.21 * x_{i1} - 0.98 * x_{i2}$$