Principal Component Analysis

ECE/CS 498 DS U/G

Lecture 11

Ravi K. Iyer

Dept. of Electrical and Computer Engineering

University of Illinois at Urbana Champaign

Announcements

- MP2 checkpoint 1 due today, Feb 25
- In class activity 3 on Wednesday, Feb 27
  - Principal Component Analysis, Clustering

- Today : Dr. Weinshilbaum MD Mayo Clinic Center for Individualized Medicine
- Intro to Principal Components Analysis (PCA)

Dimensionality Reduction

- Can your data be explained with fewer dimensions?
  - Available data may have high dimensionality
  - Actual information of interest may be explained by a smaller number of dimensions/features

- Goal of dimensionality reduction is to explain the data with as few dimensions as possible while retaining the underlying "structure" in the data

- We use the terms "feature" and "dimension" interchangeably

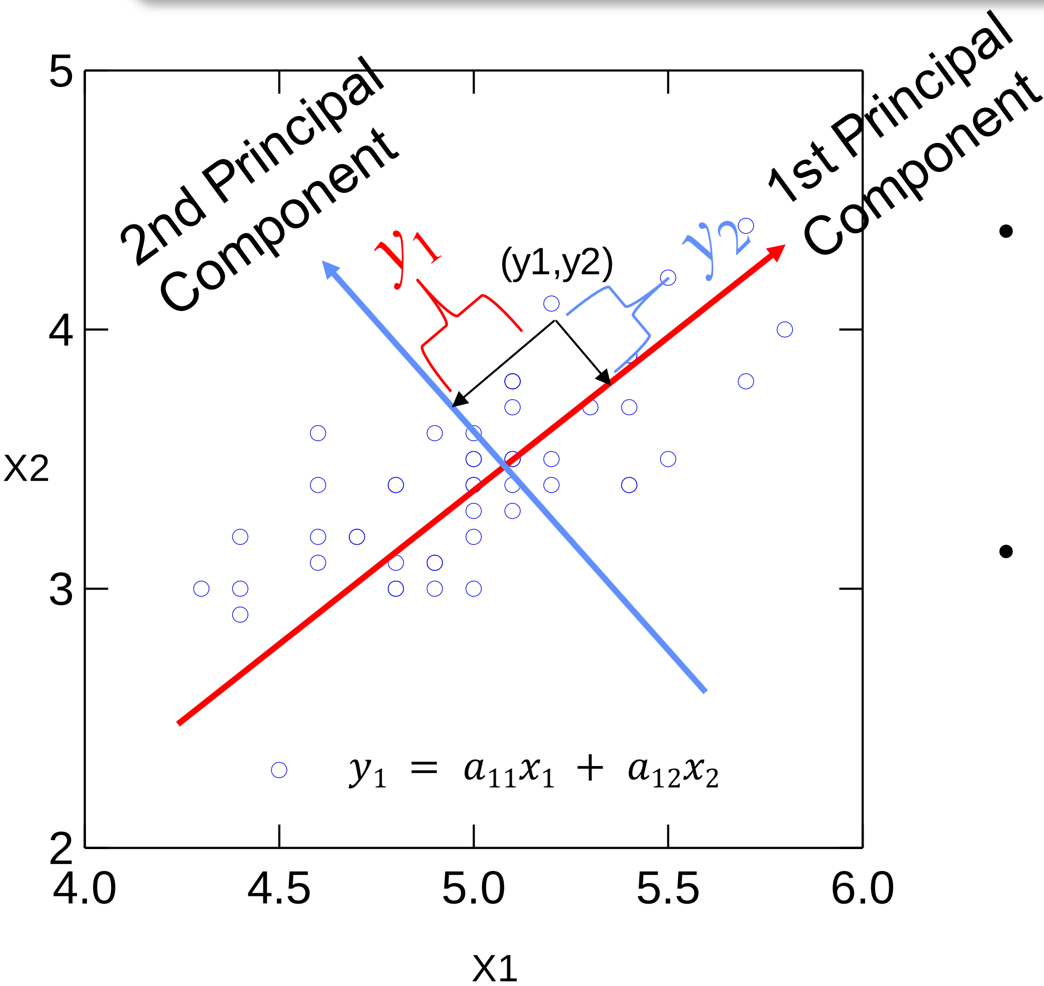
- Several ways to reduce dimension of the data
  - Drop unimportant dimensions using e.g. domain knowkedge
  - Take a (linear) combination of features*

Principal Components Analysis (PCA)

- Principal Components Analysis (PCA)
  - In PCA, "structure" refers to the variance in the data
  - Goal is to reduce dimensionality (to m) while explaining the most variance in the data so that with $m \ll d$, most of the data can be explained
  - The way we extract relevant features is by taking linear combinations of existing dimensions
  - Thus ***PCA is a statistical technique to analyze the relationships among a large number of variables and to explain these variables using smaller number of variables that we call its principal components***

- To define principal components
  - Center the data
  - Chose as the 1^{st} direction, the direction of maximum variance in the data
  - 2^{nd} direction is chosen to be perpendicular to the first , that explains the maximum remaining variance in the data
  - And so on (Keeping successive directions orthogonal)

PCA: Dimensionality Reduction Method



- What is a good feature?
  - Simplify the explanation of the input
  - Reduce dimensionality
- Why pick the direction that maximizes variability?

Principal Component Analysis

- From p random vectors (features in the dataset) $X = [X_1, X_2, \dots, X_p]$

  Produce *p* new variables: $y_1, y_2, \dots, y_p$:

$$y_1 = a_{11}x_1 + a_{12}x_2 + \dots + a_{1p}x_p$$
$$y_2 = a_{21}x_1 + a_{22}x_2 + \dots + a_{2p}x_p$$
$$\dots$$
$$y_p = a_{p1}x_1 + a_{p2}x_2 + \dots + a_{pp}x_p$$

$$\boldsymbol{Y = A'X}$$

y_j's are **Principal Components**

There is no intercept, but a_{j1}, a_{j2}, ..., a_{jp} can be viewed as **regression coefficients.**

such that:

- y_j's are uncorrelated (orthogonal) covariance among each pair of the principal axes is zero
- y_1 explains as much as possible of original variance in data set, y_2 explains as much as possible of remaining variance, and so on.

PCA Applications

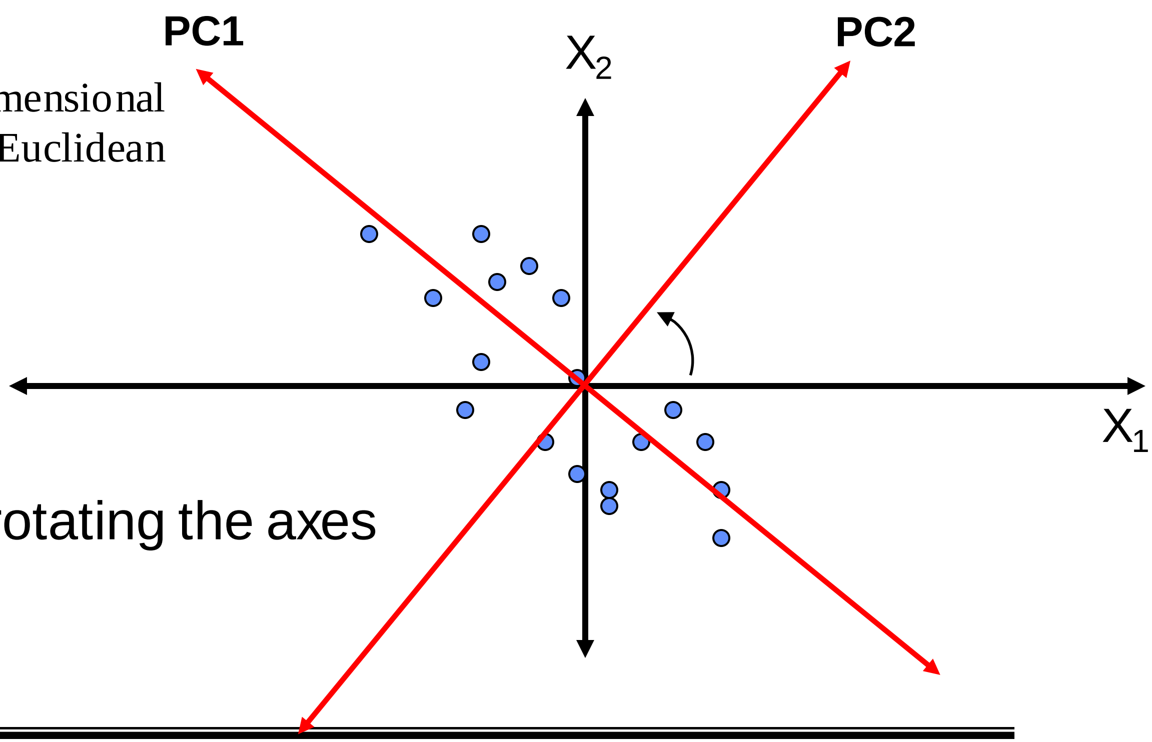
- Uses:
  - Data Visualization
  - Data Reduction
  - Data Classification
  - Trend Analysis
  - Factor Analysis
  - Noise Reduction
  - Regression
  - Clustering

- Examples:
  - How many unique “sub-sets” are in the sample?
  - How are they similar / different?
  - What are the underlying factors that influence the samples?
  - Which time / temporal trends are (anti)correlated?
  - Which measurements are needed to differentiate?
  - How to best present what is “interesting”?
  - Which “sub-set” does this new sample rightfully belong?

Trick: Rotate Coordinate Axes

Suppose we have a population measured on p random variables $X_1,\ldots,X_p$. Note that these random variables are represented on a p-axes Cartesian coordinate system. Our goal is to develop a new set of p axes (linear combinations of the original p axes) in the directions of greatest variability:

PCA derives the best possible *k* dimensional ($k < p$) representation of the Euclidean distances among observations



This is accomplished by rotating the axes

Principal Component Analysis (eigenvalues and eigenvectors)

$(a_{11}, a_{12}, \ldots, a_{1p})$ is 1st **Eigenvector (direction of PC in the rotated axis)** of correlation/covariance matrix, and **coefficients** of first principal component

Eigen values corresponding to that Eigen vector explain how much variance is in the data in the chosen direction

$(a_{21}, a_{22}, \ldots, a_{2p})$ is 2nd **Eigenvector** of correlation/covariance matrix, and **coefficients** of 2nd principal component

…

$(a_{p1}, a_{p2}, \ldots, a_{pp})$ is *p*th **Eigenvector** of correlation/covariance matrix, and **coefficients** of *p*th principal component

The Algebra of PCA: Covariance Matrix

- First step is to calculate the variance-covariance among every pair of the *p* features/dimensions in the dataset of n observations

$$S = Covariance\ (X) = (X - \bar{x})^T (X - \bar{x})$$

- Square, symmetric matrix
- Diagonals are the variances, off-diagonals are the covariance

	X_1	X_2
X_1	6.6707	3.4170
X_2	3.4170	6.2384

Variance-covariance Matrix

Trace (sum of diagonals): 12.9091

- Sum of the diagonals of the variance-covariance matrix is called the trace and it represents the total variance in the data

The Algebra of PCA

Finding the principal components and their explained variance involves eigen analysis of the covariance or correlation matrix (S)

$$\boldsymbol{Sa} = \boldsymbol{\lambda a}$$

Covariance Matrix

eigenvalue eigenvector

- First eigenvector (corresponding to largest eigenvalue) is the first principal component
- Second eigenvector (corresponding to the second largest eigenvalue) is the second principal component
- And so…
- The eigenvalue is the variance explained by direction of corresponding eigenvector

The Algebra of PCA: Eigenvalues

- Eigenvalues (latent roots) of S are solutions (λ) to the characteristic equation

$$\left|\mathbf{S} - \boldsymbol{\lambda}\mathbf{I}\right| = \mathbf{0} \Rightarrow \begin{vmatrix} s_{11} - \lambda & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} - \lambda & \cdots & s_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} - \lambda \end{vmatrix} = 0$$

- the eigenvalues, $\lambda_1, \lambda_2, \ldots \lambda_p$ are the variances of the coordinates on each principal component axis
- the sum of all p eigenvalues equals the trace of S (the sum of the variances of the original variables)

The Algebra of PCA: Eigenvalues

- Computing the eigenvalues of the covariance matrix

$$S = \begin{bmatrix} 6.6707 & 3.4170 \\ 3.4170 & 6.2384 \end{bmatrix}$$

$$|\mathbf{S} - \lambda \mathbf{I}| = \mathbf{0} \quad \Rightarrow \quad \left| \begin{bmatrix} 6.6707 & 3.4170 \\ 3.4170 & 6.2384 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right| = 0$$

Trace = 12.9091

$$\Rightarrow \quad \begin{vmatrix} 6.6707 - \lambda & 3.4170 \\ 3.4170 & 6.2384 - \lambda \end{vmatrix} = 0$$

$$\Rightarrow \quad (6.6707 - \lambda)(6.2384 - \lambda) - 3.4170 * 3.4170 = 0$$

$$\Rightarrow \quad \lambda^2 - 12.9091\lambda + 29.934 = 0$$

$$\Rightarrow \quad \boldsymbol{\lambda_1 = 9.8783, \lambda_2 = 3.0308}$$

Note: $\lambda_1 + \lambda_2 = 12.9091$

- **After selecting $k < p$ components, the total variance in the dataset is not equal to the trace of the Covariance matrix**

The Algebra of PCA: Eigenvectors

- Each **eigenvector** consists of *p* values which represent the "contribution" of each variable to the **principal component** axis
- Eigenvectors are uncorrelated (orthogonal)
  - their dot product $a_i^T a_j = 0$ if $i \neq j$
- Eigenvectors can be obtained using the following equation

$$\boldsymbol{S a_i = \lambda_i a_i}$$

for all $i \in \{1,2,\ldots,p\}$

The Algebra of PCA: Eigenvectors

Computing the eigenvectors of the covariance matrix S using the calculated eigenvalues:

$$S = \begin{bmatrix} 6.6707 & 3.4170 \\ 3.4170 & 6.2384 \end{bmatrix}$$

$$\lambda_1 = 9.8783 \quad \lambda_2 = 3.0308$$

Let us look at the first eigenvector:

$$Sa_1 = \lambda_1 a_1 \quad \Rightarrow \quad (S - \lambda_1 I)a_1 = 0$$

$$\Rightarrow \begin{bmatrix} 6.6707 - 9.8783 & 3.4170 \\ 3.4170 & 6.2384 - 9.8783 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} = 0$$

$$\Rightarrow \begin{bmatrix} -3.2076 & 3.4170 \\ 3.4170 & -3.6399 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} = 0 \quad \Rightarrow \quad \begin{matrix} -3.2076a_{11} + 3.4170a_{12} = 0 \text{ (Eq2)} \\ 3.4170a_{11} - 3.6399a_{12} = 0 \text{ (Eq1)} \end{matrix}$$

Solving Eq1 and Eq2 simultaneously, we get: $a_{11} = 1.0653, a_{12} = 1$

Similarly, can solve for a_2. Eigenvectors are: $\boldsymbol{a_1} = \frac{1}{\sqrt{1.0653^2+1^2}} \begin{bmatrix} \mathbf{1.0653} \\ \mathbf{1} \end{bmatrix}, \boldsymbol{a_2} = \frac{1}{\sqrt{0.9387^2+1^2}} \begin{bmatrix} \mathbf{-0.9387} \\ \mathbf{1} \end{bmatrix}$

The Algebra of PCA: Eigenvectors

- Eigenvectors are uncorrelated (orthogonal)
  - their dot product $a_i^T a_j = 0$ if $i \neq j$

- From the example, we get

	Eigenvectors $\downarrow$ $\boldsymbol{a_1}$	$\downarrow$ $\boldsymbol{a_2}$
$\boldsymbol{X_1}$	1.0653	-0.9387
$\boldsymbol{X_2}$	1	1

- Checking for orthogonality:

$$a_1^T a_2 = 1.0653 * (-0.9387) + 1 = 0$$

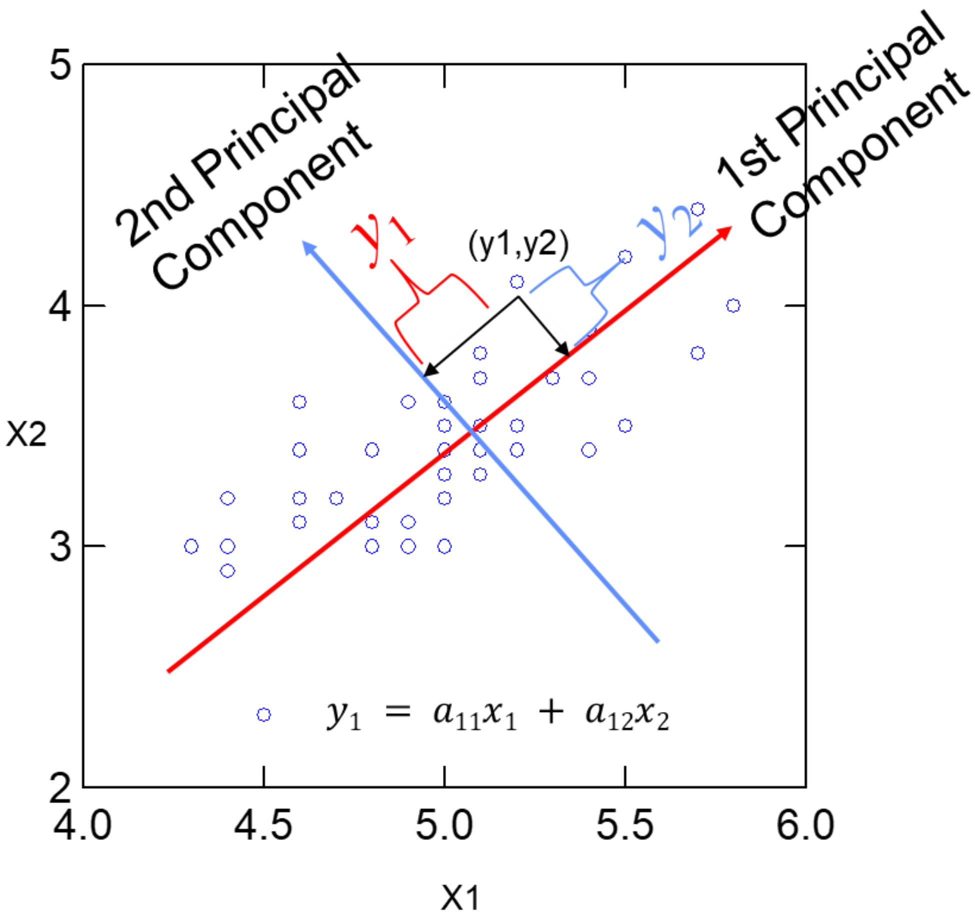
The Algebra of PCA

- Coordinates of each observation on the j^{th} principal axis, known as the **scores** on PC j, are computed as

$$y_j = a_{j1}x_1 + a_{j2}x_2 + \ldots + a_{jk}x_k$$
$$E.g, \qquad y_1 = 1.0653x_1 + 1x_2$$

- variance of the scores on each PC axis is equal to the corresponding eigenvalue for that axis
- the eigenvalue represents the variance displayed ("explained" or "extracted") by the kth axis
- the sum of the first k eigenvalues is the variance explained by the k-dimensional ordination.

The Algebra of PCA



The covariance matrix on *p* principal axes has a simple form:

- all off-diagonal values are zero (the principal axes are uncorrelated)
- the diagonal values are the eigenvalues.

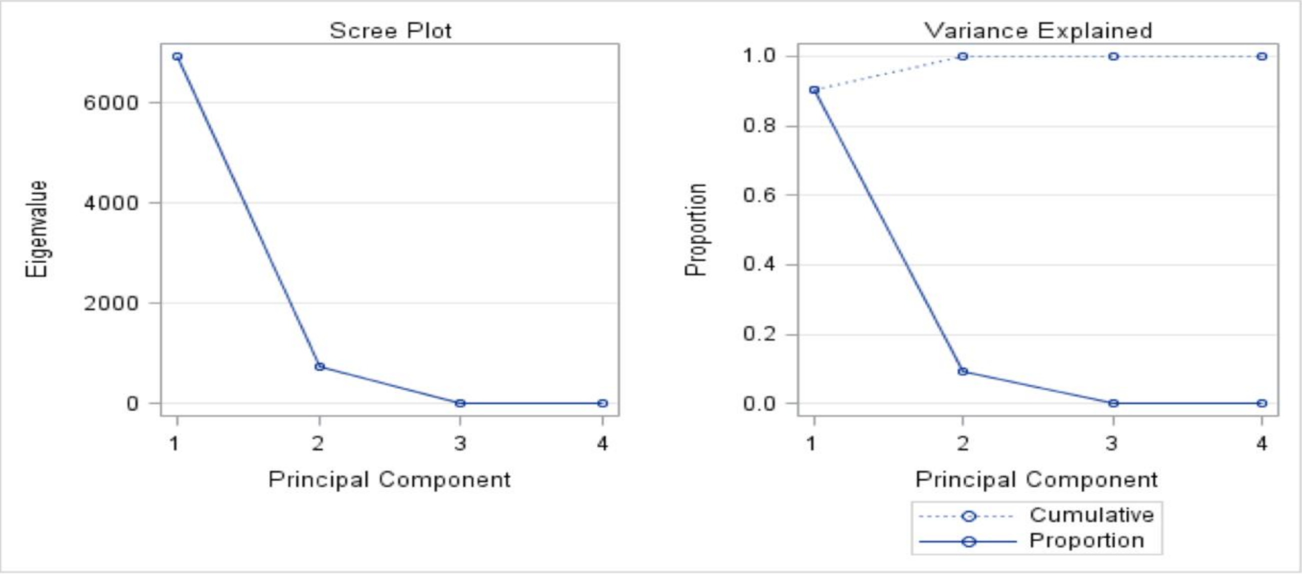
	PC_1	PC_2
PC_1	9.8783	0.0000
PC_2	0.0000	3.0308

Variance-covariance Matrix of the PC axes

Number of Dimensions

- If input was p-dimensions, how many dimensions do we keep
  – No solid answer, heuristics exists

- Look at Eigen values
  – They show variance of each component at some point they will be small



The Algebra of PCA: Covariance/Correlation Matrix

- PCA can be found using the covariance matrix OR the correlation matrix
- ***Covariance Matrix:***
  - Variables must be in same units
  - Emphasizes variables with most variance
  - **Using covariance's among variables only makes sense if they are measured in the same units**
- ***Correlation Matrix:***
  - Variables are standardized (mean 0.0, SD 1.0)
  - Variables can be in different units
  - All variables have same impact on analysis

$$r_{ij} = \frac{C_{ij}}{\sqrt{V_i V_j}}$$

Covariance of variables *i* and *j*

Correlation between variables *i* and *j*

Variance of variable *j*

	X_1	X_2
X_1	6.6707	3.4170
X_2	3.4170	6.2384

Variance-covariance Matrix

	X_1	X_2
X_1	1.0000	0.5297
X_2	0.5297	1.0000

Correlation Matrix

Trace (sum of diagonals): 12.9091

Trace (sum of diagonals): 2.0

Additional Resources

- Textbook "The Elements of Statistical Learning" , Section 14.5 Principal Components, Curves and Surfaces

- Roweis, Sam T. "EM algorithms for PCA and SPCA." *Advances in neural information processing systems*. 1998.