Principal Component Analysis

Lecture 12:

Principal Component Analysis

ECE/CS 498 DS

Professor Ravi K. Iyer

Department of Electrical and Computer Engineering

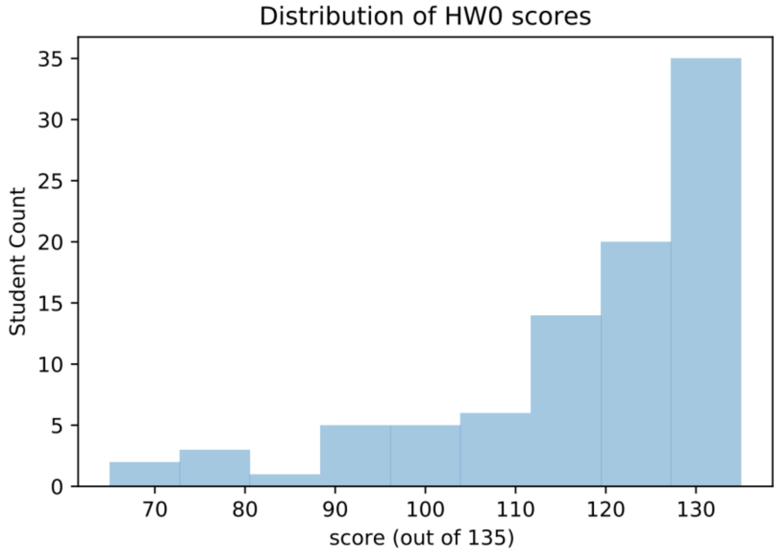
University of Illinois

Announcements

- HW 2 due tonight **Mar 2 @ 11:59 PM on Compass2G**
- MP2 Checkpoint 0.5 due tonight **Mar 2 @ 11:59 PM**
  - Submit via Google Form: https://forms.gle/uk4Meac85Va9HnAQ6
  - Provide update on work done since MP2 release
- ICA 3 this **Wed Mar 4** during class
  - Covers clustering and PCA
- Grad Students: Project proposal due this **Friday Mar 6 @ 11:59 PM** on Compass2G
  - Make sure to include all the requested components (listed in final project announcement on website)
- Midterm exam will take place on **Wed March 11th**
  - **Place TBD**
  - **One closed book no electronic devises (calc, laptops, phones, watches etc)**
  - **One 8X11 sheet**

HW 0 Grades

- Grade distribution for HW 0
  - Average: **118/135 points (87%)**
  - Standard deviation: **16 points (12%)**
- Lowest scoring questions
  - 5d: Finding a conditional PDF
  - 12: Comparing arrival time of two buses (uniform distributions)
    - c: Finding PDF and mean of later arrival time
    - d: Finding mean of earlier arrival time
    - e: Finding probability both buses are together at stop



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
- terms "feature" and "dimension" interchangeably
- Several ways to reduce dimension of the data
  - Drop unimportant dimensions using e.g. domain knowledge
  - 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 d (down 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 1st direction, the direction of maximum variance in the data
  - 2nd 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 Example: Food Habits

- Average consumption of 17 different types of food was tracked in 4 different countries in the UK.

- Measurements are reported in grams per person per week

- **Do any of the countries seem to have unusual consumption patterns?**

	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

http://setosa.io/ev/principal-component-analysis/

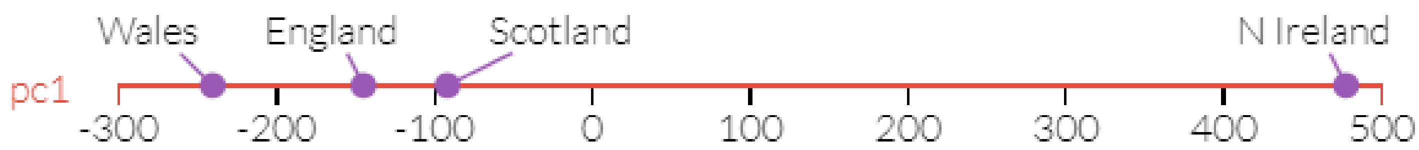
PCA Example: Food Habits

- In this setup, we have 17-dimensional data point $\boldsymbol{X} = (X_1, X_2, \ldots, X_{17})$
  - E.g. X_1=Alcoholic Drinks, X_2=Beverages, … , X_{17}=Sugars
- PCA reduces the number of dimensions of the data points by projecting each point onto different axes called principal components
  - Each successive principal component explains the maximum remaining variance in the data set, and is orthogonal to the other components
  - Each projection is a linear combination of the original features/dimensions
  - We refer to the projected points on the principal components as coordinates
  - In our example, the coordinate for the first principal component can be computed as

$$\begin{aligned}&-0.46X_1 - 0.026X_2 + 0.048X_3 - 0.048X_4 - 0.057X_5 - 0.030X_6 \\ &- 0.0052X_7 - 0.084X_8 - 0.63X_9 + 0.40X_{10} - 0.15X_{11} - 0.26X_{12} \\ &- 0.24X_{13} - 0.027X_{14} - 0.036X_{15} + 0.23X_{16} - 0.038X_{17}\end{aligned}$$

PCA Example: Food Habits

- We project each sample (17-D datapoint) onto the first principal component and plot the projections

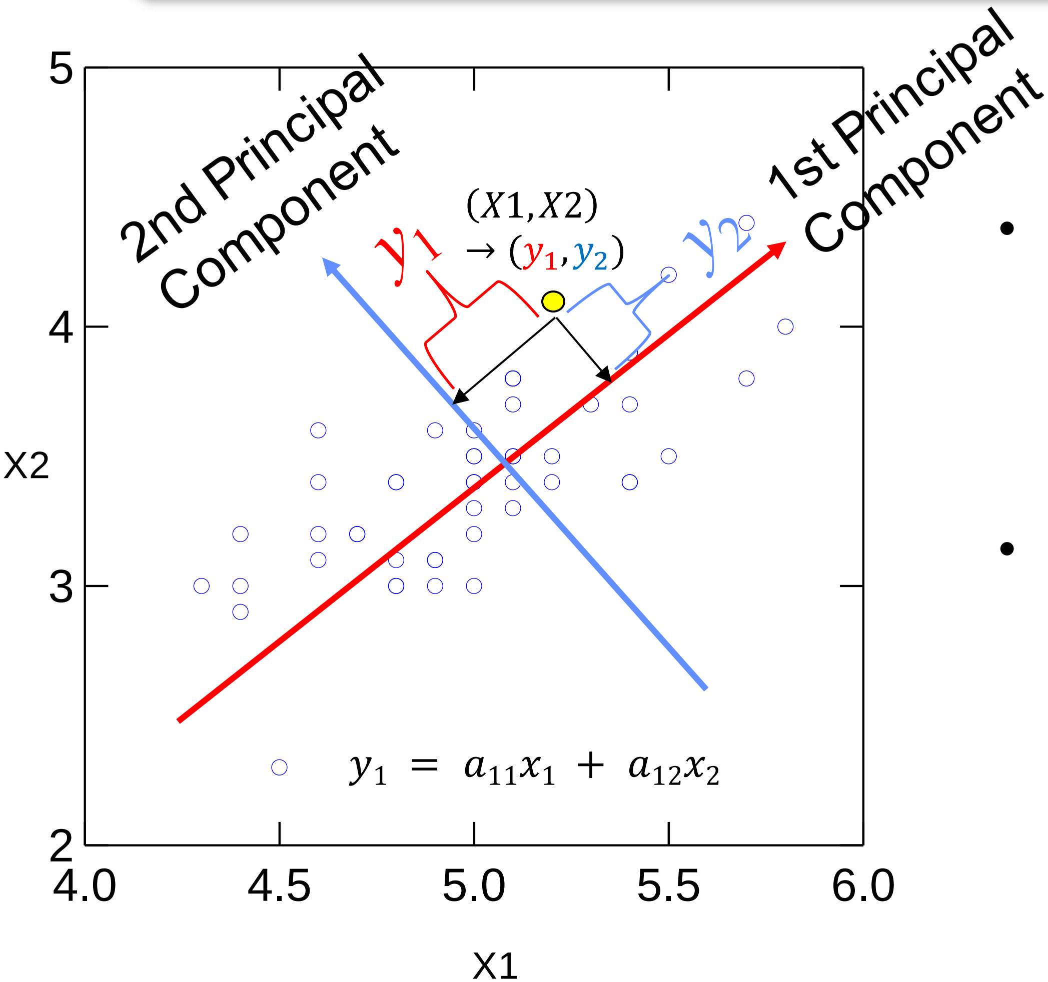


- From this plot, we can see that N Ireland’s food habits are notably different from those of the other UK countries.
  - This wasn’t as apparent from examining the raw data
  - Upon closer examination, N Ireland on average consumes more fresh potatoes and less fresh fruits, cheese, fish and alcoholic drinks
  - ? this makes sense since N Ireland

	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

http://setosa.io/ev/principal-component-analysis/

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) $X = [X_1, X_2, \ldots, X_p]$
- Produce *p* new variables: $y_1, y_2, \ldots, y_p$:

$$\left.\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p \\ &\ldots \\ y_p &= a_{p1}x_1 + a_{p2}x_2 + \ldots + a_{pp}x_p \end{aligned}\right\}$$

- y_j's are **principal components**
- $a_{j1}, a_{j2}, \ldots, a_{jp}$ are **regression coefficients**
- There are no intercepts (since we centered data)

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

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

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

PCA Applications

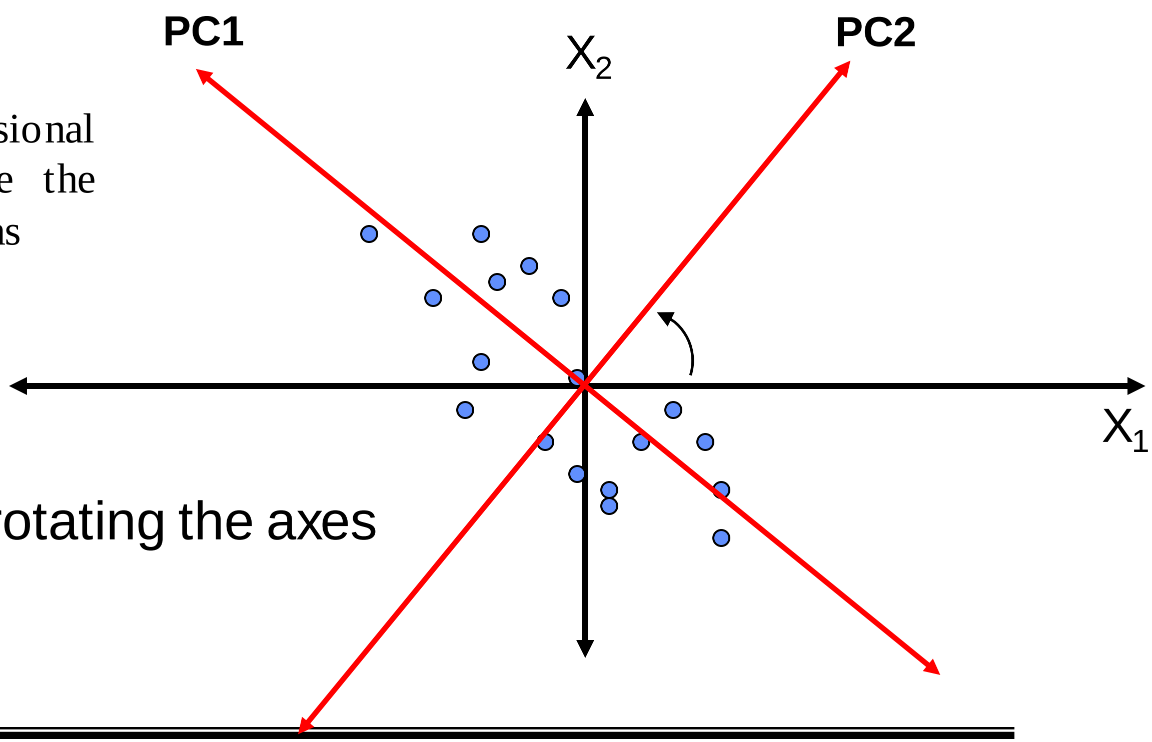
- Uses:
  - Data Visualization
  - Data Reduction
  - compression
  - Data Classification
  - Trend Analysis
  - Noise Reduction
  - Regression
  - Clustering

- Examples:
  - How to best present what is “interesting”?
  - Dimensionality reduction technique in domains like facial recognition, computer vision and image compression.
  - Finding patterns in data of high dimensions in finance, data mining, bioinformatics, psychology
  - How many unique “sub-sets” are in the sample?
  - How are they similar / different?
  - What measurements are needed to differentiate?
  - Which “sub-set” does a 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 to minimize the Euclidean distances among observations



This is accomplished by rotating the axes

Principal Component Analysis (eigenvalues and eigenvectors)

- Let M be either the correlation or covariance matrix of the original data
  - We will discuss later whether correlation or covariance matrix should be used for a dataset

- The **First Principal Component** $(a_{11},a_{12},...,a_{1p})$ is the eigenvector corresponding to the largest eigenvalue of M
  - The direction is specified by the normalized eigenvector
  - The magnitude is specified by the largest eigenvalue of M – **this reflects how much variance in the data is explained by this principal component**

- The **Second Principal Component** $(a_{21},a_{22},...,a_{2p})$ is the eigenvector corresponding to the second-largest eigenvalue of M

…

- The **p^{th} Principal Component** $(a_{p1},a_{p2},...,a_{pp})$ is the eigenvector corresponding to the p^{th}-largest eigenvalue of M

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

$$S = Covariance\,(X) = \frac{1}{n}(X - \bar{x})^T(X - \bar{x})$$

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

	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…
- An eigenvalue divided by the trace of S defines the percent of variance in the data explained by the principal component corresponding to that eigenvalue

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

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

Trace = 12.9091

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

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

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

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

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

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:

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

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

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

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

	a_1	a_2
X_1	1.0653	-0.9387
X_2	1	1

- Checking for orthogonality:

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

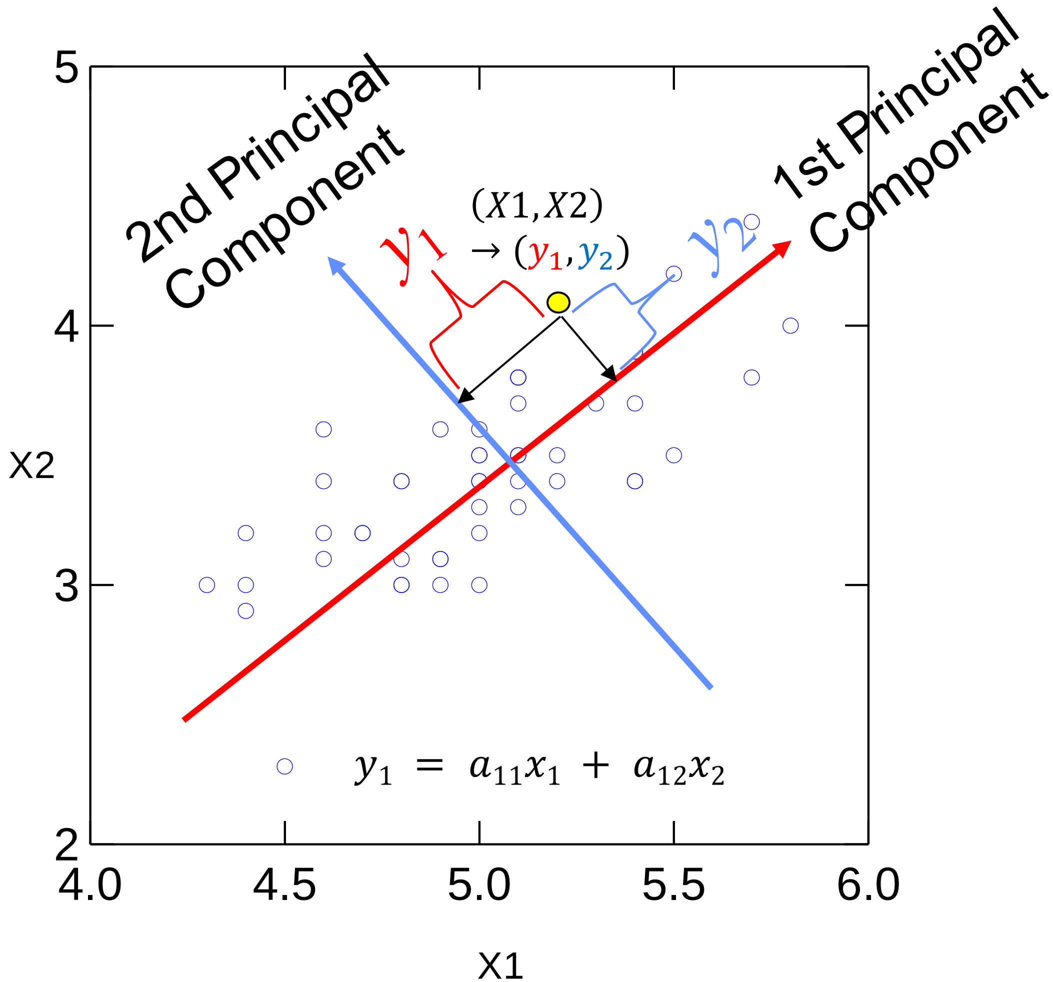
The Algebra of PCA

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

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

- 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.

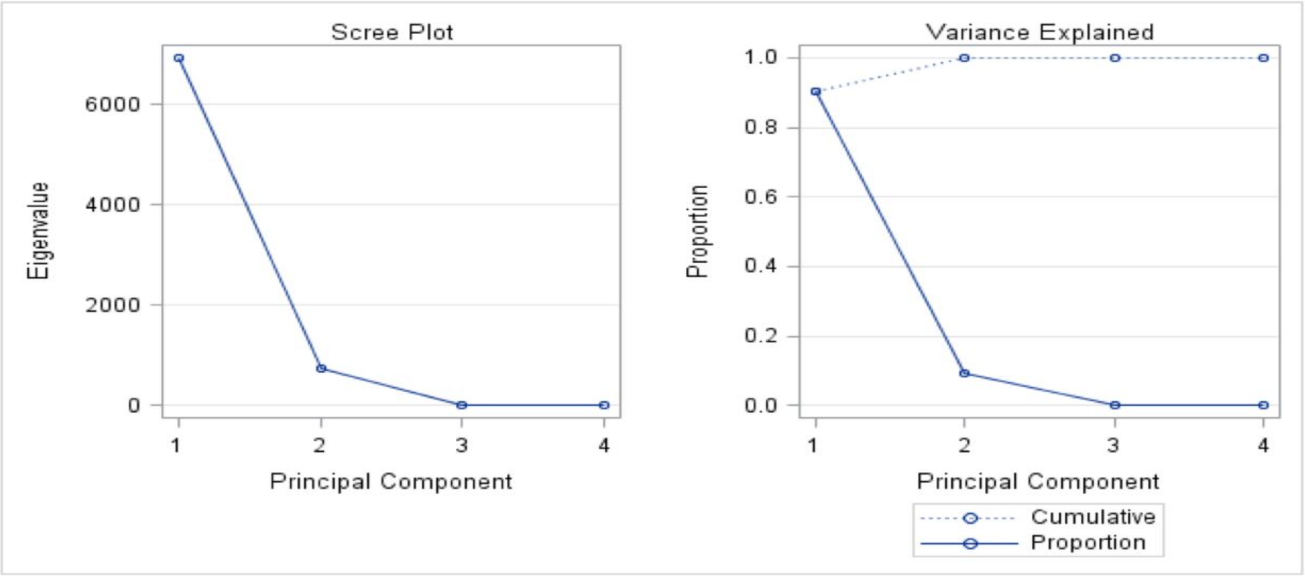
	PC_1	PC_2
PC_1	9.8783	0.0000
PC_2	0.0000	3.0308

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

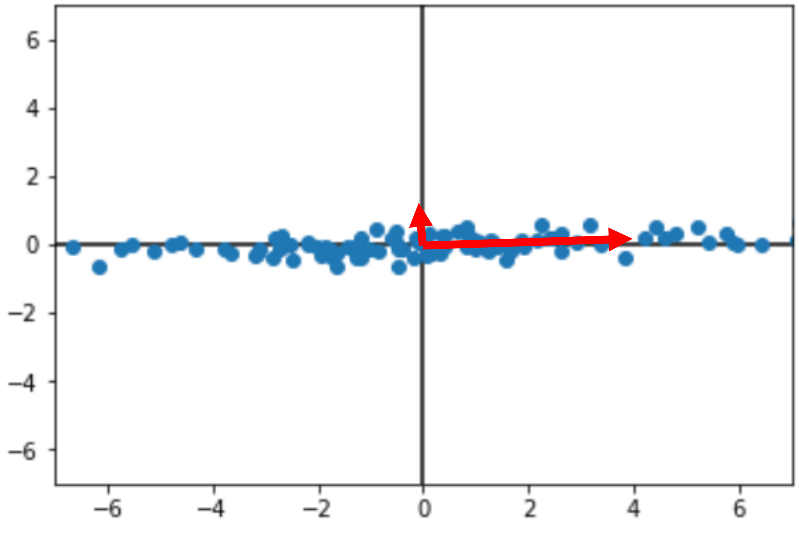
	X_1	X_2
X_1	1.0000	0.5297
X_2	0.5297	1.0000

Correlation Matrix

Trace (sum of diagonals): 12.9091

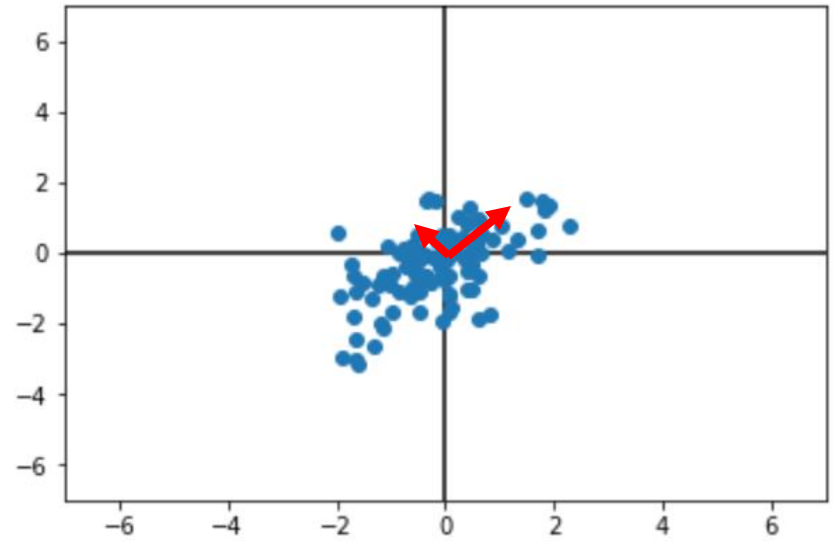
Trace (sum of diagonals): 2.0

The Algebra of PCA: Covariance/Correlation Matrix



$$\begin{bmatrix} 10 & 0.5 \\ 0.5 & 0.1 \end{bmatrix}$$

Covariance matrix



$$\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

Correlation matrix

- If variance of features is not on comparable scale, then principal components have high contribution from features with large variance

PCA with Correlation Matrix

- Compute correlation matrix from covariance matrix:

Correlation between variables *i* and *j*

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

Covariance of variables *i* and *j*

Variance of variable *j*

- Solve eigenvalue equation: $S_{cor} a = \lambda a$

Correlation Matrix

- Compute eigenvalues by solving: $|S_{cor} - \lambda I| = 0$

- Compute eigenvectors (principal components) by solving the following for each eigenvalue λ_i: $(S_{cor} - \lambda_i I) a_i = 0$

- Principal components may be different for correlation matrix and covariance matrix

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.

Additional Slides

Eigenvalues and Eigenvectors

Mahalanobis Distance

- Recall that when calculating Mahalanobis distance, we transformed and rescaled the datapoints before calculating the Euclidean distance between them
  - Transformation was done to eliminate covariance between distinct features
  - Rescaling was done so that each feature has variance of 1

- We were able to accomplish this as follows:
  - Transformation: Use the eigenvectors of the covariance matrix as the new axes
  - Rescaling: Scale each new axis i by the respective eigenvalue $(1/\sqrt{\lambda_i})$

Eigenvalues and Eigenvectors: Alternate Interpretation

- A matrix S has an eigenvalue λ_i with corresponding eigenvector $\boldsymbol{a_i}$ if the following holds true:

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

- For example, suppose you have a matrix $S = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}$, and you know that one of its two eigenvectors is $\boldsymbol{a_1} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$. Then you can solve for λ_1:

$$S\boldsymbol{a_i} = \lambda_1 \boldsymbol{a_i} \Rightarrow \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \lambda_1 \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

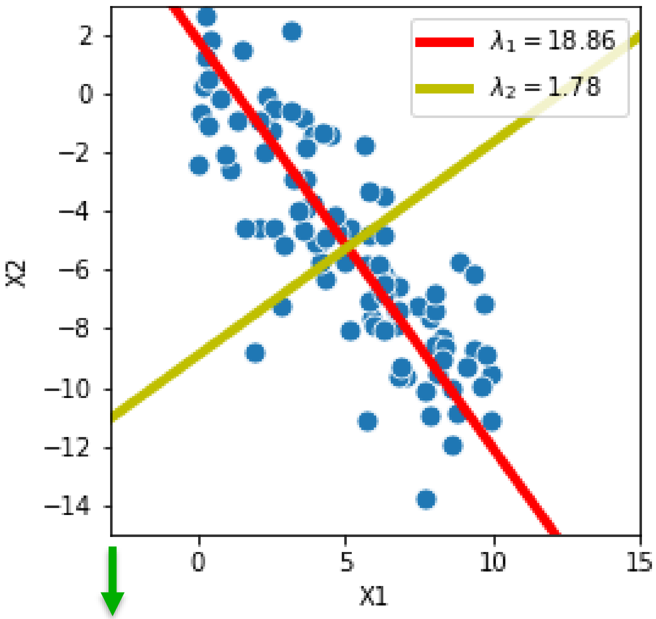
$$\Rightarrow \begin{bmatrix} 12 \\ 8 \end{bmatrix} = \lambda_1 \begin{bmatrix} 3 \\ 2 \end{bmatrix} \Rightarrow \lambda_1 = 4$$

- Typically, the eigenvectors are scaled to have unit length. In our example,

$$\widehat{\boldsymbol{a_1}} = \frac{a_1}{|a_1|} = \frac{1}{\sqrt{3^2 + 2^2}} \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 3/\sqrt{13} \\ 2/\sqrt{13} \end{bmatrix}$$

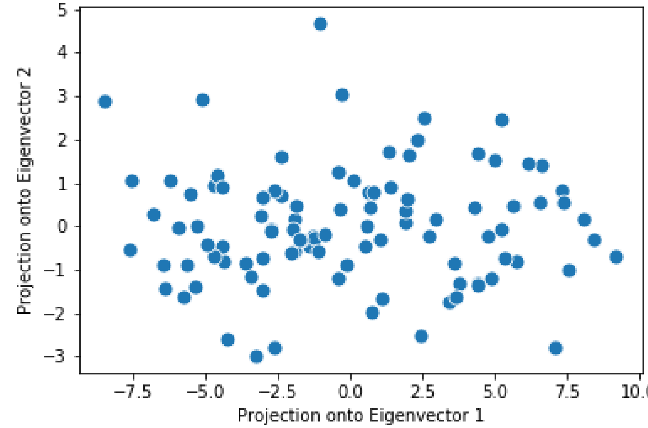
- We can double check: $\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 3/\sqrt{13} \\ 2/\sqrt{13} \end{bmatrix} = \begin{bmatrix} 12/\sqrt{13} \\ 8/\sqrt{13} \end{bmatrix} = 4 \begin{bmatrix} 3/\sqrt{13} \\ 2/\sqrt{13} \end{bmatrix}$

Another Example



Use eigenvectors as new coordinate axes

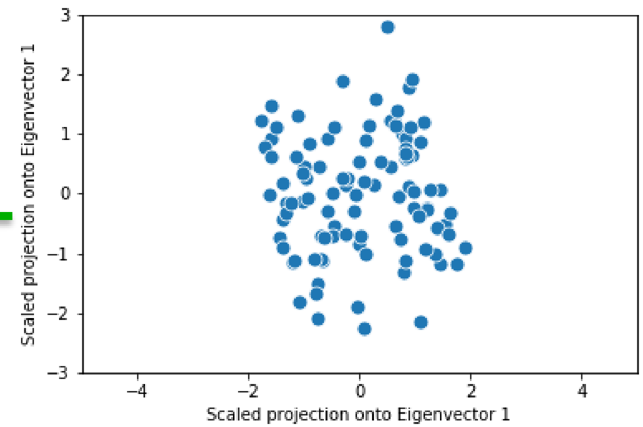
Removes covariance



Shrink each axis by $\sqrt{\lambda_i}$
λ_i = eigenvalue

Standard deviation set to 1 along each axis

Correlation (X1, X2) = - 0.81
Var (X1) = 7.56
Var (X2) = 12.87



Correlation between transformed projections = 0
Variation along eigenvector1 = 1
Variation along eigenvector2 = 1

Covariance vs Correlation in PCA

Covariance vs Correlation

- It's often useful to analyze how two variables change together
  - Doing so provides insight on the relationship between the variables' behavior in a system
- Commonly, the linear relationship between the two variables is examined (i.e. look at Y vs X instead of Y vs X^2)
- Two frequently used statistical measures to quantify such relationships are <u>covariance</u> and <u>correlation</u>, which handle variable scaling differently

Variable Scaling

- In order to identify how variables change, we need to consider the *scaling* of both variables
- Suppose we have two vectors A and B, each containing measurement data for a different feature
  - If the units of measurements for A and B are comparable, then directly comparing $range(A)$ and $range(B)$ can tell us which feature changes more
    - e.g. If A and B contain height measurements (in cm) over two years for Alice and Bob respectively, then knowing $|range(A)| > |range(B)|$ implies that Alice grew more over the two year period
  - However, if A and B have incomparable units, then the magnitude of the ranges for each vector doesn't necessarily convey information about which variable changed more
    - e.g. If $A = [150, 156, 160]$ contains Alice's height measurements in cm over two years, and $B = [1.3, 1.4, 1.6]$ contains Bob's height measurements in meters over two years, then Bob grew more than Alice even though $|range(A)| = 10 > |range(B)| = 0.3$

Covariance vs Correlation

- **<u>Covariance</u>**: How much do two different variables vary together?
  - Sensitive to scaling of both variables
  - Unbounded
  - Unit of covariance is product of units of both variables
- **<u>Correlation</u>**: When does a change in one variable result in a change of the other variable?
  - Normalized value of covariance – not affected by variable scaling
  - Bounded between -1 and 1
  - Correlation is unitless

Covariance

- **<u>Covariance</u>**: How much do two different variables vary together?
  - Sensitive to scaling of both variables
  - Unbounded
  - Unit of covariance is product of units of both variables
- Let
  - X and Y be n-dimensional variables
  - $\bar{X} = E[X]$
  - $\bar{Y} = E[Y]$

  Then,

$$Covariance\,(X,Y) = \frac{1}{n}\sum_{j\in\{1,2,\dots,n\}}\sum_{i\in\{1,2,\dots,n\}}(X_i-\bar{X})(Y_j-\bar{Y})$$

Correlation

- **Correlation**: When does a change in one variable result in a change of the other variable?
  - Normalized value of covariance – not affected by variable scaling
  - Bounded between -1 and 1
    - Positive value ⇒ increase in one variable results in increase of other
    - Negative value ⇒ increase in one variable results in decrease of other
    - Zero value ⇒ no linear relationship between variables
  - Correlation is unitless
- Let
  - X and Y be n-dimensional variables
  - $Cov(X, Y)$ be the covariance of X and Y
  - σ_X and σ_Y be the standard deviations of of X and Y respectively

  Then,

$$Correlation\ (X, Y) = \rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Correlation vs Covariance in PCA

- It's up to the analyst to decide whether it is more appropriate to use the correlation or covariance matrices during PCA analysis
- Using the covariance matrix is only appropriate if both of the following are true:
  - (i) Variables all have same units
  - (ii) You wish to emphasize variables with the most variance
- Using the correlation matrix is appropriate if
  - (i) Variables are reported in different units
  - (ii) Variables are reported in same units, but you wish to emphasize them all equally during PCA
- Since real-world variables are usually not reported in the same units, the correlation matrix is typically used more for PCA

PCA Example: New Cars

- We have a real dataset on 387 new cars that were introduced in the year 2004
- For each car, we have the 11 features shown to the right
- Notice that the features with the largest standard deviation are **retail price**, **invoice price**, and **weight**

Feature	Units	Standard Deviation
Retail price	US dollars	19699.13
Invoice price	US dollars	17878.04
Engine size	Liters	1.01
Number of cylinders	-	1.49
Power	Horsepower	70.17
City fuel efficiency	Miles/Gallon	5.26
Highway fuel efficiency	Miles/Gallon	5.63
Weight	Pounds	705.09
Wheelbase	Inches	7.08
Length	Inches	13.22
Width	Inches	3.36

http://jse.amstat.org/jse_data_archive.htm

PCA Example

- We perform PCA on the dataset separately with the covariance and correlation matrices
- When using the covariance matrix, the features with the highest variance (retail price, invoice price, and weight) are emphasized the most
  - Their component scores are multiple orders of magnitudes higher
  - As a result of this, the projections are much more spread out
  - Notice the larger scales on the pc1 and pc2 axes in the first graph!
- In this case, since the features have different units, PCA should be performed with the correlation matrix