

Analiza wariancji i metody klasyfikacyjne **Analysis of variance and classification** **methods**

lecture 6

25 November 2019

Ilona Anna Urbaniak (PK)

Marcin Wolter (IFJ PAN)

e-mail: marcin.wolter@ifj.edu.pl, phone: 12 662 8024

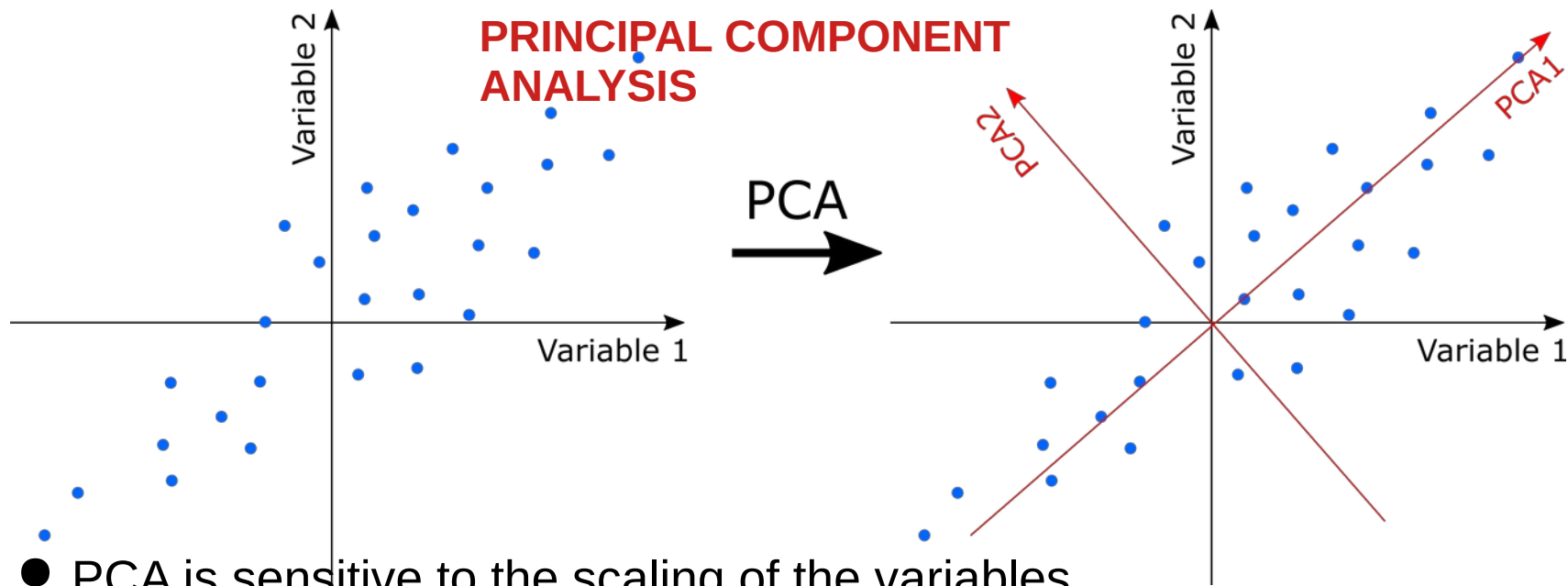
Slides: <https://indico.ifj.edu.pl/event/271/>

Principal Component Analysis PCA

Analiza składowych głównych

PCA summary

- Suppose we have a population measured on p random variables X_1, \dots, X_p .
- Goal: a new set of p axes (linear combinations of the original p axes) in the directions of greatest variability:



- PCA is sensitive to the scaling of the variables.

Two examples in python

- **principal_component_analysis.ipynb**
 - Principal component analysis on famous IRIS dataset
 - PCA is done once manually and once using sklearn package
 - Sklearn is a machine learning package
- **plot_digits_simple_classif.ipynb**
 - Analyze hand-written digits - 8x8 pixel maps
 - PCA performed on 64 input variables
 - Naive Bayes method used for classification on n first principal components
 - Digits visualized on 2D space

Did you manage to run these examples using other datasets?

Just two points

mean subtracted : $x_1 = -x_2 = x$

$y_1 = -y_2 = y$

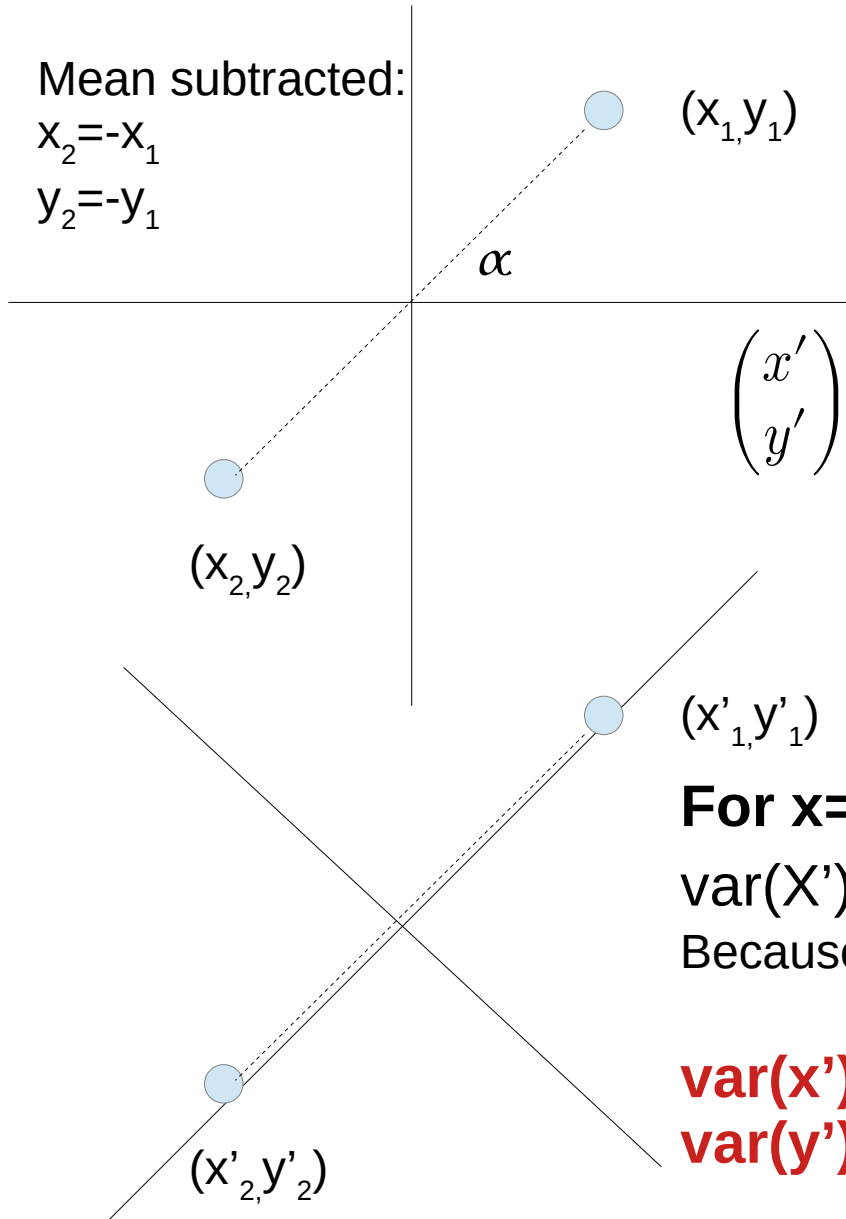
$\text{var}(X) = x^2$

$\text{var}(Y) = y^2$

Mean subtracted:

$x_2 = -x_1$

$y_2 = -y_1$



After rotation by an angle α :

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \cos(\alpha) + y \sin(\alpha) \\ x \sin(\alpha) - y \cos(\alpha) \end{pmatrix}$$

$$\text{Var}(X') = x'^2 = (x \cos(\alpha) + y \sin(\alpha))^2$$

$$\text{Var}(Y') = y'^2 = (-x \sin(\alpha) + y \cos(\alpha))^2$$

For $x=y$ maximum $\text{var}(X')$ at $\alpha=45^\circ$

$$\text{var}(X') = x^2(\cos(\alpha) + \sin(\alpha))^2 = x^2(1 + \sin(2\alpha))$$

Because:

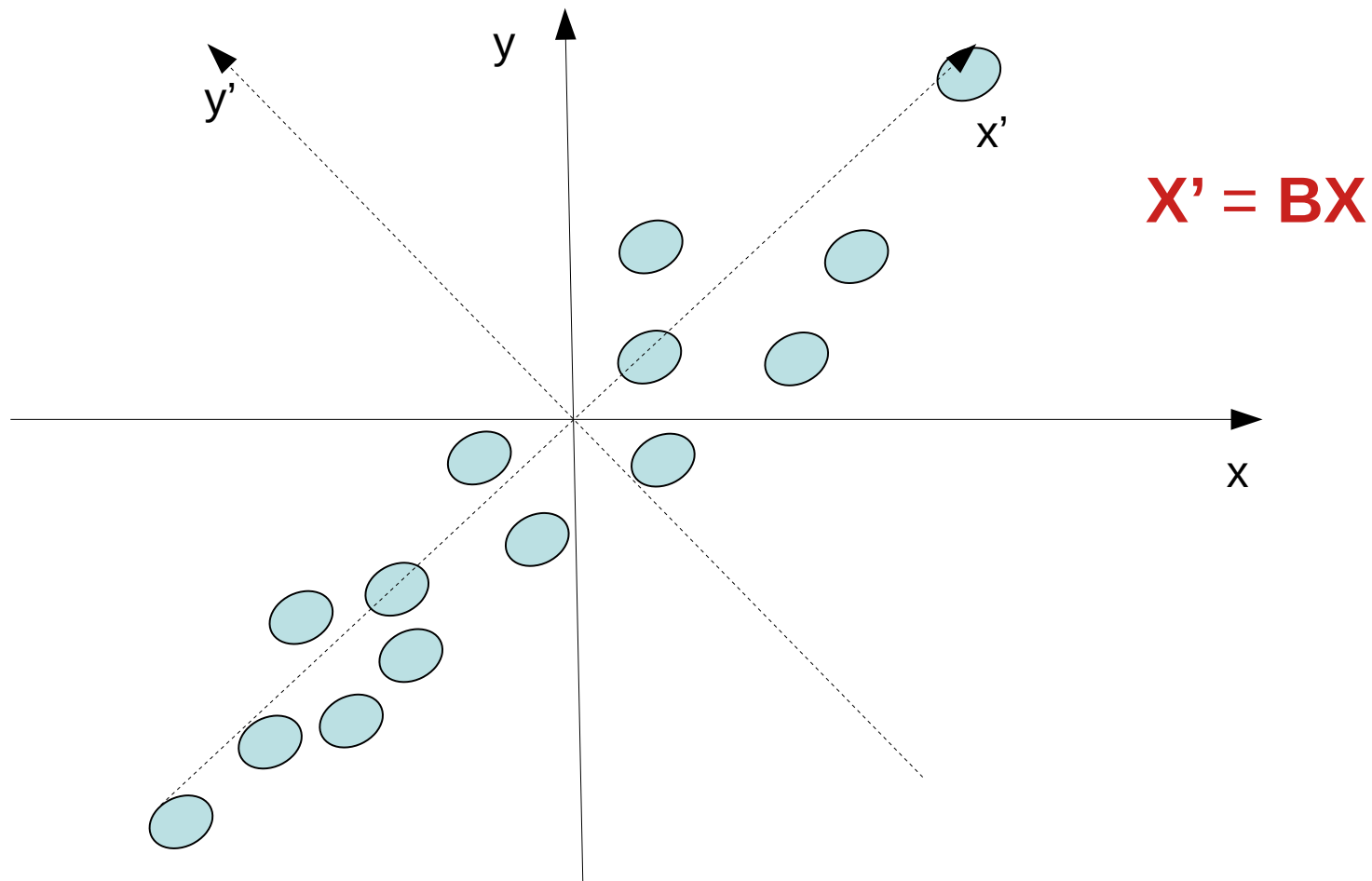
$$\sin(2\theta) = 2 \sin \theta \cos \theta = \frac{2 \tan \theta}{1 + \tan^2 \theta}$$

$$\text{var}(x') = 2 \text{Var}(x)$$

$$\text{var}(y') = 0$$

Formulate the problem

- We have a set of data points \mathbf{X} having a covariance matrix Σ
- The transformation (rotation) \mathbf{B} transforms \mathbf{X} to \mathbf{X}' : $\mathbf{X}' = \mathbf{B}\mathbf{X}$



So some calculations

- Σ – covariance matrix of a data \mathbf{X}
- Σ has the eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ and associated eigenvectors are $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_p$.
- $\mathbf{X}' = \mathbf{B}\mathbf{X}$ – transformation of \mathbf{X} to the new coordinate system
- thus covariance $\text{Cov}(\mathbf{x}'_1) = \text{Cov}(\mathbf{B}_{11}\mathbf{x}_1 + \dots + \mathbf{B}_{1p}\mathbf{x}_p) = \mathbf{B}_1^T \Sigma \mathbf{B}_1$, where $\mathbf{B}_1 = (\mathbf{B}_{11}, \mathbf{B}_{12}, \dots, \mathbf{B}_{1p})$

We know that
(see next slide) $\max_{\mathbf{x} \neq 0} \frac{\mathbf{x}^T \mathbf{B} \mathbf{x}}{\mathbf{x}^T \mathbf{x}} = \lambda_1$ (attained when $\mathbf{x} = \mathbf{e}_1$)

So: $\max_{\mathbf{B}_1 \neq 0} \frac{\mathbf{B}_1^T \Sigma \mathbf{B}_1}{\mathbf{B}_1^T \mathbf{B}_1} = \lambda_1$ (\mathbf{B}_1 eigenvector of Σ , λ_1 eigenvalue)

$$\lambda_1 = \frac{\mathbf{e}_1^T \Sigma \mathbf{e}_1}{\mathbf{e}_1^T \mathbf{e}_1} = \mathbf{e}_1^T \Sigma \mathbf{e}_1 = \text{Var}(\mathbf{X}'_1)$$

Def. of λ
 $\mathbf{e}_1^T \mathbf{e}_1 = 1$

What we wanted to show!

Max. variance of $\mathbf{X}'_1 = \lambda_1$ - 1st eigenvalue of covariance matrix Σ ,

The 1st PCA axis is the eigenvector \mathbf{e}_1 of covariance matrix Σ

Lemma

Maximization of Quadratic Forms. Let \mathbf{B} ($p \times p$) be a positive definite matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ and associated normalized eigenvectors are $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_p$. Then:

$$\max_{\mathbf{x} \neq 0} \frac{\mathbf{x}^T \mathbf{B} \mathbf{x}}{\mathbf{x}^T \mathbf{x}} = \lambda_1 \quad (\text{attained when } \mathbf{x} = \mathbf{e}_1)$$

$$\min_{\mathbf{x} \neq 0} \frac{\mathbf{x}^T \mathbf{B} \mathbf{x}}{\mathbf{x}^T \mathbf{x}} = \lambda_p \quad (\text{attained when } \mathbf{x} = \mathbf{e}_p)$$

Proof: Let \mathbf{P} ($p \times p$) be the orthogonal matrix whose columns are the eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_p$ and $\mathbf{\Lambda}$ be the diagonal matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ along the main diagonal. Let $\mathbf{B}^{1/2} = \mathbf{P} \mathbf{\Lambda}^{1/2} \mathbf{P}^T$ and $\mathbf{y} = \mathbf{P}^T \mathbf{x}$ (sizes: $\mathbf{y}(p \times 1), \mathbf{x}(p \times 1), \mathbf{P}^T(p \times p)$).

Consequently, $\mathbf{x} \neq \mathbf{0}$ implies $\mathbf{y} \neq \mathbf{0}$. Thus,

$$\begin{aligned} \frac{\mathbf{x}^T \mathbf{B} \mathbf{x}}{\mathbf{x}^T \mathbf{x}} &= \frac{\mathbf{x}^T \mathbf{B}^{1/2} \mathbf{B}^{1/2} \mathbf{x}}{\mathbf{x}^T \mathbf{P} \mathbf{P}^T \mathbf{x}} \\ &= \frac{\mathbf{x}^T \mathbf{P} \mathbf{\Lambda}^{1/2} \mathbf{P}^T \mathbf{P} \mathbf{\Lambda}^{1/2} \mathbf{P}^T \mathbf{x}}{\mathbf{y}^T \mathbf{y}} = \frac{\mathbf{y}^T \mathbf{\Lambda} \mathbf{y}}{\mathbf{y}^T \mathbf{y}} \\ &= \frac{\sum_{i=1}^p \lambda_i y_i^2}{\sum_{i=1}^p y_i^2} \leq \lambda_1 \frac{\sum_{i=1}^p y_i^2}{\sum_{i=1}^p y_i^2} = \lambda_1 \end{aligned}$$

Just two points again

- Try to do it using matrix calculations

$$\text{Data : } B = \begin{vmatrix} x & y \\ -x & -y \end{vmatrix}$$

$$\text{Covariance : } \text{Cov}(B) = \frac{1}{N} B^T B = \frac{1}{2} \begin{vmatrix} 2x^2 & 2xy \\ 2xy & 2y^2 \end{vmatrix} = \begin{vmatrix} x^2 & xy \\ xy & y^2 \end{vmatrix}$$

$$\text{Eigenvalues : } \text{Det}(\text{Cov}(B) - \lambda I) = \text{Det} \left(\begin{vmatrix} x^2 - \lambda & xy \\ xy & y^2 - \lambda \end{vmatrix} \right) = 0$$

$$(x^2 - \lambda)(y^2 - \lambda) - x^2 y^2 = 0 \implies \lambda_1 = x^2 + y^2, \lambda_2 = 0$$

Eigenvalues are: $\lambda_1 = x_1^2 + y_1^2, \lambda_2 = 0$

These eigenvalues are our two variances we have found in the previous example!

Corresponding eigenvectors:
(modulo normalization) $e_1 = \begin{pmatrix} 1/y \\ 1/x \end{pmatrix} \quad e_2 = \begin{pmatrix} 1/x \\ -1/y \end{pmatrix}$

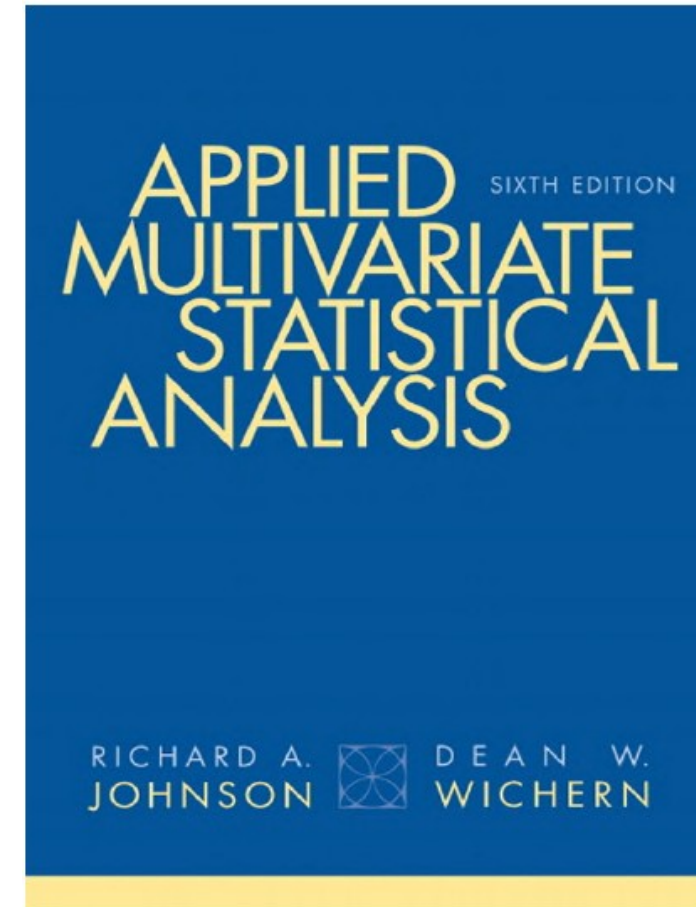
Bessel's correction

- In statistics, Bessel's correction is the use of $n - 1$ instead of n in the formula for the sample variance and sample standard deviation, where n is the number of observations in a sample.
- This method corrects the bias in the estimation of the population variance. It also partially corrects the bias in the estimation of the population standard deviation. However, the correction often increases the mean squared error in these estimations. This technique is named after Friedrich Bessel.
- In estimating the population variance from a sample when the population mean is unknown, the uncorrected sample variance is the mean of the squares of deviations of sample values from the sample mean (i.e. using a multiplicative factor $1/n$). In this case, the sample variance is a biased estimator of the population variance.

Algebraic Interpretation

Nicely explained in:

<http://docshare04.docshare.tips/files/12598/125983744.pdf>



Steps of PCA

Calculate the empirical mean

Find the empirical mean along each column $j = 1, \dots, p$.

Place the calculated mean values into an empirical mean vector u of dimensions $p \times 1$.

$$u_j = \frac{1}{n} \sum_{i=1}^n X_{ij}$$

Calculate the deviations from the mean

Subtract the mean from each data point

Find the covariance matrix

Find the $p \times p$ empirical covariance matrix \mathbf{C} from matrix \mathbf{B} (data):

$$\mathbf{C} = \frac{1}{N} \mathbf{B}^T \mathbf{B}$$

Find eigenvalues by solving:

$$\det(\mathbf{C} - \lambda \mathbf{1}) = 0$$

This means solving a characteristic polynomial.

Steps of PCA

Find the eigenvectors and eigenvalues of the covariance matrix

Compute the matrix **V** of eigenvectors which diagonalizes the covariance matrix **C**:

$$\mathbf{C}\vec{v}_i = \lambda_i \vec{v}_i$$

Rearrange the eigenvectors and eigenvalues

Sort the columns of the eigenvector matrix **V** and eigenvalue matrix **D** in order of decreasing eigenvalue.

Select a subset of the eigenvectors as basis vectors

The goal is to choose a value of **L** as small as possible while achieving a reasonably high value of **g** on a percentage basis.

Project the z-scores of the data onto the new basis

You have reduced the dimensionality of your explaining as much variance as possible.

Steps of PCA

Find how much of variance is explained by n first principal components

In case of PCA, "variance" means *summative variance* or *multivariate variability* or *overall variability* or *total variability*. Below is the covariance matrix of some 3 variables. Their variances are on the diagonal, and the sum of the 3 values (3.448) is the overall variability.

1.343730519	-.160152268	.186470243
-.160152268	.619205620	-.126684273
.186470243	-.126684273	1.485549631

Now, PCA replaces original variables with new variables, called principal components, which are orthogonal (i.e. they have zero covariations) and have variances (called eigenvalues) in decreasing order. So, the covariance matrix between the principal components extracted from the above data is this:

1.651354285	.0000000000	.0000000000
.0000000000	1.220288343	.0000000000
.0000000000	.0000000000	.576843142

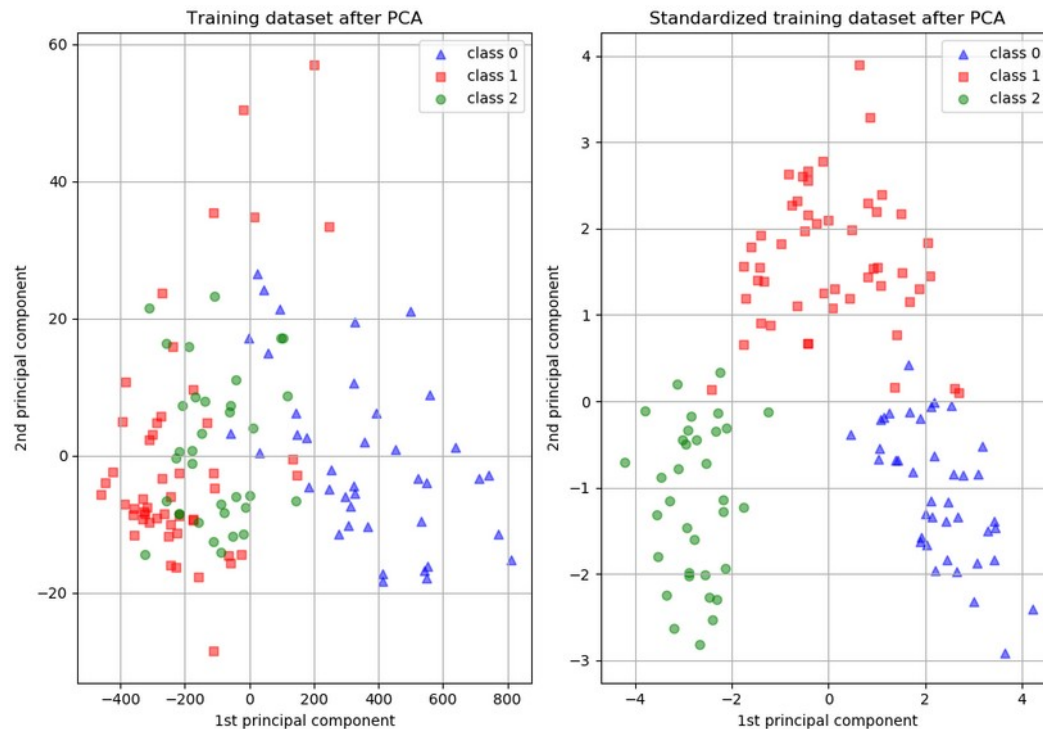
Note that the diagonal sum is still 3.448, which says that all 3 components account for all the multivariate variability. The 1st principal component accounts for or "explains" $1.651/3.448 = 47.9\%$ of the overall variability; the 2nd one explains $1.220/3.448 = 35.4\%$ of it; the 3rd one explains $.577/3.448 = 16.7\%$ of it.

Steps of PCA

Give some interpretation to the principal components

What would happen If I did PCA without normalization? Why do we normalize data?

In PCA we are interested in the components that maximize the variance. If one component (e.g. human height) varies less than another (e.g. weight) because of their respective scales (meters vs. kilos), PCA might determine that the direction of maximal variance more closely corresponds with the 'weight' axis, if those features are not scaled. As a change in height of one meter can be considered much more important than the change in weight of one kilogram, this is clearly incorrect.



The dataset used is the Wine Dataset available at UCI. This dataset has continuous features that are heterogeneous in scale due to differing properties that they measure (i.e alcohol content, and malic acid).

Correlation matrix

The **correlation matrix** refers to the symmetric array of numbers

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{21} & 1 & r_{23} & \cdots & r_{2p} \\ r_{31} & r_{32} & 1 & \cdots & r_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & r_{p3} & \cdots & 1 \end{pmatrix}$$

where

$$r_{jk} = \frac{s_{jk}}{s_j s_k} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \sqrt{\sum_{i=1}^n (x_{ik} - \bar{x}_k)^2}}$$

is the Pearson correlation coefficient between variables \mathbf{x}_j and \mathbf{x}_k .

Correlation or covariance matrix?

- Mean-centering is unnecessary if performing a principal components analysis on a correlation matrix, as the data are already centered after calculating correlations.
- We tend to use the covariance matrix when the variable scales are similar and the correlation matrix when variables are on different scales.

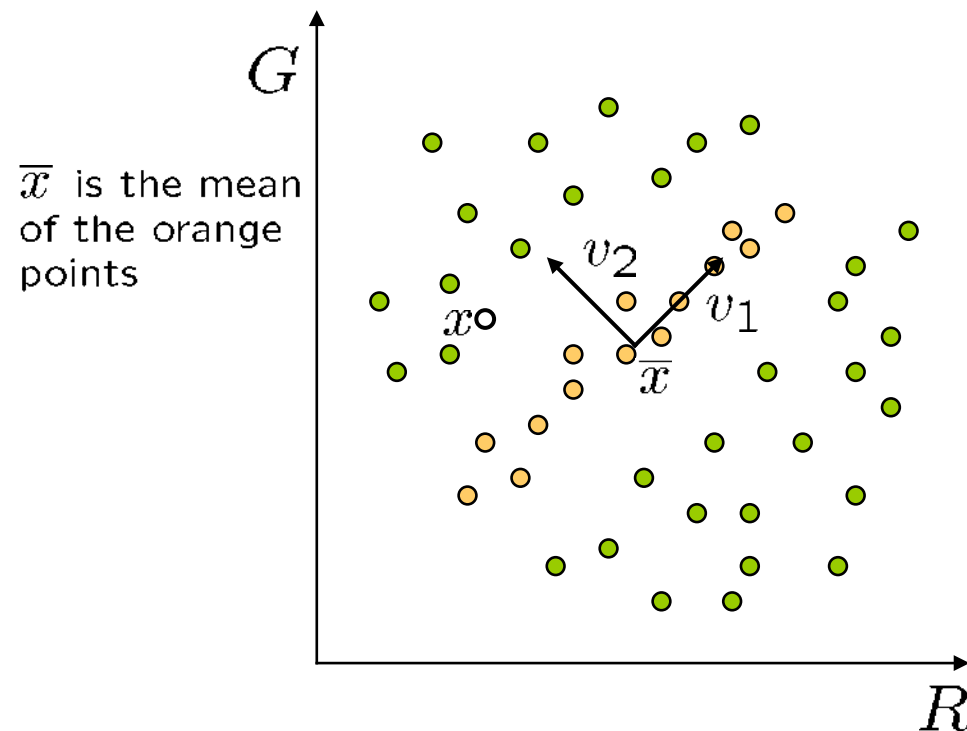
$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}$$

↓
↓
↓

Correlation between X and Y
 Standard deviation of X
 Standard deviation of Y

} Covarianced normalized by Standard Deviation

Dimensionality Reduction



- Dimensionality reduction
 - We can represent the orange points with *only* their \mathbf{v}_1 coordinates
 - since \mathbf{v}_2 coordinates are all essentially 0
 - This makes it much cheaper to store and compare points
 - A bigger deal for higher dimensional problems

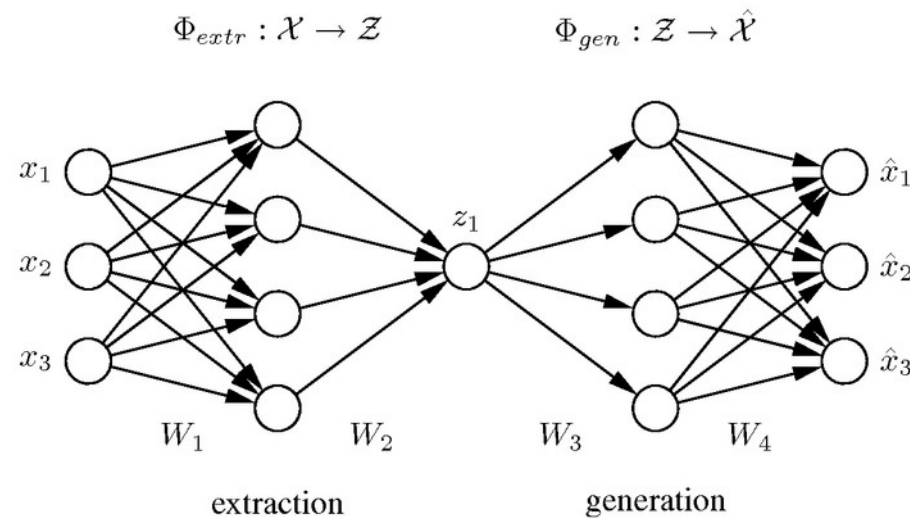
Higher Dimensions

- Suppose each data point is N-dimensional
 - Same procedure applies:

$$\begin{aligned} \text{var}(\mathbf{v}) &= \sum_{\mathbf{x}} \|(\mathbf{x} - \bar{\mathbf{x}})^T \cdot \mathbf{v}\|^2 \\ &= \mathbf{v}^T \mathbf{A} \mathbf{v} \quad \text{where } \mathbf{A} = \sum_{\mathbf{x}} (\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T \end{aligned}$$

- The eigenvectors of \mathbf{A} define a new coordinate system
 - eigenvector with largest eigenvalue captures the most variation among training vectors \mathbf{x}
 - eigenvector with smallest eigenvalue has least variation
- We can compress the data by only using the top few eigenvectors
 - corresponds to choosing a “linear subspace”
 - represent points on a line, plane, or “hyper-plane”
 - these eigenvectors are known as the ***principal components***

Non-linear PCA

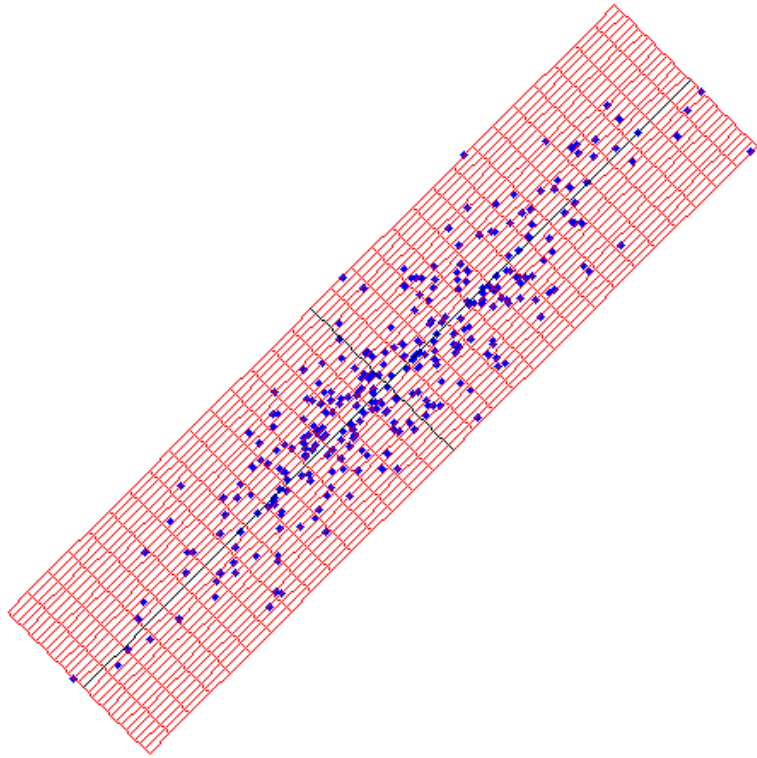


Nonlinear principal component analysis (NLPCA) is a nonlinear generalization of PCA. It generalizes the principal components from straight lines to curves (nonlinear).

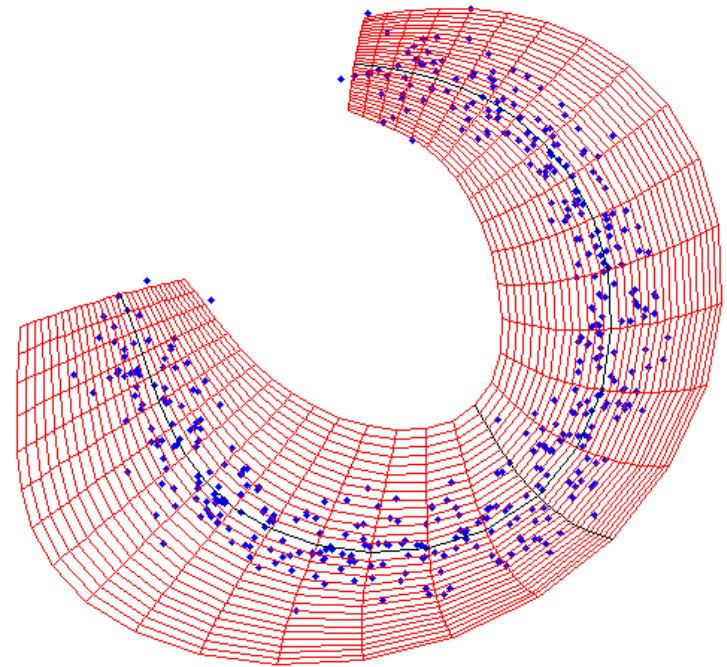
Nonlinear PCA can be achieved by using a neural network with an autoassociative architecture also known as autoencoder. Such autoassociative neural network is a multi-layer perceptron that performs an identity mapping, meaning that the output of the network is required to be identical to the input.

In the middle of the network is a layer that works as a bottleneck in which a reduction of the dimension of the data is enforced. This bottleneck-layer provides the desired component values.

Non-linear PCA



Linear PCA

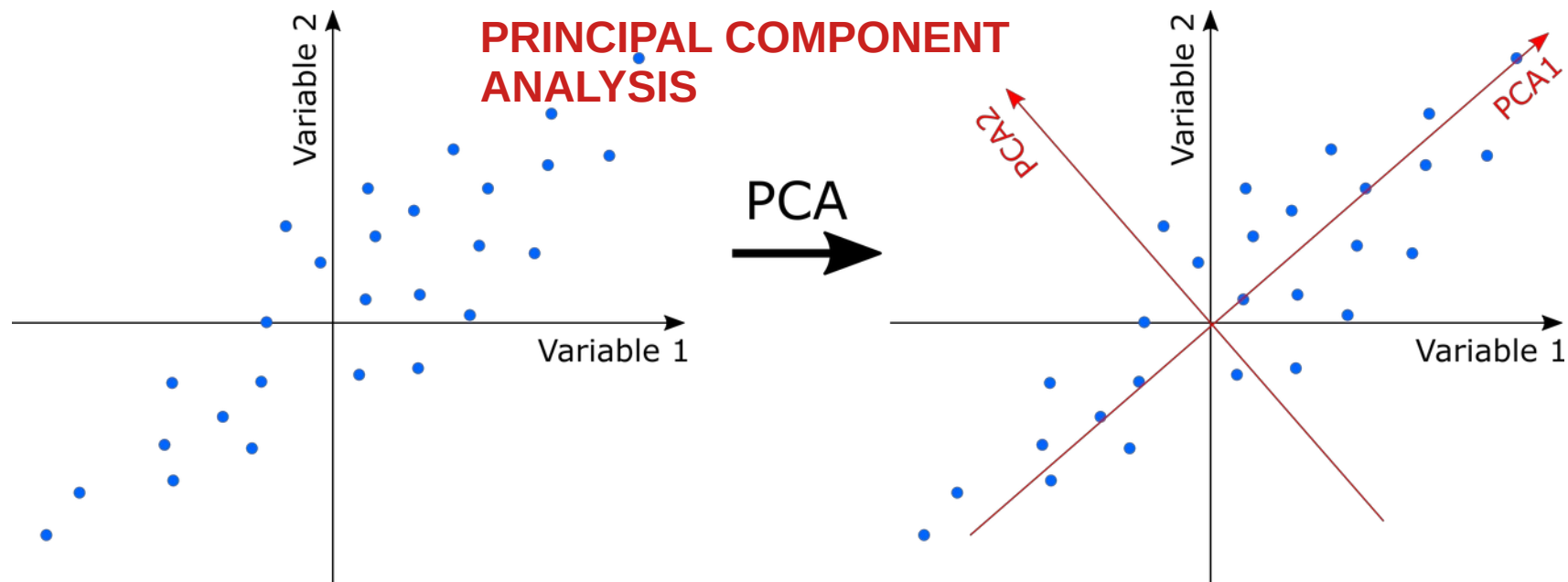


Non-linear PCA

Summary

Principal Component Analysis:

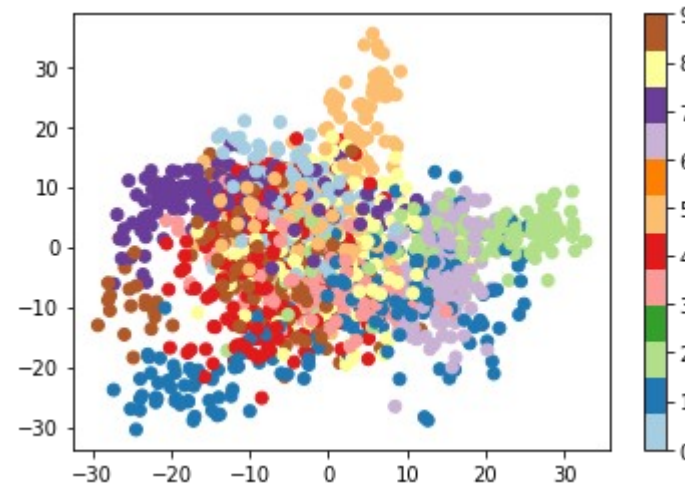
- Idea: rotate the axes to get maximal variance at the first axes, then second maximal at the second etc.



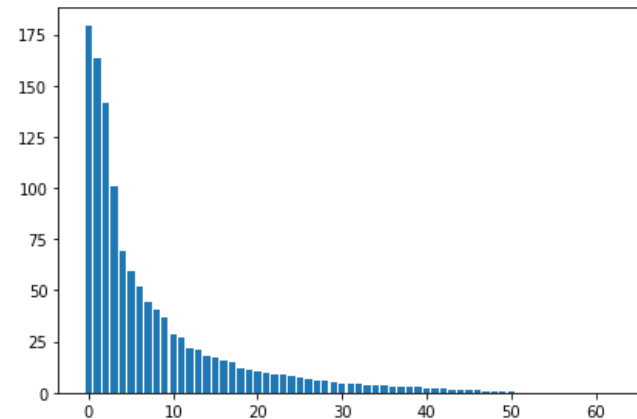
- Drop the least important axes to reduce the data dimensionality losing as little information as possible.
- Principal components (eigenvalues and eigenvectors) can be found analytically

Summary

- PCA helps to visualize the multidimensional data in 2D:



- Few components can explain most of the variance:



- Further analysis / classification might be much easier.

A super-simple example

Data:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \left\{ \begin{pmatrix} -20 \\ -8 \end{pmatrix}, \begin{pmatrix} -10 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 10 \\ 1 \end{pmatrix}, \begin{pmatrix} 20 \\ 8 \end{pmatrix} \right\}$$

Calculate the means:

$$E(x) = 0, \quad E(y) = 0$$

Covariance matrix:

$$\mathbf{C} = \frac{1}{5} \mathbf{X} \mathbf{X}^T = \begin{pmatrix} 200 & 68 \\ 68 & 26 \end{pmatrix}$$

Eigenvalues:

$$\mathbf{C} = \begin{pmatrix} 200 & 68 \\ 68 & 26 \end{pmatrix}$$

$$\det(\mathbf{C} - \lambda \mathbf{1}) = 0$$

$$\begin{pmatrix} 200 - \lambda & 68 \\ 68 & 26 - \lambda \end{pmatrix} = 0$$

$$(200 - \lambda)(26 - \lambda) - 68^2 = 0$$

$$\lambda_1 = 223.42, \quad \lambda_2 = 2.58$$

One eigenvalue much bigger than the other.

Eigenvectors:

$$\mathbf{C}\vec{v}_i = \lambda_i \vec{v}_i$$

$$\begin{aligned} 1. & \quad \mathbf{C}\vec{v}_1 = \begin{pmatrix} 200 & 68 \\ 68 & 26 \end{pmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = 223.42 \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} \\ 2. & \quad \mathbf{C}\vec{v}_2 = \begin{pmatrix} 200 & 68 \\ 68 & 26 \end{pmatrix} \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = 2.58 \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \\ 3. & \quad x_1^2 + y_1^2 = 1, \quad x_2^2 + y_2^2 = 1 \end{aligned}$$

From 1. we get: $68y_1 = 23.42x_1$; $197.42y_1 = 68x_1 \Rightarrow y_1 = 0.34x_1$

Since $x_1^2 + y_1^2 = 1$ we get: $x_1 = 0.95$; $y_1 = 0.33$

From 2. we get: $68y_2 = -197.42x_2$; $23.42y_2 = -68x_2 \Rightarrow y_2 = -2.90x_2$

Since $x_2^2 + y_2^2 = 1$ we get: $x_2 = 0.33$; $y_2 = -0.95$

$$\vec{v}_1 = \begin{pmatrix} 0.95 \\ 0.33 \end{pmatrix}$$

$$\vec{v}_2 = \begin{pmatrix} 0.33 \\ -0.95 \end{pmatrix}$$

Both vectors are perpendicular and normalized.

A 2D Numerical Example

PCA Example –STEP 1

- Subtract the mean

from each of the data dimensions. All the x values have \bar{x} subtracted and y values have \bar{y} subtracted from them. This produces a data set whose mean is zero.

Subtracting the mean makes variance and covariance calculation easier by simplifying their equations. The variance and co-variance values are not affected by the mean value.

PCA Example –STEP 1

<http://kybele.psych.cornell.edu/~edelman/Psych-465-Spring-2003/PCA-tutorial.pdf>

DATA:

<u>x</u>	<u>y</u>
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

ZERO MEAN DATA:

<u>x</u>	<u>y</u>
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

PCA Example –STEP 1

<http://kybele.psych.cornell.edu/~edelman/Psych-465-Spring-2003/PCA-tutorial.pdf>

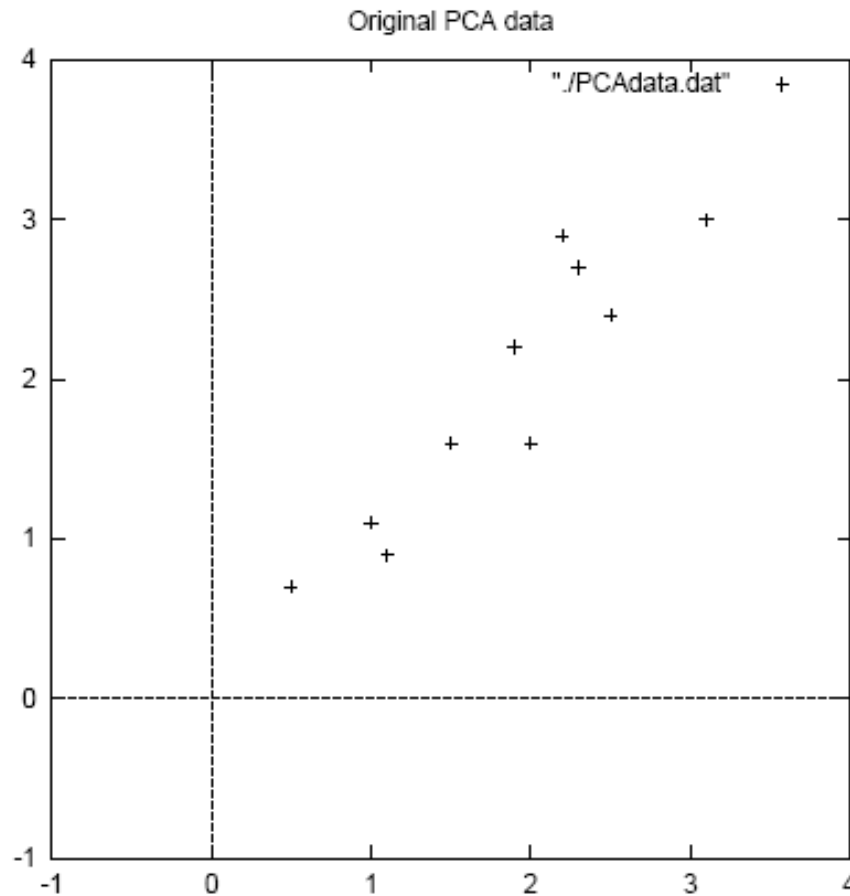


Figure 3.1: PCA example data, original data on the left, data with the means subtracted on the right, and a plot of the data

- Calculate the covariance matrix

$$\text{cov} = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

- Formula used: $\text{cov}(\mathbf{B}) = 1/(\mathbf{N}-1) \mathbf{B}^T \mathbf{B}$
Bessel's correction
- since the non-diagonal elements in this covariance matrix are positive, we should expect that both the x and y variable increase together.

PCA Example –STEP 3

- Calculate the eigenvectors and eigenvalues of the covariance matrix

eigenvalues = 0.0490833989, 1.28402771

$$\text{eigenvectors} = \begin{pmatrix} -0.735178656 \\ 0.677873399 \end{pmatrix} \begin{pmatrix} -0.677873399 \\ -0.735178656 \end{pmatrix}$$

PCA Example –STEP 3

<http://kybele.psych.cornell.edu/~edelman/Psych-465-Spring-2003/PCA-tutorial.pdf>

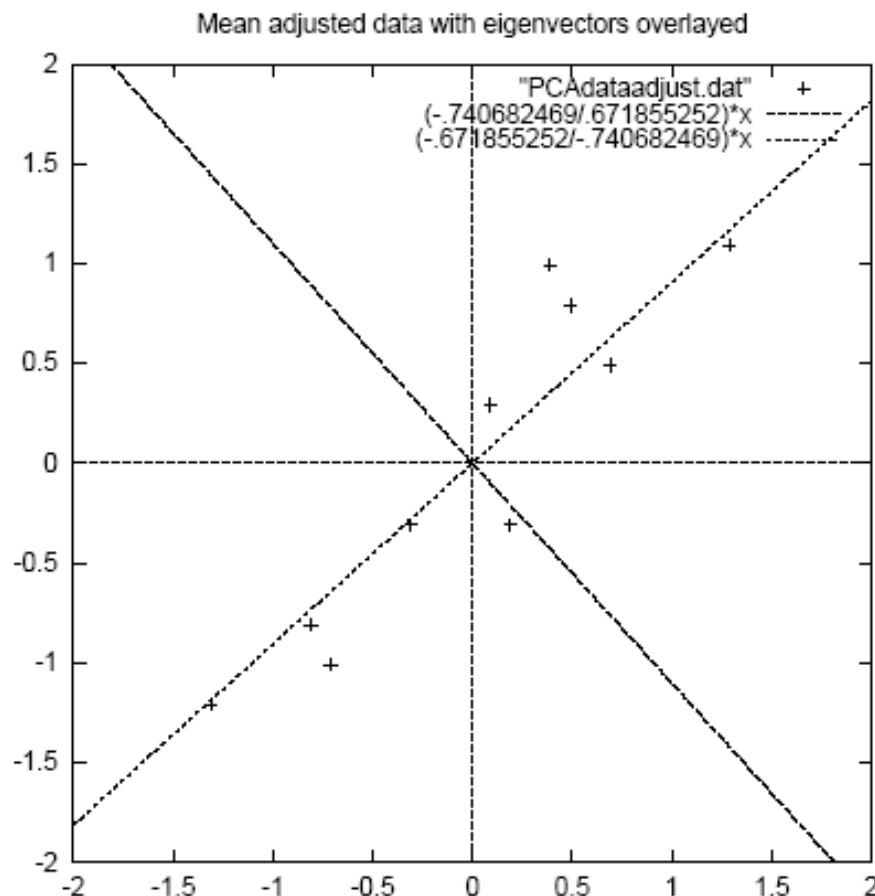


Figure 3.2: A plot of the normalised data (mean subtracted) with the eigenvectors of the covariance matrix overlayed on top.

- eigenvectors are plotted as diagonal dotted lines on the plot.

- Note they are perpendicular to each other.

- Note one of the eigenvectors goes through the middle of the points, like drawing a line of best fit.

- The second eigenvector gives us the other, less important, pattern in the data, that all the points follow the main line, but are off to the side of the main line by some amount.

PCA Example –STEP 4

- Reduce dimensionality and form *feature vector*
the eigenvector with the *highest* eigenvalue is the *principle component* of the data set.

In our example, the eigenvector with the largest eigenvalue was the one that pointed down the middle of the data.

Once eigenvectors are found from the covariance matrix, the next step is to **order them by eigenvalue**, highest to lowest. This gives you the components in order of significance.

PCA Example –STEP 4

Now, if you like, you can decide to *ignore* the components of lesser significance.

You do *lose some information*, but if the eigenvalues are small, you don't lose much

- n dimensions in your data
 - calculate n eigenvectors and eigenvalues
 - choose only the first p eigenvectors
 - final data set has only p dimensions.

- Feature Vector

$$\text{FeatureVector} = (\text{eig}_1 \text{ eig}_2 \text{ eig}_3 \dots \text{eig}_n)$$

We can either form a feature vector with both of the eigenvectors:

$$\begin{pmatrix} -.677873399 & -.735178656 \\ -.735178656 & .677873399 \end{pmatrix}$$

or, we can choose to leave out the smaller, less significant component and only have a single column:

$$\begin{pmatrix} -.677873399 \\ -.735178656 \end{pmatrix}$$

PCA Example –STEP 5

- Deriving the new data

$$\text{FinalData} = \text{RowFeatureVector} \times \text{RowZeroMeanData}$$

RowFeatureVector is the matrix with the eigenvectors in the columns *transposed* so that the eigenvectors are now in the rows, with the most significant eigenvector at the top

RowZeroMeanData is the mean-adjusted data *transposed*, ie. the data items are in each column, with each row holding a separate dimension.

PCA Example –STEP 5

FinalData transpose: dimensions
along columns

x	y
-.827970186	-.175115307
1.77758033	.142857227
-.992197494	.384374989
-.274210416	.130417207
-1.67580142	-.209498461
-.912949103	.175282444
.0991094375	-.349824698
1.14457216	.0464172582
.438046137	.0177646297
1.22382056	-.162675287

PCA Example –STEP 5

<http://kybele.psych.cornell.edu/~edelman/Psych-465-Spring-2003/PCA-tutorial.pdf>

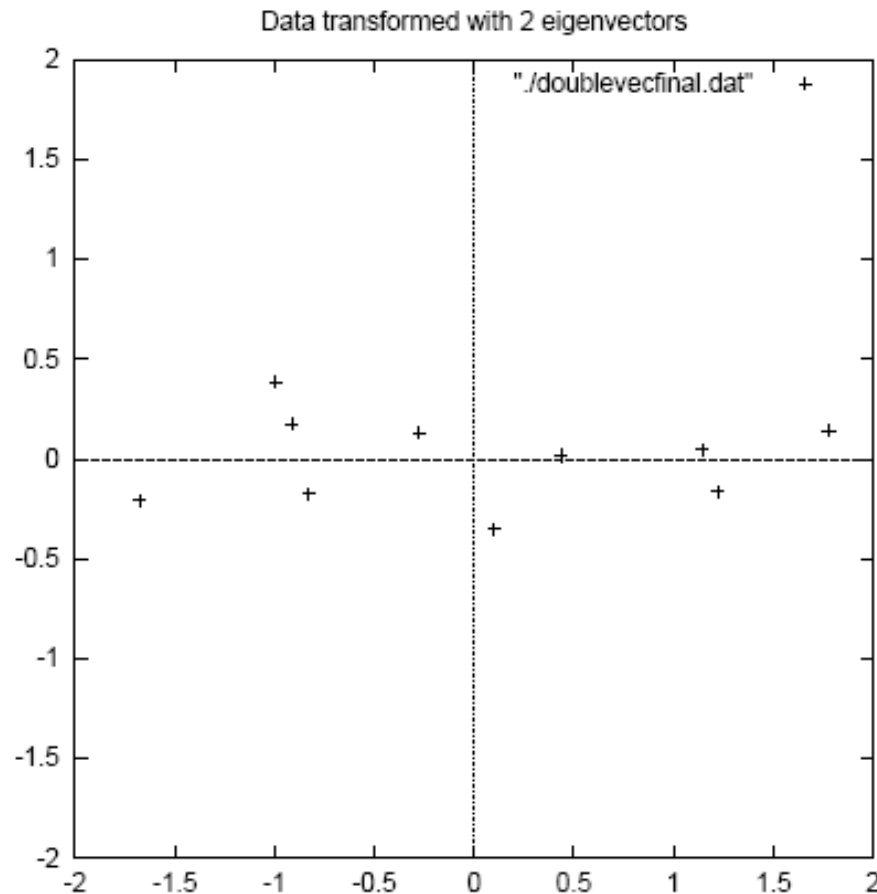


Figure 3.3: The table of data by applying the PCA analysis using both eigenvectors, and a plot of the new data points.

Reconstruction of original Data

- If we reduced the dimensionality, obviously, when reconstructing the data we would lose those dimensions we chose to discard. In our example let us assume that we considered only the x dimension...

Reconstruction of original Data

<http://kybele.psych.cornell.edu/~edelman/Psych-465-Spring-2003/PCA-tutorial.pdf>

X

-.827970186
 1.77758033
 -.992197494
 -.274210416
 -1.67580142
 -.912949103
 .0991094375
 1.14457216
 .438046137
 1.22382056

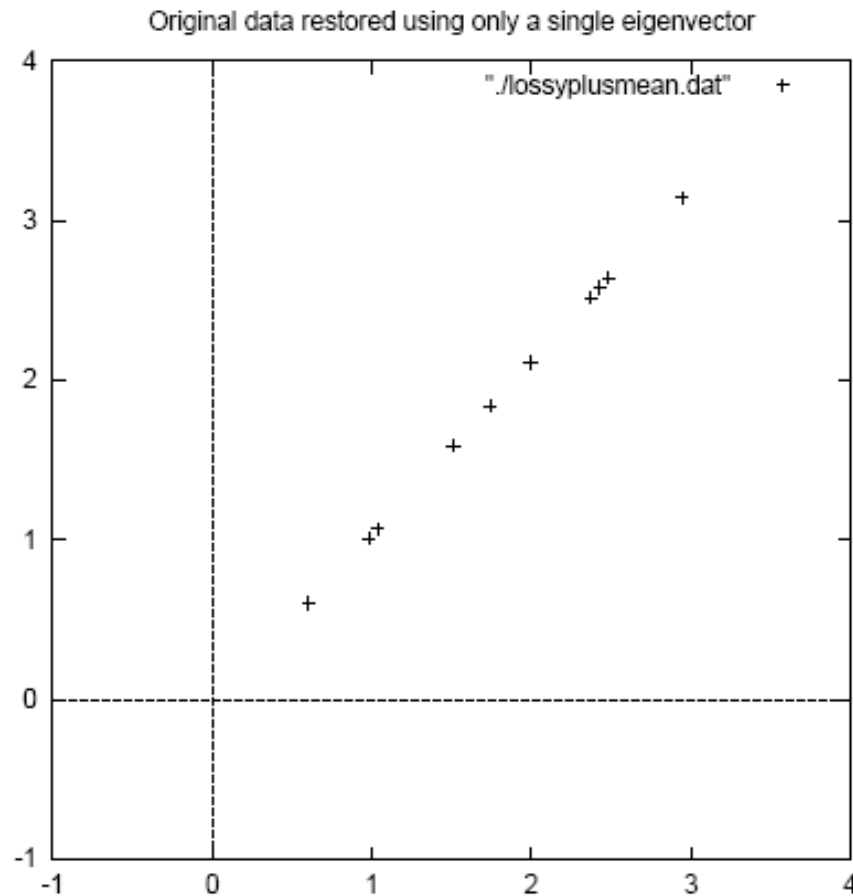


Figure 3.5: The reconstruction from the data that was derived using only a single eigenvector

Homework

1. Example of PCA and interpretation of principal components. Read with understanding.

<http://docshare04.docshare.tips/files/12598/125983744.pdf>

Table 8.4 Stock-Price Data (Weekly Rate Of Return)

Week	J P Morgan	Citibank	Wells Fargo	Royal Dutch Shell	Exxon Mobil
1	0.01303	-0.00784	-0.00319	-0.04477	0.00522
2	0.00849	0.01669	-0.00621	0.01196	0.01349
3	-0.01792	-0.00864	0.01004	0	-0.00614
4	0.02156	-0.00349	0.01744	-0.02859	-0.00695
5	0.01082	0.00372	-0.01013	0.02919	0.04098
6	0.01017	-0.01220	-0.00838	0.01371	0.00299
7	0.01113	0.02800	0.00807	0.03054	0.00323
8	0.04848	-0.00515	0.01825	0.00633	0.00768
9	-0.03449	-0.01380	-0.00805	-0.02990	-0.01081
10	-0.00466	0.02099	-0.00608	-0.02039	-0.01267
...
94	0.03732	0.03593	0.02528	0.05819	0.01697
95	0.02380	0.00311	-0.00688	0.01225	0.02817
96	0.02568	0.05253	0.04070	-0.03166	-0.01885
97	-0.00606	0.00863	0.00584	0.04456	0.03059
98	0.02174	0.02296	0.02920	0.00844	0.03193
99	0.00337	-0.01531	-0.02382	-0.00167	-0.01723
100	0.00336	0.00290	-0.00305	-0.00122	-0.00970
101	0.01701	0.00951	0.01820	-0.01618	-0.00756
102	0.01039	-0.00266	0.00443	-0.00248	-0.01645
103	-0.01279	-0.01437	-0.01874	-0.00498	-0.01637

The eigenvalues and corresponding normalized eigenvectors of \mathbf{R} , determined by a computer, are

$$\hat{\lambda}_1 = 2.437, \quad \hat{\mathbf{e}}_1' = [.469, .532, .465, .387, .361]$$

$$\hat{\lambda}_2 = 1.407, \quad \hat{\mathbf{e}}_2' = [-.368, -.236, -.315, .585, .606]$$

$$\hat{\lambda}_3 = .501, \quad \hat{\mathbf{e}}_3' = [-.604, -.136, .772, .093, -.109]$$

$$\hat{\lambda}_4 = .400, \quad \hat{\mathbf{e}}_4' = [.363, -.629, .289, -.381, .493]$$

$$\hat{\lambda}_5 = .255, \quad \hat{\mathbf{e}}_5' = [.384, -.496, .071, .595, -.498]$$

Using the standardized variables, we obtain the first two sample principal components:

$$\hat{y}_1 = \hat{\mathbf{e}}_1' \mathbf{z} = .469z_1 + .532z_2 + .465z_3 + .387z_4 + .361z_5$$

$$\hat{y}_2 = \hat{\mathbf{e}}_2' \mathbf{z} = -.368z_1 - .236z_2 - .315z_3 + .585z_4 + .606z_5$$

These components, which account for

$$\left(\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{p} \right) 100\% = \left(\frac{2.437 + 1.407}{5} \right) 100\% = 77\%$$

of the total (standardized) sample variance, have interesting interpretations. The first component is a roughly equally weighted sum, or "index," of the five stocks. This component might be called a *general stock-market component*, or, simply, a *market component*.

The second component represents a contrast between the banking stocks (JP Morgan, Citibank, Wells Fargo) and the oil stocks (Royal Dutch Shell, Exxon-Mobil). It might be called an *industry component*. Thus, we see that most of the variation in these stock returns is due to market activity and uncorrelated industry activity. This interpretation of stock price behavior also has been suggested by King [12].

The remaining components are not easy to interpret and, collectively, represent variation that is probably specific to each stock. In any event, they do not explain much of the total sample variance. ■

Homework

2. Solve this:

- 8.6.** Data on x_1 = sales and x_2 = profits for the 10 largest companies in the world were listed in Exercise 1.4 of Chapter 1.
From Example 4.12

$$\bar{\mathbf{x}} = \begin{bmatrix} 155.60 \\ 14.70 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 7476.45 & 303.62 \\ 303.62 & 26.19 \end{bmatrix}$$

- (a) Determine the sample principal components and their variances for these data. (You may need the quadratic formula to solve for the eigenvalues of \mathbf{S} .)
 - (b) Find the proportion of the total sample variance explained by \hat{y}_1 .
- 8.7.** Convert the covariance matrix \mathbf{S} in Exercise 8.6 to a sample correlation matrix \mathbf{R} .
- (a) Find the sample principal components \hat{y}_1, \hat{y}_2 and their variances.
 - (b) Compute the proportion of the total sample variance explained by \hat{y}_1 .
 - (c) Compute the correlation coefficients $r_{\hat{y}_1, x_k}, k = 1, 2$. Interpret \hat{y}_1 .
 - (d) Compare the components obtained in Part a with those obtained in Exercise 8.6(a). Given the original data displayed in Exercise 1.4, do you feel that it is better to determine principal components from the sample covariance matrix or sample correlation matrix? Explain.

Homework

Some hints:

If X and Y are two random variables, with **means** (expected values) μ_X and μ_Y and standard deviations σ_X and σ_Y , respectively, then their covariance and correlation are as follows:

covariance $\text{cov}_{XY} = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$

correlation $\text{corr}_{XY} = \rho_{XY} = E[(X - \mu_X)(Y - \mu_Y)] / (\sigma_X \sigma_Y),$

so that

$$\rho_{XY} = \sigma_{XY} / (\sigma_X \sigma_Y)$$

Page 449 of the book: Standardizing the Sample Principal Components

Example of conversion of covariance matrix to correlation matrix:

$$\text{Cov.matrix} = \begin{pmatrix} 1.0 & 1.0 & 8.1 \\ 1.0 & 16.0 & 18.0 \\ 8.1 & 18.0 & 81.0 \end{pmatrix}$$

Corr. Matrix =

R		
1	0.25	0.9
0.25	1	0.5
0.9	0.5	1

Homework

- **Extra problem:**

In the python script `plot_digits_simple_classif.ipynb` try to reconstruct the digits using limited number of principal components.