university of groningen

Principal Component Analysis

Seminar in Methodology and Statistics

Nafid Haque

EM-LCT

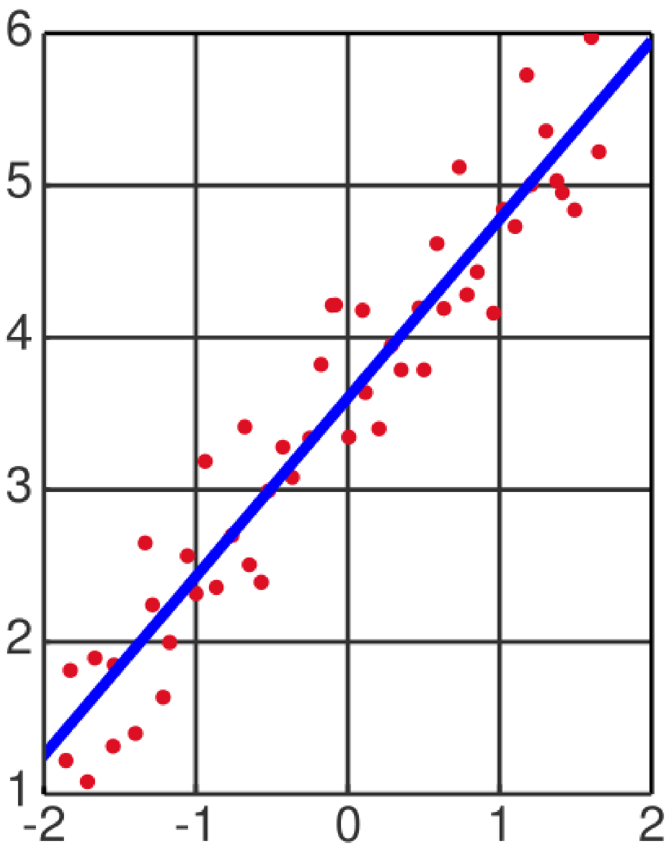
N.Haque@student.rug.nl

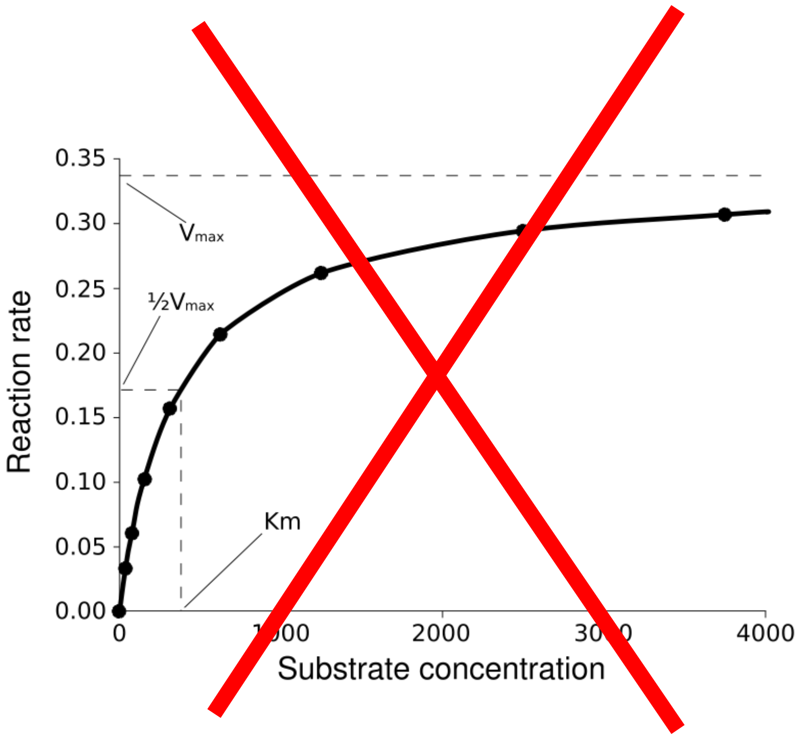
Outline

- What is PCA?
- Steps for performing PCA
- Conclusion
- Discussion

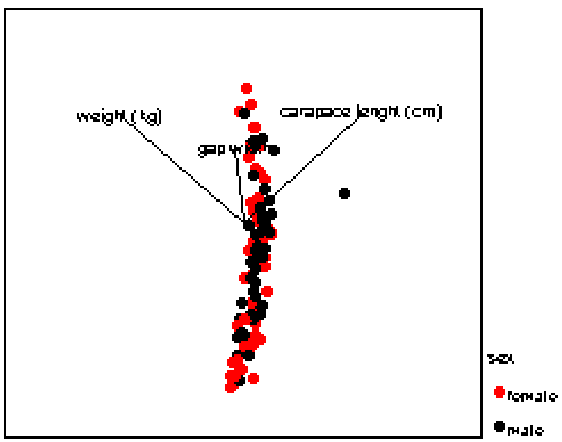
What is PCA?

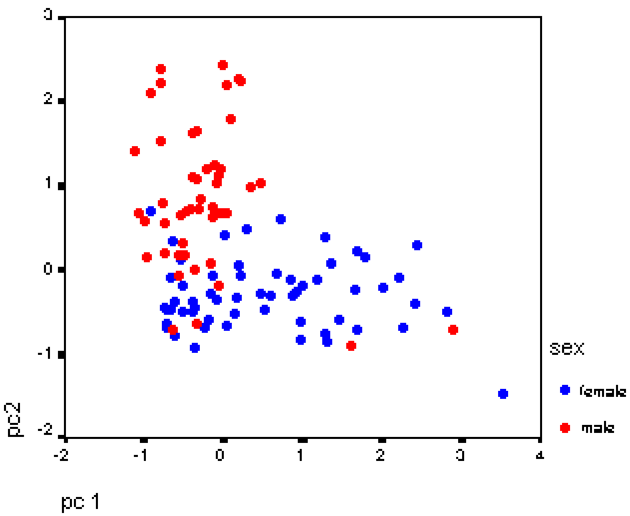
- A statistical method for exploring and making sense of datasets
- It is used to ‘summarize’ the data (not to ‘cluster’ data)
- Only used for linear data
- Its goal is to reduce the dimensionality of the original data set

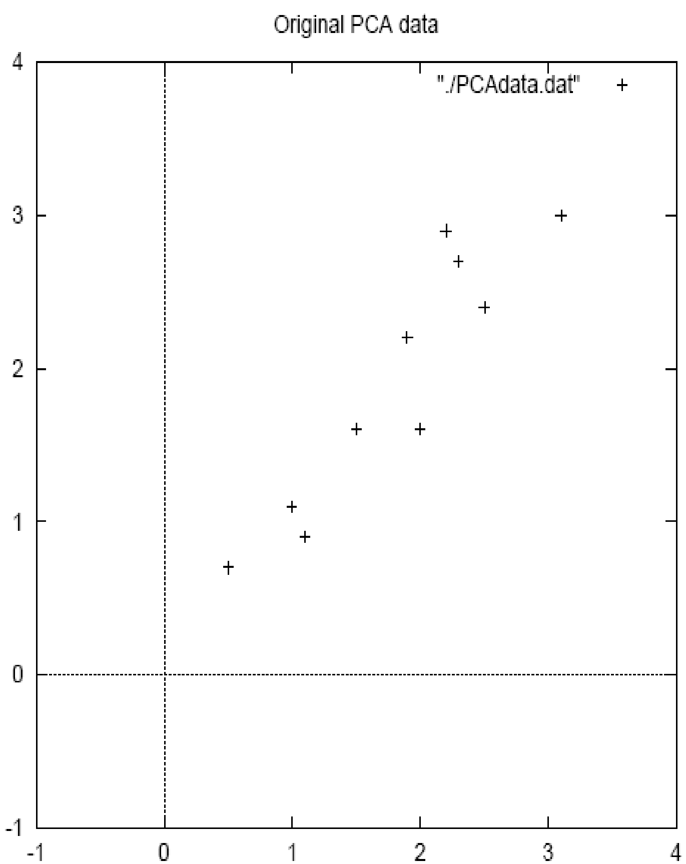

6
5
4
3
2
1
-2
-1
0
1
2

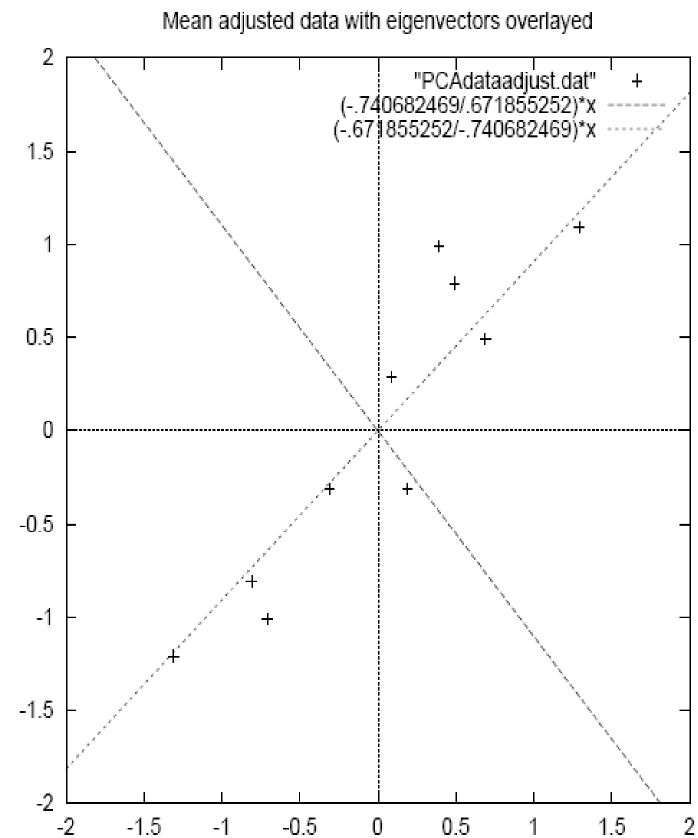

0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Vmax
½Vmax
Km
Reaction rate
0
2000
4000
Substrate concentration

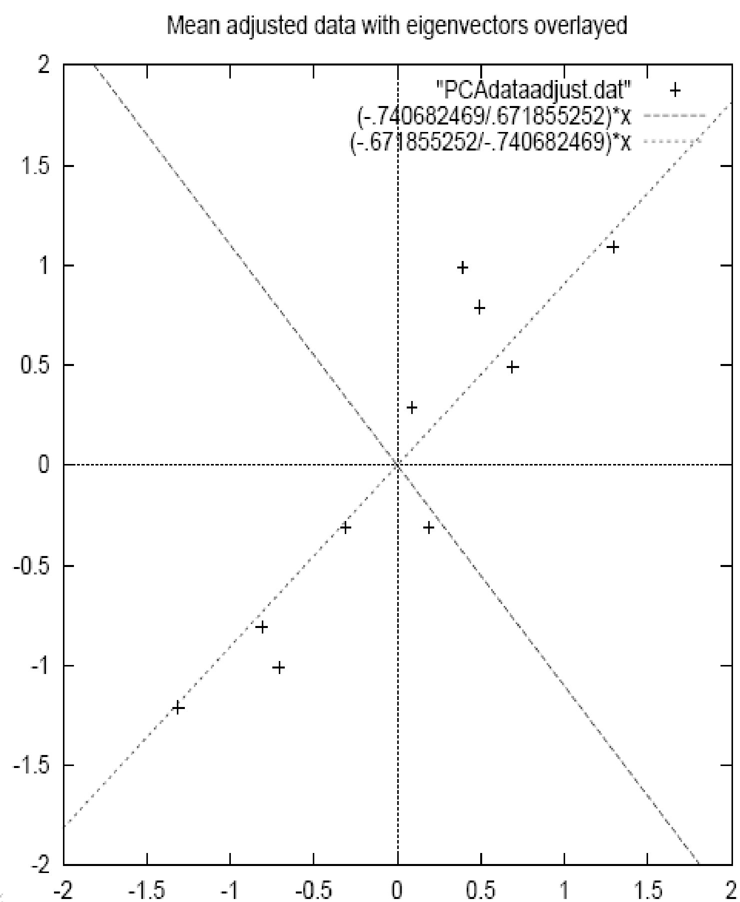
case	ht (x_1)	wt(x_2)	age(x_3)	sbp(x_4)	heart rate (x_5)
1	175	1225	25	117	56
2	156	1050	31	122	63
n	202	1350	58	154	67

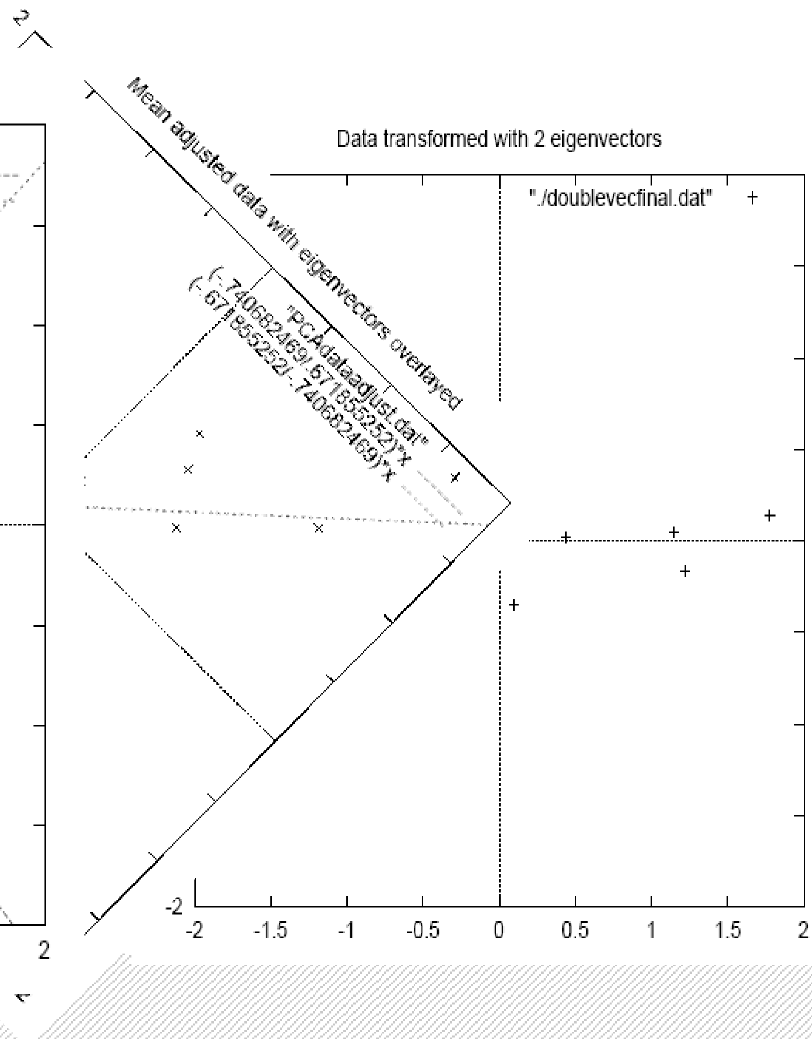





Original PCA data
"./PCAdata.dat"
-1
0
1
2
3
4


Mean adjusted data with eigenvectors overlayed
"PCAdataadjust.dat"
(-.740682469/.671855252)*x
(-.671855252/-.740682469)*x
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2


Mean adjusted data with eigenvectors overlayed
"PCAdataadjust.dat"
(-.740682469/.671855252)*x
(-.671855252/-.740682469)*x


Data transformed with 2 eigenvectors
"./doublevecfinal.dat"
Mean adjusted data with eigenvectors overlayed
"PCAdataadjust.dat"
(-.740682469/.671855252)*x
(-.671855252/-.740682469)*x

The steps to carry out PCA on a dataset

- Step 1: Get some data
- Step 2: Normalize/Adjust the data (Subtract the mean)
- Step 3: Calculate the covariance matrix
- Step 4: Calculate the eigenvectors and eigenvalues of the covariance matrix
- Step 5: Choosing components and forming a feature vector
- Step 6: Deriving the new dataset

Step 1: Get some data

X	Y
2.50	2.40
0.50	0.70
2.20	2.90
1.90	2.20
3.10	3.00
2.30	2.70
2.00	1.60
1.00	1.10
1.50	1.60
1.10	0.90

Step 2: Normalize/Adjust the data (Subtract the mean)

X	Y	X_i-X_m	Y_i-Y_m
2.50	2.40	0.69	0.49
0.50	0.70	-1.31	-1.21
2.20	2.90	0.39	0.99
1.90	2.20	0.09	0.29
3.10	3.00	1.29	1.09
2.30	2.70	0.49	0.79
2.00	1.60	0.19	-0.31
1.00	1.10	-0.81	-0.81
1.50	1.60	-0.31	-0.31
1.10	0.90	-0.71	-1.01
18.10	19.10	0.00	0.00
1.81	1.91	0.00	0.00

Step 3: Calculate the covariance matrix

- Covariance is
  - How two variables change with respect to each other (so 2 dimensions)
  - (Variance operate only on 1 dimension)
  - We have 2 dimensional data so we need to calculate *cov* (X,Y)

Step 3.1 (a): How to calculate *cov* (X,Y)

- Variance

$$var(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{(n-1)}$$

- Covariance

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

Step 3.1 (b): How to calculate *cov* (X,Y)

X	Y	X_i-X_m	(Xi-Xm)(Xi-Xm)	Y_i-Y_m	(Yi-Ym)(Yi-Ym)	(Xi-Xm)(Yi-Ym)
2.50	2.40	0.69	0.48	0.49	0.24	0.34
0.50	0.70	-1.31	1.72	-1.21	1.46	1.59
2.20	2.90	0.39	0.15	0.99	0.98	0.39
1.90	2.20	0.09	0.01	0.29	0.08	0.03
3.10	3.00	1.29	1.66	1.09	1.19	1.41
2.30	2.70	0.49	0.24	0.79	0.62	0.39
2.00	1.60	0.19	0.04	-0.31	0.10	-0.06
1.00	1.10	-0.81	0.66	-0.81	0.66	0.66
1.50	1.60	-0.31	0.10	-0.31	0.10	0.10
1.10	0.90	-0.71	0.50	-1.01	1.02	0.72
18.10	19.10	0.00	5.55	0.00	6.45	5.54
1.81	1.91	0.00	0.62	0.00	0.72	0.62

Step 3.2 (a): How to find the covariance matrix

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$

Step 3.2 (b): How to find the covariance matrix

X	Y	X_i-X_m	(Xi-Xm)(Xi-Xm)	Y_i-Y_m	(Yi-Ym)(Yi-Ym)	(Xi-Xm)(Yi-Ym)
2.50	2.40	0.69	0.48	0.49	0.24	0.34
0.50	0.70	-1.31	1.72	-1.21	1.46	1.59
2.20	2.90	0.39	0.15	0.99	0.98	0.39
1.90	2.20	0.09	0.01	0.29	0.08	0.03
3.10	3.00	1.29	1.66	1.09	1.19	1.41
2.30	2.70	0.49	0.24	0.79	0.62	0.39
2.00	1.60	0.19	0.04	-0.31	0.10	-0.06
1.00	1.10	-0.81	0.66	-0.81	0.66	0.66
1.50	1.60	-0.31	0.10	-0.31	0.10	0.10
1.10	0.90	-0.71	0.50	-1.01	1.02	0.72
18.10	19.10	0.00	5.55	0.00	6.45	5.54
1.81	1.91	0.00	0.62	0.00	0.72	0.62

Step 3.2 (c): How to find the covariance matrix

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

Step 4: Calculate the eigenvectors and eigenvalues of the covariance matrix

Let A be an $n \times n$ matrix. The number λ is an **eigenvalue** of A if there exists a non-zero vector $\mathbf{v}$ such that

$$A\mathbf{v} = \lambda\mathbf{v}.$$

In this case, vector $\mathbf{v}$ is called an **eigenvector** of A corresponding to λ.

Step 4.1 (a): Examples of eigenvectors and eigenvalues

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 11 \\ 5 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4 \times \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 6 \\ 4 \end{pmatrix} = \begin{pmatrix} 24 \\ 16 \end{pmatrix} = 4 \times \begin{pmatrix} 6 \\ 4 \end{pmatrix}$$

Example of one non-eigenvector and one eigenvector

Step 4.1 (b): How to compute the eigenvectors and eigenvalues

We can rewrite the condition $A\mathbf{v} = \lambda\mathbf{v}$ as

$$(A - \lambda I)\mathbf{v} = \mathbf{0}.$$

where I is the $n \times n$ identity matrix. Now, in order for a *non-zero* vector $\mathbf{v}$ to satisfy this equation, $A - \lambda I$ must *not* be invertible.

That is, the determinant of $A - \lambda I$ must equal 0.

We call $p(\lambda) = \det(A - \lambda I)$ the **characteristic polynomial** of A. The eigenvalues of A are simply the roots of the characteristic polynomial of A.

Step 4.1.1 : What is a determinant of a matrix?

- For 2 by 2,

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad \det(A) = ad - bc.$$

- For 3 by 3,

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}.$$

$$\begin{aligned} \det(A) &= a\begin{vmatrix} e & f \\ h & i \end{vmatrix} - b\begin{vmatrix} d & f \\ g & i \end{vmatrix} + c\begin{vmatrix} d & e \\ g & h \end{vmatrix} \\ &= aei - afh - bdi + bfg + cdh - ceg \\ &= (aei + bfg + cdh) - (gec + hfa + idb), \end{aligned}$$

Step 4.2 : Finally the eigenvectors and the eigenvalues for our example

$$eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

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

Step 5: Choosing components and forming a feature vector

$$FeatureVector = (eig_1\ eig_2\ eig_3\ ....\ eig_n)$$

Given our example set of data, and the fact that we have 2 eigenvectors, we have two choices. 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:

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

Step 6: Deriving the new dataset

$$FinalData = RowFeatureVector \times RowDataAdjust,$$

where $RowFeatureVector$ 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, and $RowDataAdjust$ is the mean-adjusted data *transposed*, ie. the data items are in each column, with each row holding a separate dimension.

Step 6.1 (a): Deriving the new dataset

Transformed Data=

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

Step 6.1 (b): Deriving the new dataset

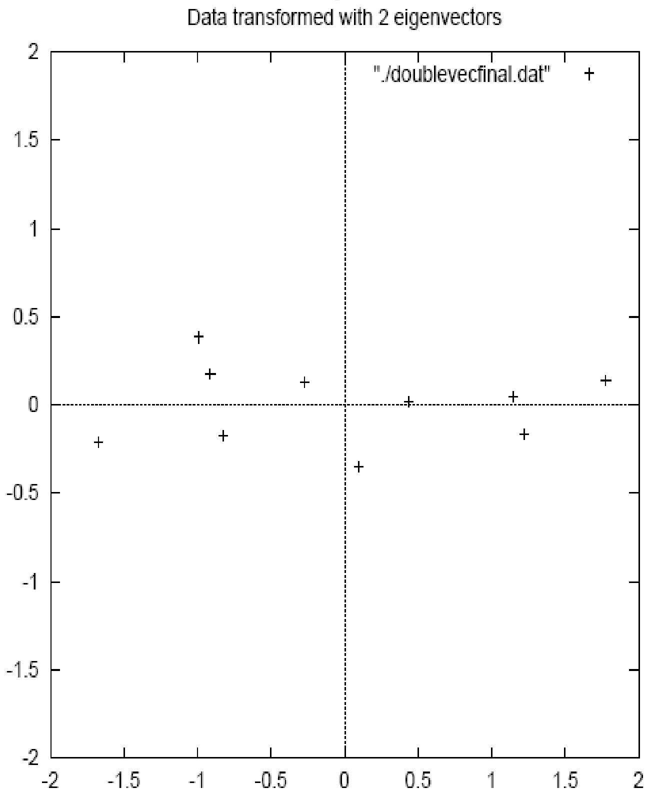


Figure 3.3: The table of data by applying the PCA analysis using both eigenvectors, and a plot of the new data points.

Conclusion (1)

› So PCA gives new variables (dimensions) that are linear combination of the original ones

› The new variables are derived in decreasing order of importance

› How many PCs to keep?

- Enough to keep a cumulative variance explained by the PCs
- (Kaiser Criterion- keep PCs>1)
- (Scree plot)

Conclusion (2)

- PCA is basically useful for finding new, more informative, uncorrelated features
- PCA reduces dimensionality by rejecting low variance features

university of groningen

References:

- Ahmed Rebai, Presentation of PCA-ICA
- Harvey Mudd College Math Tutorial: Eigenvalues and Eigenvectors
- Lindsay I Smith, A tutorial on Principal Components Analysis
- Giorgos Korfiatis, Presentation of PCA