Principal Component Analysis

(PCA) is a technique that can be used to simplify a dataset. It is a linear transformation that chooses a new coordinate system for the data set such that greatest variance by any projection of the data set comes to lie on the first axis (then called the first principal component), the second greatest variance on the second axis, and so on.

By finding the eigenvalues and eigenvectors of the covariance matrix, we find that the eigenvectors with the largest eigenvalues correspond to the dimensions that have the strongest correlation in the dataset. This is the principal component.

Why PCA

- To interpret data
- Standardizing the different range of features
- Dimension reduction
  - Instead of more features use Less significant features – express more variations of Classes
  - Reduction of irrelevant features
  - Transforming a large set of variables into a smaller one that still contains most of the information in the large set.
  - Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process.
- Find patterns in high-dimensional data
- Visualize data of high dimensionality

Applications

- – Face recognition
- – Image compression
- – Gene expression analysis

PCA –Dimension reduction

The data with D features, may be the data points mainly lie in a linear subspace of dimension lower than D. But not exactly in some lower-dimensional subspace, we might be able to find such reduced subspace of dimension d<D which maintains most of the variability in data. There will be d new variables defining the subspace. The new variables, which form a new coordinate system are called principal components (PCs) and denoted as PC_1, PC_2…PC_D. First component

have the maximum variance, ie capturing as much of the variability in $x_1,x_2,..x_D$ as possible. Subsequent PC will take up successively smaller parts the total variability.

For example, in figure 1, suppose that the triangles represent a two variable data set which we have measured in the X-Y coordinate system. The principal direction in which the data varies is shown by the U axis and the second most important direction is the V axis orthogonal to it. If we place the U − V axis system at the mean of the data it gives us a compact representation. If we transform each (X, Y) coordinate into its corresponding (U, V) value, the data is de-correlated, meaning that the co-variance between the U and V variables is zero. For a given set of data, principal component analysis finds the axis system defined by the principal directions of variance (ie the U − V axis system in figure 1). The directions U and V are called the principal components.

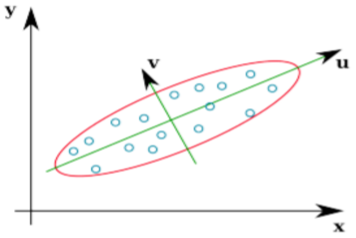


Figure 1: PCA for Data Representation

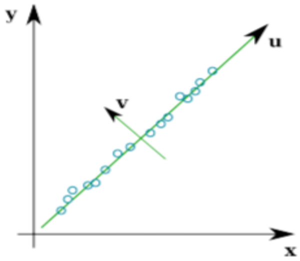


Figure 2: PCA for Dimension Reduction

If the variation in a data set is caused by some natural property, or is caused by random experimental error, then we may expect it to be normally distributed. In this case we show the nominal extent of the normal distribution by a hyper-ellipse (the two dimensional ellipse in the example). The hyper ellipse encloses data points that are thought of as belonging to a class. It is drawn at a distance beyond which the probability of a point belonging to the class is low, and can be thought of as a class boundary. If the variation in the data is caused by some other relationship then PCA gives us a way of reducing the dimensionality of a data set. Consider two variables that are nearly related linearly as shown in figure 2. As in figure 1 the principal direction in which the data varies is shown by the U axis, and the secondary direction by the V axis. However in this case all the V coordinates are all very close to zero. We may assume, for example, that they are only non zero because of experimental noise. Thus in the U − V axis system we can represent the data set by one variable U and discard V . Thus we have reduced the dimensionality of the problem by 1.

PCA Toy Example

Consider following 3D points

$$\begin{matrix} 1 & 2 & 4 & 3 & 5 & 6 \\ 2 & 4 & 8 & 6 & 10 & 12 \\ 3 & 6 & 12 & 9 & 15 & 18 \end{matrix}$$

If each component is stored in a byte then we need (3×6)=18 bytes

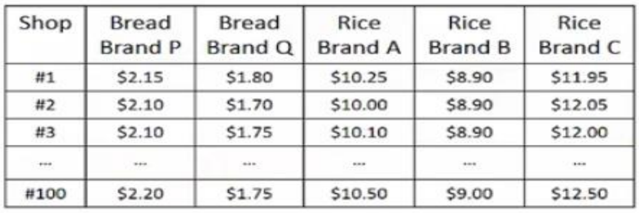
1	2	4	3	5	6
2	4	8	6	10	12
3	6	12	9	15	18
I	$I \times 2$	$I \times 4$	$I \times 3$	$I \times 5$	$I \times 6$

3 bytes for I and 6 bytes for multiplying constants =>9 bytes

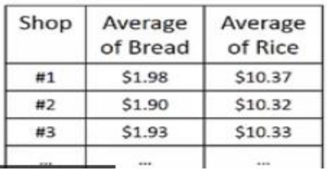
50% reduction.

Example 2:

Although n components are required to reproduce the total system variability, often much of this variability can be accounted for by a small number *p* of the principal components. If so, there is as much as information in the *p* components as there is in the original n variables. The *p* principal components can then replace the initial n variables and the original data set, consisting of N measurements on n variables is reduced to a data set consisting of N measurements on *p* principal components.



Shop	Bread Brand P	Bread Brand Q	Rice Brand A	Rice Brand B	Rice Brand C
#1	$2.15	$1.80	$10.25	$8.90	$11.95
#2	$2.10	$1.70	$10.00	$8.90	$12.05
#3	$2.10	$1.75	$10.10	$8.90	$12.00
...	...	...	...	...	...
#100	$2.20	$1.75	$10.50	$9.00	$12.50



Shop	Average of Bread	Average of Rice
#1	$1.98	$10.37
#2	$1.90	$10.32
#3	$1.93	$10.33
...	...	...

PCA-Algorithm

1. Let $\bar{X}$ be the **mean** vector (taking the mean of all rows)

   X=$a_{1,}a_{2,}\ldots a_d$ set of N×d vectors and let $\bar{X}$ be their average of d attributes.

   $$X = \begin{bmatrix} a_{11} & \ldots & a_{1d} \\ \cdot & \ldots & \cdot \\ a_{N1} & \ldots & a_{Nd} \end{bmatrix} \qquad \bar{X} = \frac{1}{N}\sum_{i=1}^{i=N} a_{ij,}\, j = 1\ to\ d$$

2. Adjust the original data by the mean X' = X – $\bar{X}$

   Let X be the N×d matrix with columns $X = [a_1 - \bar{X} \quad a_2 - \bar{X} \quad \ldots\ a_d - \bar{X}]$

   Subtracting the mean is equivalent to translating the coordinate system to the location of the mean.

3. Compute the **covariance** matrix C of adjusted X

$$C = \frac{1}{N-1}XX^T = \frac{1}{N-1}\begin{bmatrix}(a_1-\bar{X})^T \\ (a_2-\bar{X})^T \\ : \\ . \\ (a_d-\bar{X})^T\end{bmatrix}[a_1-\bar{X} \quad a_2-\bar{X} \quad \dots\ a_d-\bar{X}]$$

And C= $\begin{bmatrix} x_{11} & \dots & x_{1d} \\ \cdot & \dots & \cdot \\ x_{d1} & \dots & x_{dd}\end{bmatrix}$

C is square, symmetric, Covariance matrix - Normalized centered data matrix

4. Find the **eigenvectors and eigenvalues** of C.

In computational terms the principal components are found by calculating the eigenvectors and eigenvalues of the data covariance matrix. This process is equivalent to finding the axis system in which the co-variance matrix is diagonal. The eigenvector with the largest eigenvalue is the direction of greatest variation, the one with the second largest eigenvalue is the (orthogonal) direction with the next highest variation and so on. To see how the computation is done we will give a brief review on eigenvectors/eigenvalues. The eigen values of C are defined as the roots of :

determinant (C-λI)=|(C- λI)|=0

$$\left|\begin{bmatrix} x_{11} & \dots & x_{1d} \\ \cdot & \dots & \cdot \\ x_{d1} & \dots & x_{dd}\end{bmatrix} - \lambda\begin{bmatrix} 1 & \dots & 0 \\ . & \dots & . \\ 0 & \dots & 1\end{bmatrix}\right| = 0$$

$$\left|\begin{bmatrix} x_{11} & \dots & x_{1d} \\ \cdot & \dots & \cdot \\ x_{d1} & \dots & x_{dd}\end{bmatrix} - \begin{bmatrix} \lambda_{11} & \dots & 0 \\ . & \dots & . \\ 0 & \dots & \lambda_{dd}\end{bmatrix}\right| = 0$$

– I is a d×d identity matrix, yields a polynomial (characteristic polynomial) (degree d and has d roots)

– λ be the eigenvalue (roots) of C. Then here exists a vector $\vec{e}$ such that

$$C\,\vec{e} = \lambda\,\vec{e}$$

$$(C-\lambda I)\,\vec{e} = 0$$

$$\left[\begin{bmatrix} x_{11} & \dots & x_{1d} \\ \cdot & \dots & \cdot \\ x_{d1} & \dots & x_{dd}\end{bmatrix} - \begin{bmatrix} \lambda_{11} & \dots & 0 \\ . & \dots & . \\ 0 & \dots & \lambda_{dd}\end{bmatrix}\right]\begin{bmatrix} e_1 \\ .. \\ e_d\end{bmatrix} = 0$$

The vector $\vec{e}$ be an eigenvector of C associated with eigenvalue λ. Notice that there is no unique solution for $\vec{e}$ in the above equation. It is a direction vector only and can be scaled to any magnitude. To find the numerical solution of $\vec{e}$ we need to set one of its elements to an arbitrary value, say 1, which gives us a set of simultaneous equations to solve for the other elements.

- Solve *(C-λI)* $\vec{e}$=0 for each λ (λ_1, λ_1 .. λ_d, -d roots) to obtain eigenvectors $\vec{e}_i$
- $\vec{e}_i = \vec{e}_{i1}, \vec{e}_{i2} .. \vec{e}_{id}$ are d ×d orthonormal vectors
- *eigenvectors* of *C* is $\vec{e}$ such that *C*$\vec{e}$=λ$\vec{e}$,
- *C*$\vec{v}$=λ$\vec{e}$ ⇔ (*C*-λI) $\vec{e}$=0

Expressing C interms of $\vec{e}_{i1}, \vec{e}_{i2} .. \vec{e}_{id}$ has not changed the size of the data. Eigen values λ_i corresponds to variance on each component i. Sort the eigenvectors $\vec{e}_i$ according to their eigenvalue:

$$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_d$$

The new data set can be calculation:

$$\begin{bmatrix} y_{11} & \cdots & y_{1d} \\ . & \cdots & . \\ y_{N1} & \cdots & y_{Nd} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1d} \\ . & \cdots & . \\ a_{N1} & \cdots & a_{Nd} \end{bmatrix} \begin{bmatrix} \vec{e}_{i1} \\ \vec{e}_{i2} \\ \cdots \\ \vec{e}_{id} \end{bmatrix}, \qquad i = 1 \; to \; d$$

Consider a linear combinations in which the new data is calculated. Take the first *p* components based on top *p* eigenvectors $\vec{e}_i$. These are the directions with the largest variances.

$$Y_1 = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1d}X_d$$

$$Y_2 = e_{21}X_1 + e_{22}X_2 + \cdots + e_{2d}X_d$$

.

.

$$Y_p = e_{d1}X_1 + e_{d2}X_2 + \cdots + e_{dd}X_d$$

Each of these can be thought of as a linear regression, predicting Y_i from X_1, X_2, ... , X_d. There is no intercept, but e_{i1}, e_{i2}, ..., e_{id} can be viewed as regression coefficients.

The Principal Components selection:

$$PVE\_PC_i = (\lambda_i / \sum_{i=1}^{d} \lambda_i) \times 100$$

Where PVE-Percentage of Variance Explained

$$PVE\_PC_1 \geq PVE\_PC_2 \geq \cdots PVE\_PC_p \geq \cdots \geq PVE\_PC_d$$

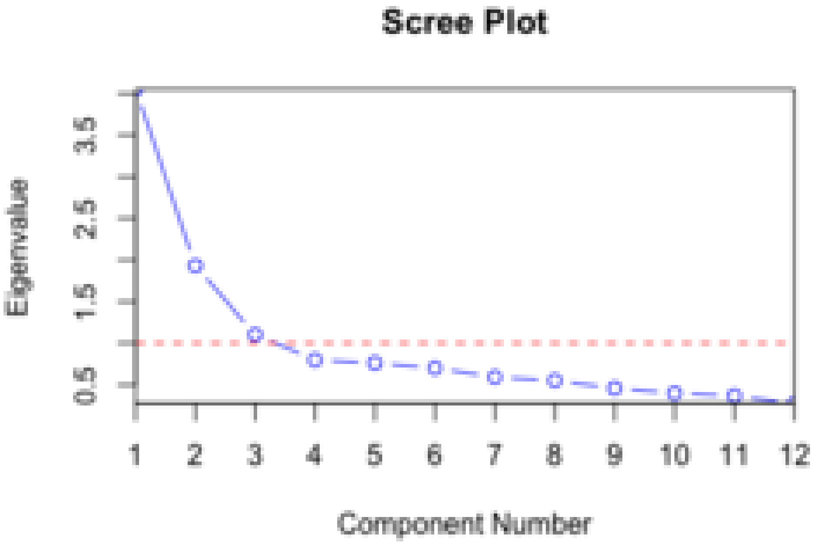
principal component (PC1)

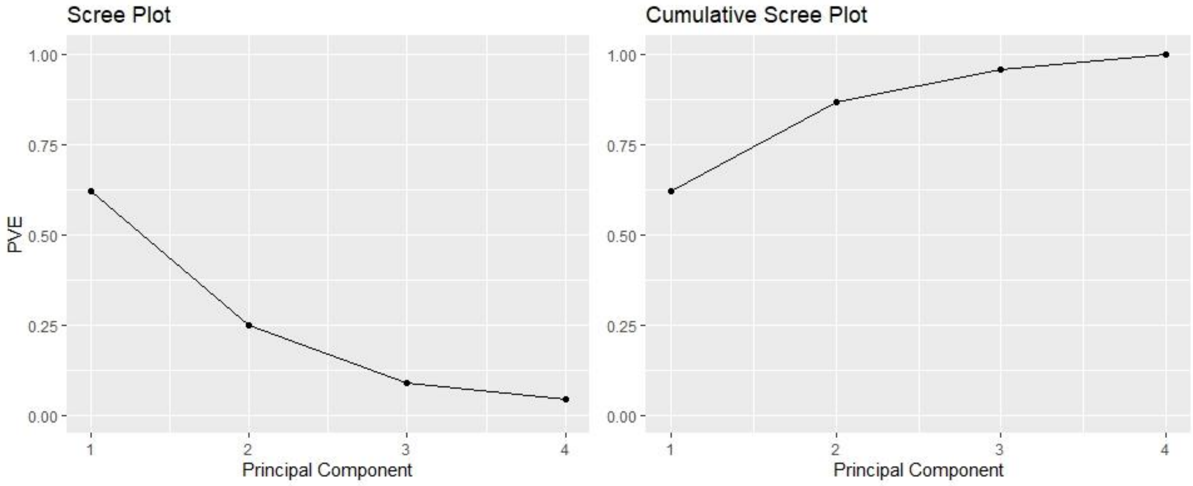
– The eigenvalue with the largest absolute value will indicate that the data have the largest variance along its eigenvector, the direction along which there is greatest variation

principal component (PC2)

- the direction with maximum variation left in data, orthogonal to the PC1. In general, only few directions manage to capture most of the variability in the data.

In the same way the d components are speaking about the variance and the percentage of variance are in descending order. From this d components select set of components which together provide the variance such 80% or 90 %.

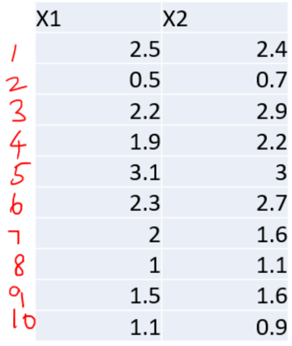




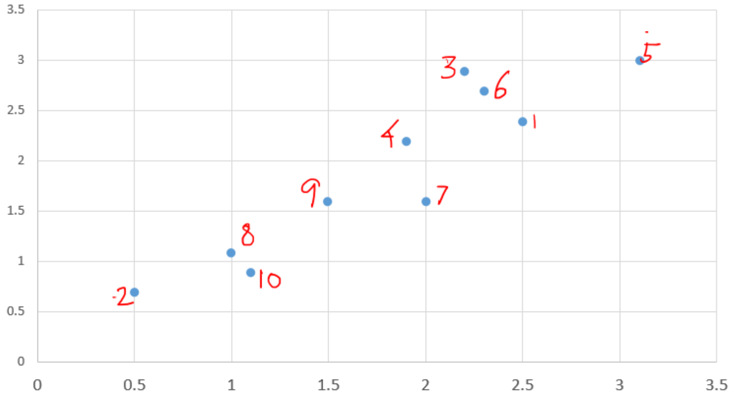
Scree plot

The scree test is derived by plotting the eigenvalues (on the Y axis) against the number of components in their order of extraction (on the X axis). The initial components extracted are larger (with high eigenvalues), followed by smaller components. Graphically, the plot will show a steep slope between the large components and the gradual trailing off of the rest of the components. The point at which the curve first begins to straighten out is considered to indicate the maximum number of components to extract. That is, those components above this point of inflection are deemed meaningful, and those below are not. As a general rule, the scree test results in at least one and sometimes two or three more components being considered significant than does the eigenvalue criterion.

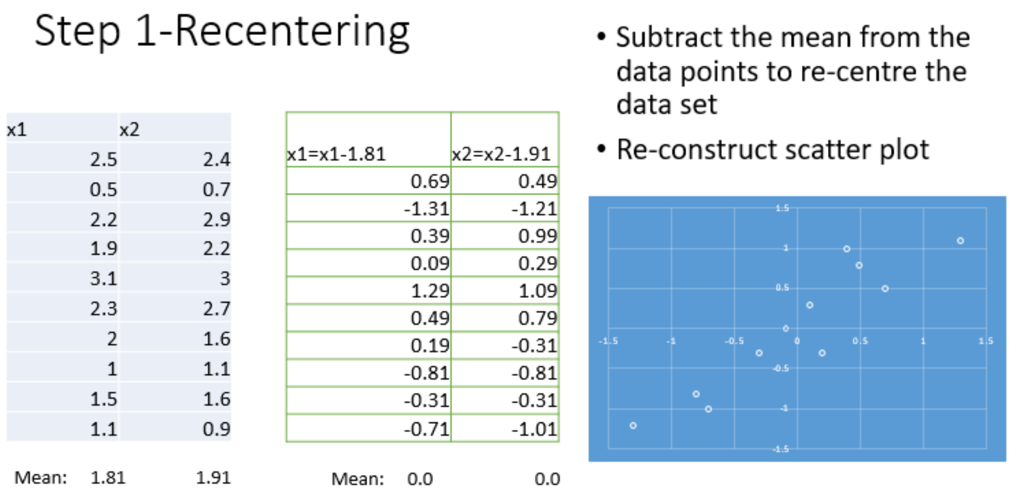
Example 1: Mathematical computation



	X1	X2
1	2.5	2.4
2	0.5	0.7
3	2.2	2.9
4	1.9	2.2
5	3.1	3
6	2.3	2.7
7	2	1.6
8	1	1.1
9	1.5	1.6
10	1.1	0.9



N=10, p=2



Step 1-Recentering

x1	x2
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9
Mean: 1.81	1.91

x1=x1-1.81	x2=x2-1.91
0.69	0.49
-1.31	-1.21
0.39	0.99
0.09	0.29
1.29	1.09
0.49	0.79
0.19	-0.31
-0.81	-0.81
-0.31	-0.31
-0.71	-1.01
Mean: 0.0	0.0

- Subtract the mean from the data points to re-centre the data set
- Re-construct scatter plot

1.5
1
0.5
0
-0.5
-1
-1.5
-1.5
-1
-0.5
0
0.5
1
1.5

Step 2-Covariance Matrix

Adjusted data matrix X

$$X = \begin{pmatrix} 0.69 & 0.49 \\ -1.31 & -0.21 \\ 0.39 & 0.99 \\ 0.09 & 0.29 \\ 1.29 & 1.09 \\ 0.49 & 0.79 \\ 0.19 & -0.31 \\ -0.81 & -0.81 \\ -0.31 & -0.31 \\ -0.71 & -0.01 \end{pmatrix}$$

10x2

- $\bar{X} = \begin{pmatrix}0\\0\end{pmatrix}$

$$C = \frac{1}{N-1}(X - 1\overline{X^T})^T(X - 1\overline{X^T})$$

$$C = \frac{1}{N-1}X^TX$$

$$C = \frac{1}{10-1}\begin{pmatrix} 0.69 & -0.31 & 0.39 & 0.09 & 1.29 & 0.49 & 0.19 & -0.81 & -0.31 & -0.71 \\ 0.49 & -0.21 & 0.99 & 0.29 & 1.09 & 0.79 & -0.31 & -0.81 & -0.31 & -0.01 \end{pmatrix}\begin{pmatrix} 0.69 & 0.49 \\ -1.31 & -0.21 \\ 0.39 & 0.99 \\ 0.09 & 0.29 \\ 1.29 & 1.09 \\ 0.49 & 0.79 \\ 0.19 & -0.31 \\ -0.81 & -0.81 \\ -0.31 & -0.31 \\ -0.71 & -0.01 \end{pmatrix}$$

2×10 10×2

$$C = \begin{pmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{pmatrix}$$

If X is not in same unit of measurements, then use correlation matrix

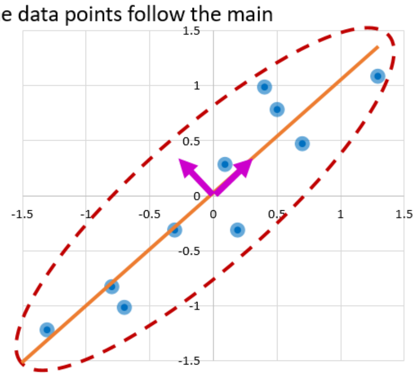
Step 3- Eigenvalues and eigenvectors

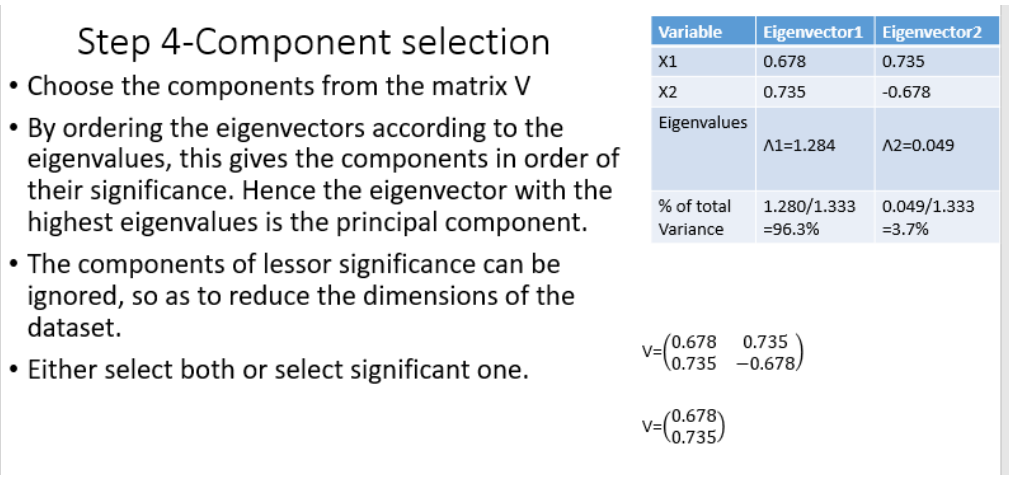
- Compute eigenvalues λi and unit or normalized eigenvectors ei of C, order the corresponding pairs from the highest to the lowest eigenvalues.
- Eigenvalues determines the radius of the ellipse
- Lines that characterise data: The first eigenvector go to the middle of the data points- best fit
- Second eigenvector gives less important pattern in the data All the data points follow the main line but are off to the side of main line by some amount.
- Total sample variance=sum of eigenvalues

$(C-\lambda I)\vec{v}=0$

Variable	Eigenvector1	Eigenvector2
X1	0.678	0.735
X2	0.735	-0.678
Eigenvalues	Λ1=1.284	Λ2=0.049
% of total Variance	1.280/1.333 =96.3%	0.049/1.333 =3.7%

1.333





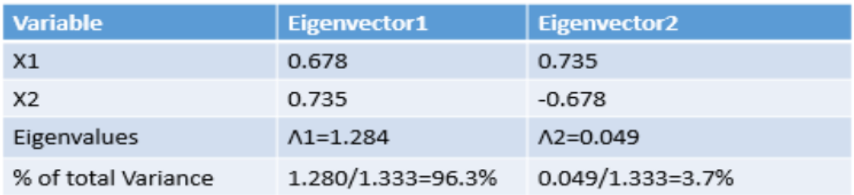
Step 4-Component selection

- Choose the components from the matrix V
- By ordering the eigenvectors according to the eigenvalues, this gives the components in order of their significance. Hence the eigenvector with the highest eigenvalues is the principal component.
- The components of lessor significance can be ignored, so as to reduce the dimensions of the dataset.
- Either select both or select significant one.

Variable	Eigenvector1	Eigenvector2
X1	0.678	0.735
X2	0.735	-0.678
Eigenvalues	Λ1=1.284	Λ2=0.049
% of total Variance	1.280/1.333 =96.3%	0.049/1.333 =3.7%

$$V=\begin{pmatrix}0.678 & 0.735\\ 0.735 & -0.678\end{pmatrix}$$

$$V=\begin{pmatrix}0.678\\ 0.735\end{pmatrix}$$



Variable	Eigenvector1	Eigenvector2
X1	0.678	0.735
X2	0.735	-0.678
Eigenvalues	Λ1=1.284	Λ2=0.049
% of total Variance	1.280/1.333=96.3%	0.049/1.333=3.7%

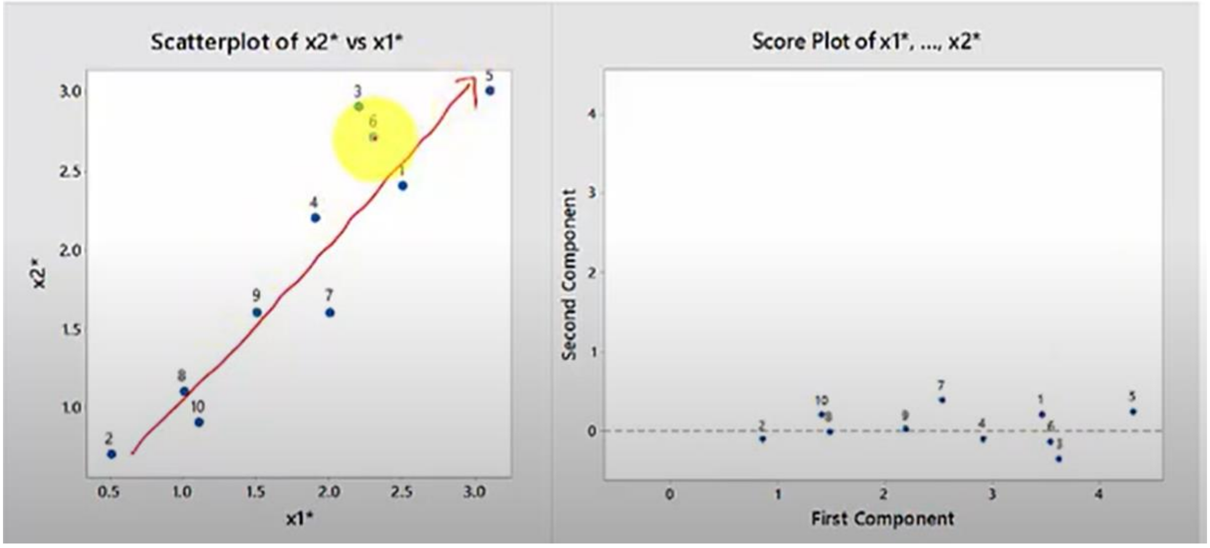
- Step 5- Derive new data set Y=XV

$$V=\begin{pmatrix}0.678\\ 0.735\end{pmatrix}$$

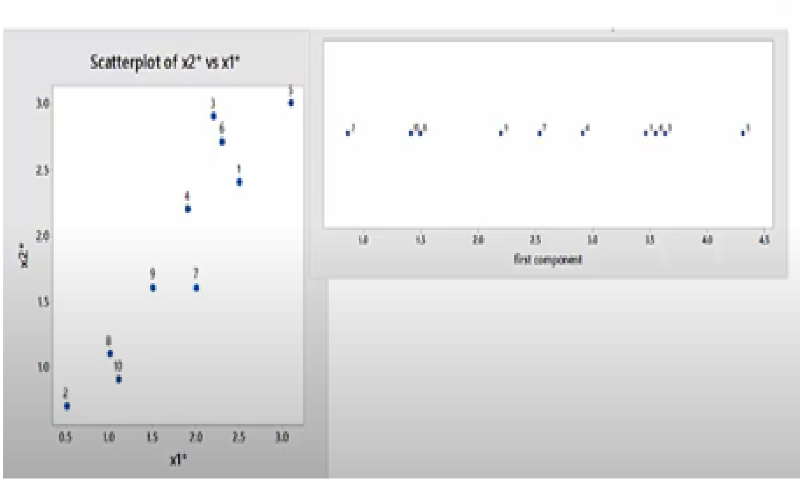
$$Y=\begin{pmatrix}2.5 & 2.4\\ 0.5 & 0.7\\ 2.2 & 2.9\\ 1.9 & 2.2\\ 3.1 & 3\\ 2.3 & 2.7\\ 2 & 1.6\\ 1 & 1.1\\ 1.5 & 1.6\\ 1.1 & 0.9\end{pmatrix}\begin{pmatrix}0.678 & 0.735\\ 0.735 & -0.678\end{pmatrix}=\begin{pmatrix}3.459 & -0.211\\ 0.854 & -0.107\\ 3.623 & -0.348\\ . & .\\ . & .\\ . & .\\ . & .\\ . & .\\ . & .\\ 1.407 & 0.199\end{pmatrix}$$

Y1=(0.678) X1+(0.735) X2
Y2=(0.735) X1+ (-0.678) X2

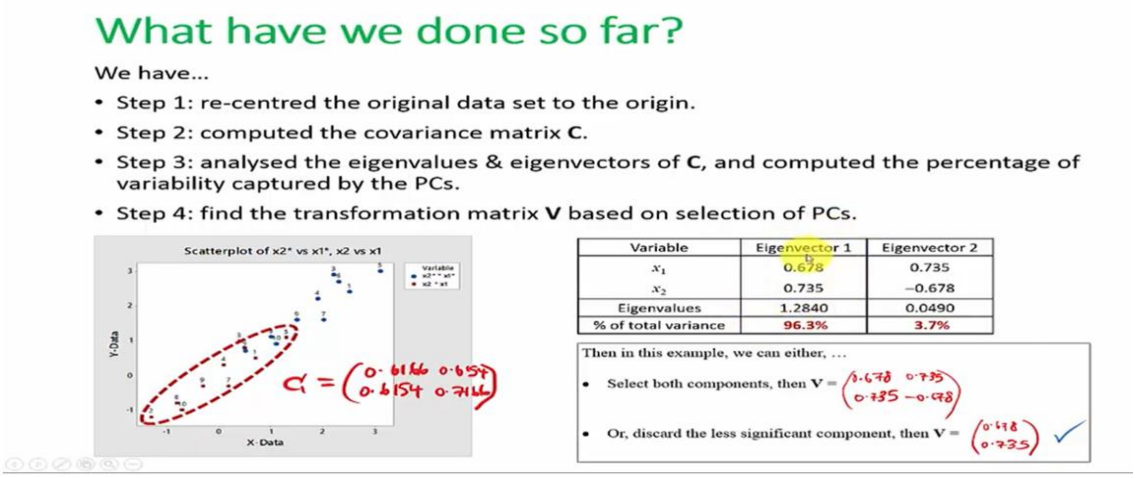
- Transformation with rotation and reflection



$$Y=\begin{pmatrix}2.5 & 2.4\\ 0.5 & 0.7\\ 2.2 & 2.9\\ 1.9 & 2.2\\ 3.1 & 3\\ 2.3 & 2.7\\ 2 & 1.6\\ 1 & 1.1\\ 1.5 & 1.6\\ 1.1 & 0.9\end{pmatrix}\begin{pmatrix}0.678\\ 0.735\end{pmatrix}=\begin{pmatrix}3.459\\ 0.854\\ 3.623\\ .\\ .\\ .\\ .\\ .\\ .\\ 1.407\end{pmatrix}$$



Y1=0.678 X1+0.735 X2



What have we done so far?

We have...

- Step 1: re-centred the original data set to the origin.
- Step 2: computed the covariance matrix **C**.
- Step 3: analysed the eigenvalues & eigenvectors of **C**, and computed the percentage of variability captured by the PCs.
- Step 4: find the transformation matrix **V** based on selection of PCs.

$C=\begin{pmatrix}0.6166 & 0.6154\\ 0.6154 & 0.7166\end{pmatrix}$

Variable	Eigenvector 1	Eigenvector 2
x_1	0.678	0.735
x_2	0.735	−0.678
Eigenvalues	1.2840	0.0490
% of total variance	96.3%	3.7%

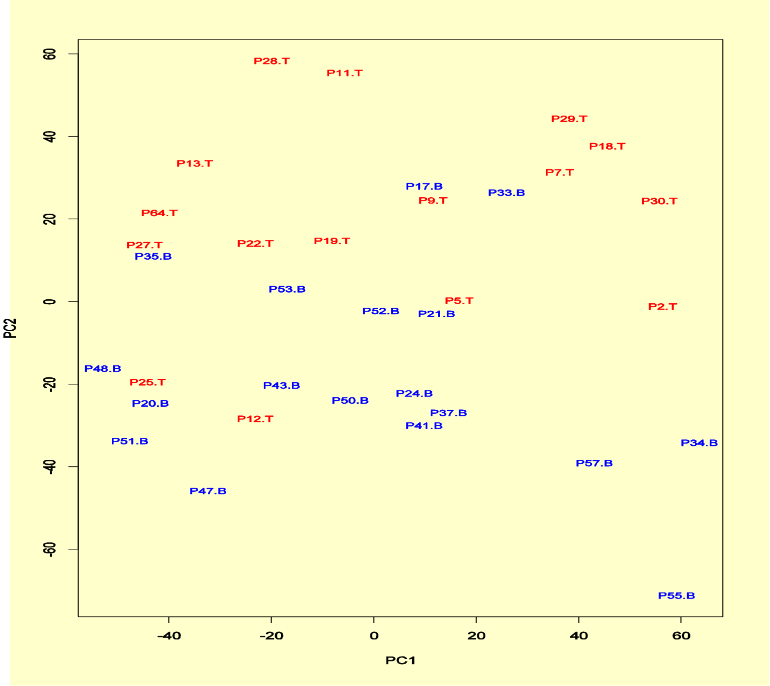
Then in this example, we can either, ...

- Select both components, then $\mathbf{V}=\begin{pmatrix}0.678 & 0.735\\ 0.735 & -0.678\end{pmatrix}$
- Or, discard the less significant component, then $\mathbf{V}=\begin{pmatrix}0.678\\ 0.735\end{pmatrix}$ ✓

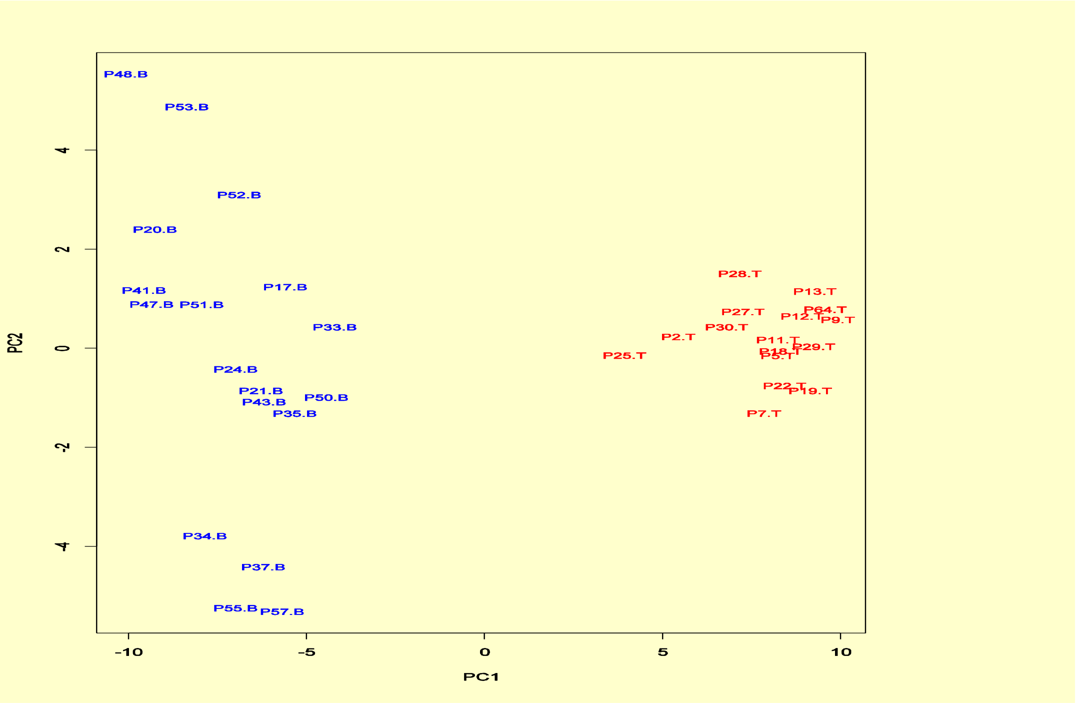
Y=XV

$$Y=\begin{pmatrix}2.5 & 2.4\\ 0.5 & 0.7\\ 2.2 & 2.9\\ 1.9 & 2.2\\ 3.1 & 3\\ 2.3 & 2.7\\ 2 & 1.6\\ 1 & 1.1\\ 1.5 & 1.6\\ 1.1 & 0.9\end{pmatrix}\begin{pmatrix}0.678\\ 0.735\end{pmatrix}=\begin{pmatrix}3.459\\ 0.854\\ 3.623\\ .\\ .\\ .\\ .\\ .\\ .\\ 1.407\end{pmatrix}$$

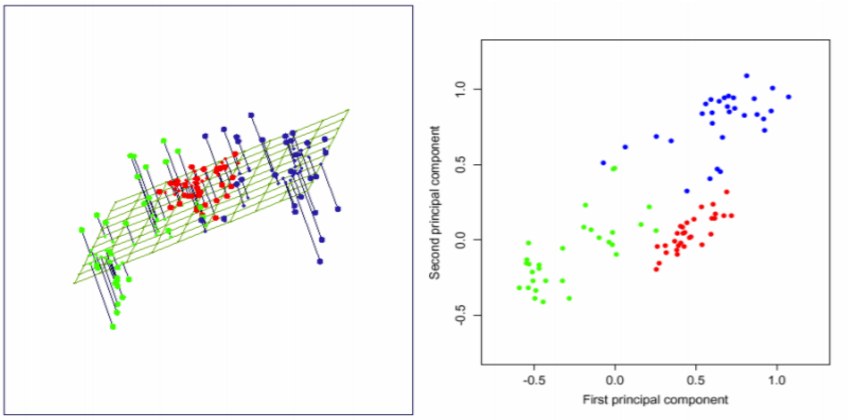
PCA on all Genes Leukemia data, precursor B and T. Plot of 34 patients, dimension of 8973 genes reduced to 2. From the plot it is obvious that good classification of the data points is not possible.



PCA on 100 top significant genes Leukemia data, precursor B and T. Plot of 34 patients, dimension of 100 genes reduced to 2. Here with 100 genes reduced to 2 component, we got good classification. This reveals that PCA cannot be directly applicable to any data set. Based on the feature patterns and information after some feature processing or reduction PCA can be applied.



Example 2: In this example we can reduce a set of 3D data to a set in only two dimensions. We do this by forming a plane from the first two principal components U_1 and U_2. We then project all of the points on to that plane. The points are shown in their reduced dimension in the figure on the right.



The best rank-two linear approximation

Real time Examples: Step by step process of different compression ie., component selection is shown below. First image with all 512 components and second drastically reduced to just 10 components and the picture quality in terms of clarity is reduced much. So in third image 30 components are used which is better than second and still need to be improved. So 50,70 and finally with 100 components, the expected quality, closer to first image is achieved.

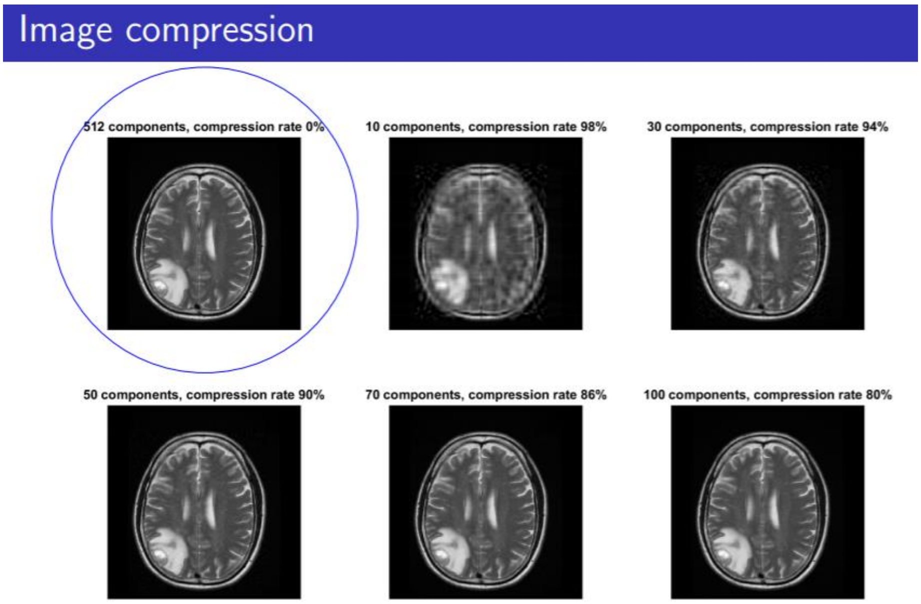


Image compression

512 components, compression rate 0%
10 components, compression rate 98%
30 components, compression rate 94%
50 components, compression rate 90%
70 components, compression rate 86%
100 components, compression rate 80%

- General about principal components
  - summary variables
  - linear combinations of the original variables
  - uncorrelated with each other
  - capture as much of the original variance as possible
  - components are Orthogonal to each other

Spectral Decomposition Theorem:

The variance-covariance matrix can be written as the sum over the *d* eigenvalues, multiplied by the product of the corresponding eigenvector times its transpose as shown in the first expression below:

$$\sum_{i=1}^{d} \boldsymbol{\lambda_i \vec{e}_i \vec{e}^{i^T}} \cong \sum_{i=1}^{p} \lambda_i \vec{e}_i \, \vec{e}^{i^T}$$

For the given example, Eigenvalues and Eigen vectors are:

$$\boldsymbol{\lambda_1 = 5.75, \lambda_2 = 475.73}$$

$$\boldsymbol{\vec{e}_1 = \begin{bmatrix} -0.999 \\ -0.035 \end{bmatrix}; \vec{e}_2 = \begin{bmatrix} 0.035 \\ -0.999 \end{bmatrix}}$$

$$\mathbf{PVE}(\boldsymbol{\lambda_1}) = \boldsymbol{5.75/(5.75 + 475.73)}$$

= 5.75/481.48=0.012*100=1.2%

$$\mathbf{PVE}(\boldsymbol{\lambda_2}) = \boldsymbol{475.73/(5.75 + 475.73)}$$

= $\boldsymbol{475.73}$ /481.48=0.988*100=98.8%

$$\mathbf{PVE}(\boldsymbol{\lambda_2}) > \mathbf{PVE}(\boldsymbol{\lambda_1})$$

And **PVE($\boldsymbol{\lambda_2}$) explains 98.8%, so, let the p value be 1.**

PC1 is formed with $\boldsymbol{\lambda_2}$. So for RHS, use $\boldsymbol{\lambda_2\ and\ \vec{e}_2}$

LHS:

$$\boldsymbol{\lambda_1 \vec{e}_1 \vec{e}_1{}^T + \lambda_2 \vec{e}_2 \vec{e}_2{}^T}$$

$$\boldsymbol{5.75 \begin{bmatrix} -0.999 \\ -0.035 \end{bmatrix} [-0.999 \quad -0.035]} + \boldsymbol{475.73 \begin{bmatrix} 0.035 \\ -0.999 \end{bmatrix} [0.035 \quad -0.999]}$$

$=5.75\begin{bmatrix}0.998 & 0.035\\0.035 & 0.001\end{bmatrix}+475.73\begin{bmatrix}0.001 & -0.035\\-0.035 & 0.998\end{bmatrix}+$

$=\begin{bmatrix}5.74 & 0.2\\0.2 & 0.006\end{bmatrix}+\begin{bmatrix}0.476 & -16.65\\-16.65 & 474.78\end{bmatrix}=\begin{bmatrix}6.22 & -16.45\\-16.45 & 474.79\end{bmatrix}$

Now RHS using first eigenvalue and its eigenvector.

PC1 is formed with $\boldsymbol{\lambda_2}$. PVE($\boldsymbol{\lambda_2}$) explains 98.8%

RHS=$\boldsymbol{\lambda_2 \vec{e}_2 \vec{e}_2{}^T}$

$=\mathbf{475.73}\begin{bmatrix}\mathbf{0.035}\\\mathbf{-0.999}\end{bmatrix}[\mathbf{0.035}\quad \mathbf{-0.999}]$

$=475.73\begin{bmatrix}0.001 & -0.035\\-0.035 & 0.998\end{bmatrix}$

$=\begin{bmatrix}0.476 & -16.65\\-16.65 & 474.78\end{bmatrix}$

LHS=RHS

This proof shows that the component selection is correct, since d component selection and p component selection values are approximately equal.

Eigenvalue verification: The total variation of X as the trace of the variance-covariance matrix, that is the sum of the variances of the individual variables. This is also equal to the sum of the eigenvalues as shown below:

$$Covariance\ matrix\ = \begin{pmatrix}0.6166 & 0.6154\\0.6154 & 0.7166\end{pmatrix}$$

Eigen values λ_1=1.284 and λ_2= 0.049.

Trace (Covariance Matrix) = summation of Eigenvalues

$\sum_{i=1}^{d} X_{ii} = \sum_{i=1}^{d} \lambda_i$

LHS : 0.6166+0.7166 =1.3332

RHS: 1.284+0.049=1.333

LHS=RHS

Components Vs. Features : Example

- We will use the Places Rated Almanac data (Boyer and Savageau) which rates 329 communities according to nine criteria:

1. Climate and Terrain
2. Housing
3. Health Care & Environment
4. Crime
5. Transportation
6. Education
7. The Arts
8. Recreation
9. Economics

Component	**Eigenvalue**	**Proportion**	**Cumulative**
1	0.3775	0.7227	0.7227
2	0.0511	0.0977	0.8204
3	0.0279	0.0535	0.8739
4	0.0230	0.0440	0.9178
5	0.0168	0.0321	0.9500
6	0.0120	0.0229	0.9728
7	0.0085	0.0162	0.9890
8	0.0039	0.0075	0.9966
9	0.0018	0.0034	1.0000
Total	0.5225		

- The correlations between the principal components and the original variables are copied into the following table for the Places Rated Example. You will also note that if you look at the principal components themselves, then there is zero correlation between the components.
- Cor(PC_i,X) using only 3 components out of 9, since only 3 components are selected.

Variable	**1**	**2**	**3**
Climate	0.190	0.017	0.207
Housing	**0.544**	0.020	0.204
Health	**0.782**	**-0.605**	0.144

Crime	0.365	0.294	**0.585**
Transportation	**0.585**	0.085	0.234
Education	0.394	-0.273	0.027
Arts	**0.985**	0.126	-0.111
Recreation	**0.520**	0.402	**0.519**
Economy	0.142	0.150	0.239

Interpretation of the principal components is based on finding which variables are most strongly correlated with each component, i.e., which of these numbers are large in magnitude, the farthest from zero in either direction.

Which numbers we consider to be large or small is of course is a subjective decision.

You need to determine at what level the correlation is of importance.

Here a correlation above 0.5 is deemed important.

These larger correlations are in boldface in the table.

Correlation PCA Vs. Covariance PCA:

If the features have varying scales of unstructured data then apply correlation in place of covariance. Structures in the sense the values of features are in same scale. For structured data either Covariance or Correlation can be applied. But most of the studies need to know the variance to interpret data. At the same time when Covariance is applied for unstructured data we don’t get perfect eigenvalues and vectors and we cannot get the significant variance explained by components. Hence for unstructured data use correlation. Whatever the scale values of data, the correlation checks the relationship between features, so irrespective data scale, correlation provides the perfect eigenvalues, eigenvectors and principal components.

	data	Normalize	eigen value	eigen vector
COV	structured	yes	4,4,4,4	0.707,-0.707,0.707,0.707
COV	unstructured	Yes	0,40004	0.9995,0.0099,0.0099,0.999
COR	structured	yes	2,0	0.707,-0.707,0.707,0.707
COR	structured	No	2,0	0.707,-0.707,0.707,0.707

COR	unstructured	Yes	2,0	0.707,-0.707,0.707,0.707
COR	unstructured	No	2,0	0.707,-0.707,0.707,0.707