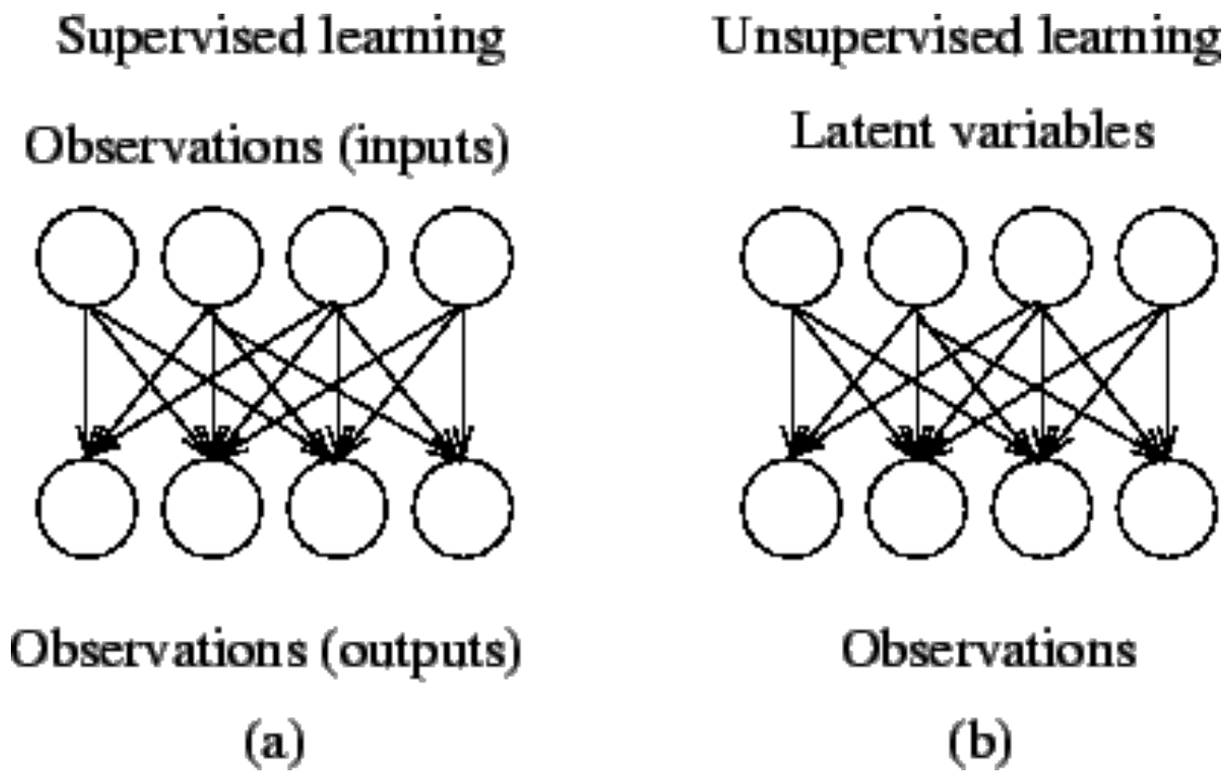
Principal Components Analysis (PCA)

Outline

- variance, covariance and covariance matrix
- PCA algorithm
- How to select K
- PCA pros and cons

Unsupervised Learning

- **Unsupervised learning** is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses



Variance, Covariance, Covariance Matrix

Variance is a measure of the spread of data in a data set. In fact it is almost identical to the standard deviation. (1-dimension)

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)}$$

Covariance is always measured between 2 dimensions.If you calculate the covariance between one dimension and itself, you get the variance. covariance between the x and y dimensions, the y and z dimensions, and the x and z dimensions.

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

The definition for the covariance matrix for a set of data with dimensions is:

$$C^{n\times n} = (c_{i,j},\ c_{i,j} = cov(Dim_i, Dim_j)),$$

An example. We'll make up the covariance matrix for an imaginary 3 dimensional data set, using the usual dimensions x, y and z. Then, the covariance matrix has 3 rows and 3 columns, and the values are this:

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

Principal Components Analysis

- PCA: an unsupervised learning tool used for data visualization or data pre-processing before supervised techniques are applied
- PCA: produces a low-dimensional representation of dataset (variable reduction)
- Procedure
  - center the data at 0 by subtracting the mean from each variable
  - compute the variances of the data
  - compute the covariance matrix of the data
  - find the eigenvector of the matrix (principal component) : orthogonal directions of highest variability
  - calculate the eigenvalues: magnitude of variance along the principal components
- PCA: transform data to k selected dimensions that preserve as much the original structure as possible

Principal Components Analysis

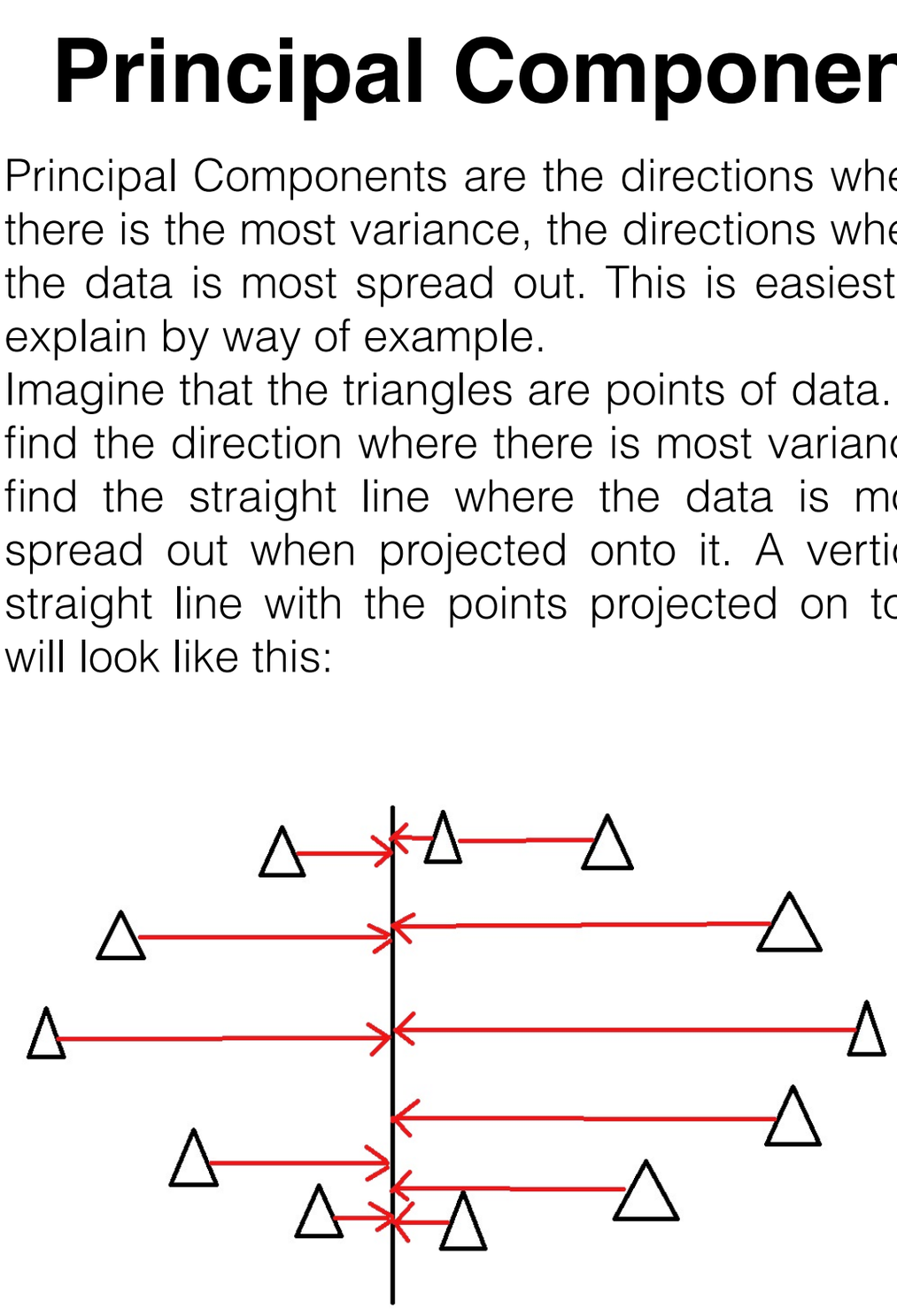
- PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.
- Consider a data matrix, X, with column-wise zero empirical mean (the sample mean of each column has been shifted to zero), where n is number of samples, and p is number of features.
- Let's say we have a set of predictors as X^1, X^2…,X^p, The principal component can be written as:

$$Z^1 = \Phi^{11}X^1 + \Phi^{21}X^2 + \Phi^{31}X^3 + ... + \Phi p^1 X^p$$

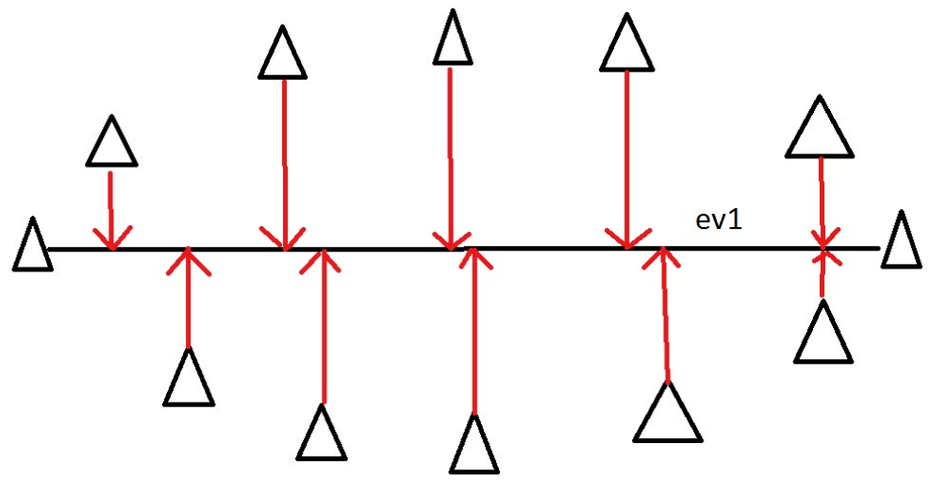
- Z^1 is first principal component
- Φp^1 is the loading vector comprising of loadings (Φ^1, Φ^2..) of first principal component. The loadings are constrained to a sum of square equals to 1. This is because large magnitude of loadings may lead to large variance. It also defines the direction of the principal component (Z^1) along which data varies the most. It results in a line in p dimensional space which is closest to the n observations. Closeness is measured using average squared euclidean distance.
- $X^1..X^p$ are standardized predictors. Standardized predictors have mean equals to zero and standard deviation equals to one.

Principal Components

Principal Components are the directions where there is the most variance, the directions where the data is most spread out. This is easiest to explain by way of example.
Imagine that the triangles are points of data. To find the direction where there is most variance, find the straight line where the data is most spread out when projected onto it. A vertical straight line with the points projected on to it will look like this:



The data isn't very spread out here, therefore it doesn't have a large variance. It is probably not the principal component.
A horizontal line are with lines projected on will look like this:



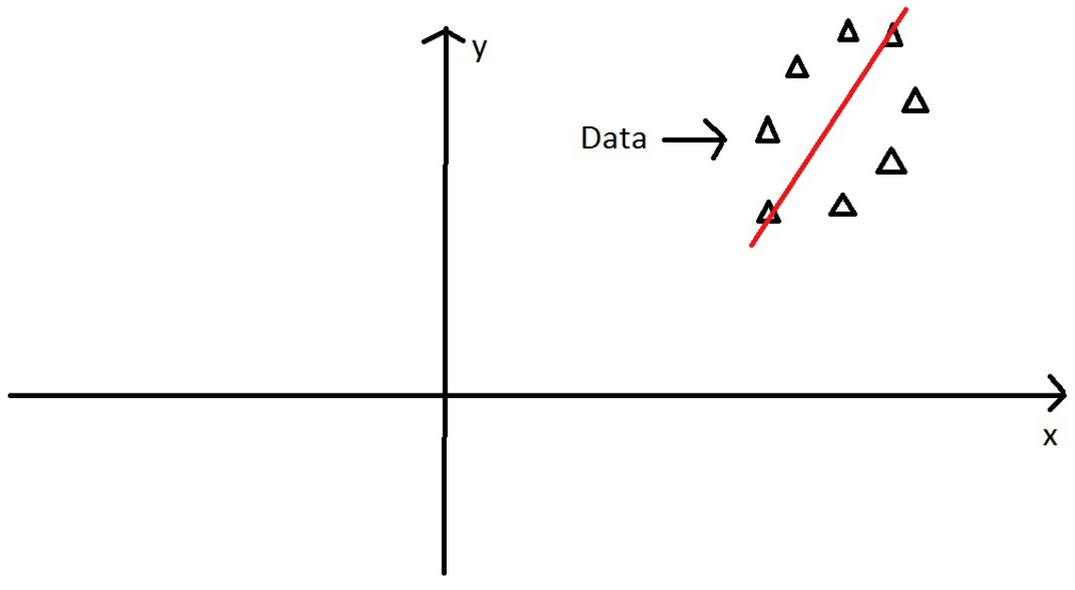
On this line the data is way more spread out, it has a large variance. In fact there isn't a straight line you can draw that has a larger variance than a horizontal one. A horizontal line is therefore the principal component in this example.

Eigenvectors and Eigenvalues

- When we get a set of data points, like the triangles above, we can deconstruct the set into eigenvectors and eigenvalues.
- Eigenvectors and values exist in pairs: every eigenvector has a corresponding eigenvalue. An eigenvector is a direction, in the example above the eigenvector was the direction of the line (vertical, horizontal, 45 degrees etc.) .
- An eigenvalue is a number, telling how much variance there is in the data in that direction, in the example above the eigenvalue is a number telling us how spread out the data is on the line.
- The eigenvector with the highest eigenvalue is therefore the principal component.
- In fact the amount of eigenvectors/values that exist equals the number of dimensions the data set has.

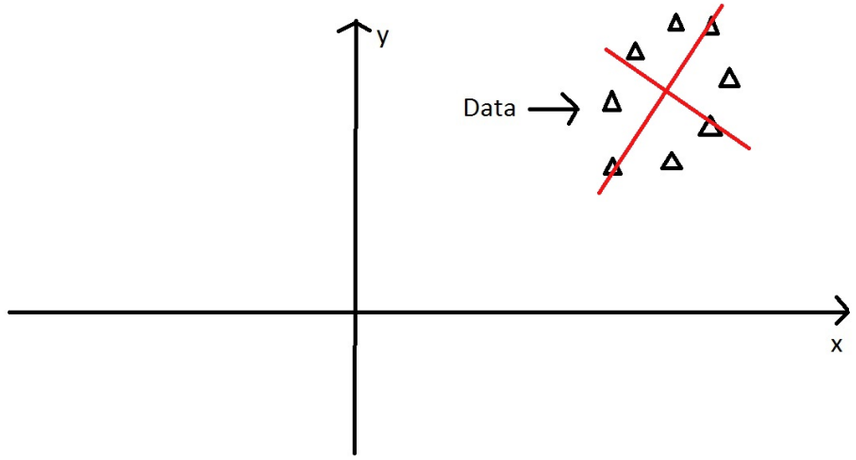
For example measuring age and hours on the internet. there are 2 variables, it's a 2 dimensional data set, therefore there are 2 eigenvectors/values. The reason for this is that eigenvectors put the data into a new set of dimensions, and these new dimensions have to be equal to the original amount of dimensions. This sounds complicated, but again an example should make it clear.

At the moment the oval is on an x-y axis. x could be age and y hours on the internet. These are the two dimensions that my data set is currently being measured in. Now remember that the principal component of the oval was a line splitting it longways. The first eigenvector looks like this:

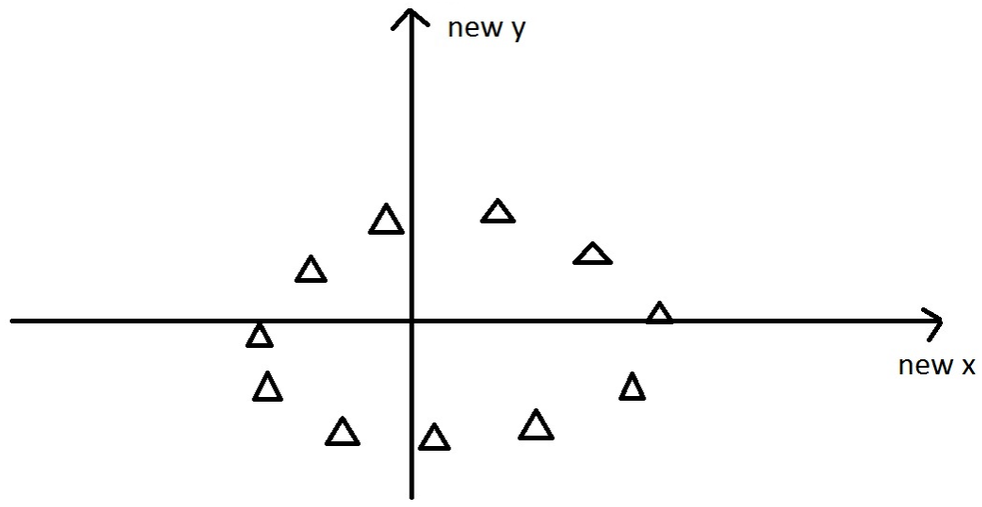


Eigenvectors and Eigenvalues

It turns out the other eigenvector is perpendicular to the first principal component. As we said, the eigenvectors have to be able to span the whole x-y area, in order to do this (most effectively), the two directions need to be orthogonal (i.e. 90 degrees) to one another. This why the x and y axis are orthogonal to each other in the first place. It would be really awkward if the y axis was at 45 degrees to the x axis. So the second eigenvector would look like this:



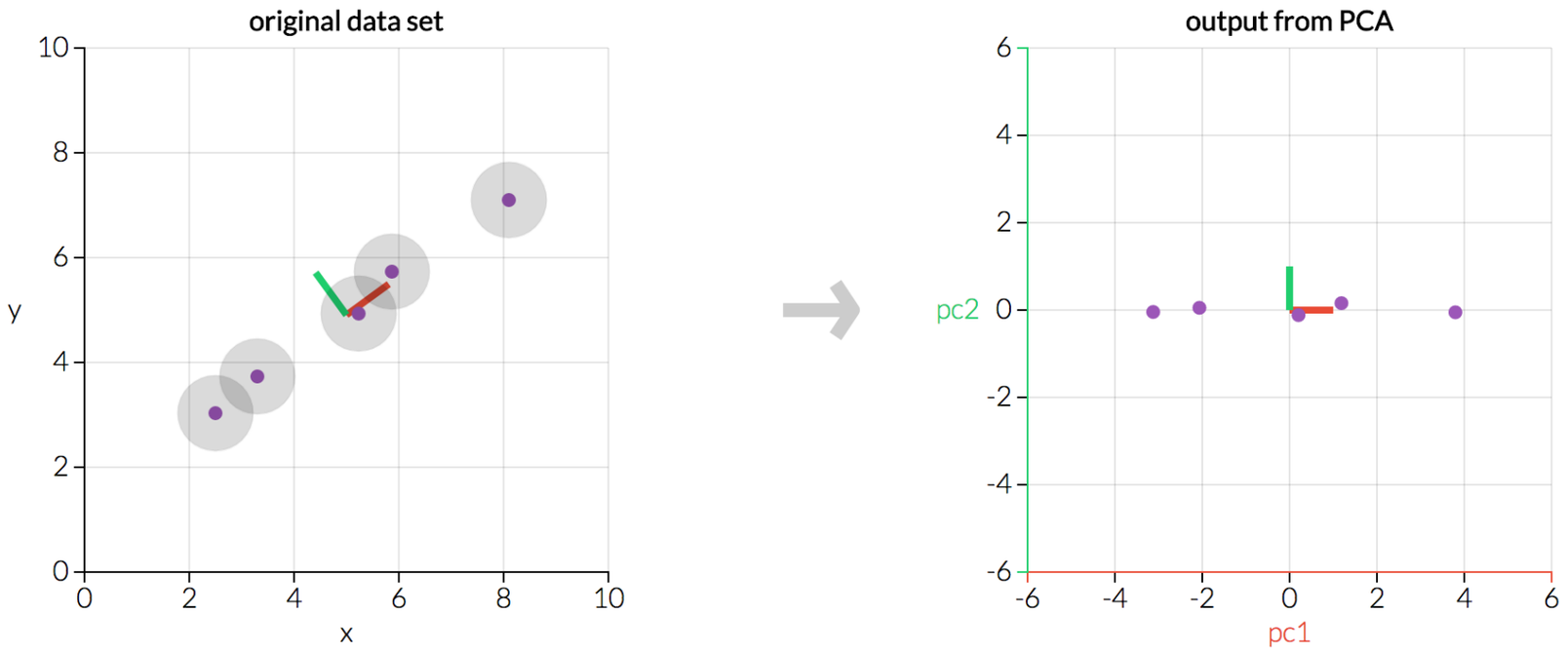
The eigenvectors have given us a much more useful axis to frame the data in. We can now re-frame the data in these new dimensions. It would look like this:



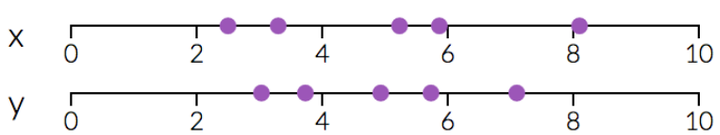
Note that nothing has been done to the data itself. We're just looking at it from a different angle. So getting the eigenvectors gets you from one set of axes to another. These axes are much more intuitive to the shape of the data now. These directions are where there is most variation, and that is where there is more information (think about this the reverse way round. If there was no variation in the data [e.g. everything was equal to 1] there would be no information, it's a very boring statistic – in this scenario the eigenvalue for that dimension would equal zero, because there is no variation).

Dimension reduction: 2D example

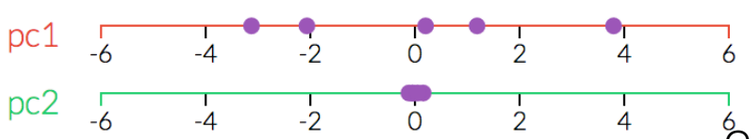
- First, consider a dataset in only two dimensions, like (height, weight). This dataset can be plotted as points in a plane. But if we want to tease out variation, PCA finds a new coordinate system in which every point has a new (x,y) value. The axes don't actually mean anything physical; they're combinations of height and weight called "principal components" that are chosen to give one axes lots of variation.
- Drag the points around in the following visualization to see PC coordinate system adjusts.



PCA is useful for eliminating dimensions. Below, we've plotted the data along a pair of lines: one composed of the x-values and another of the y-values.

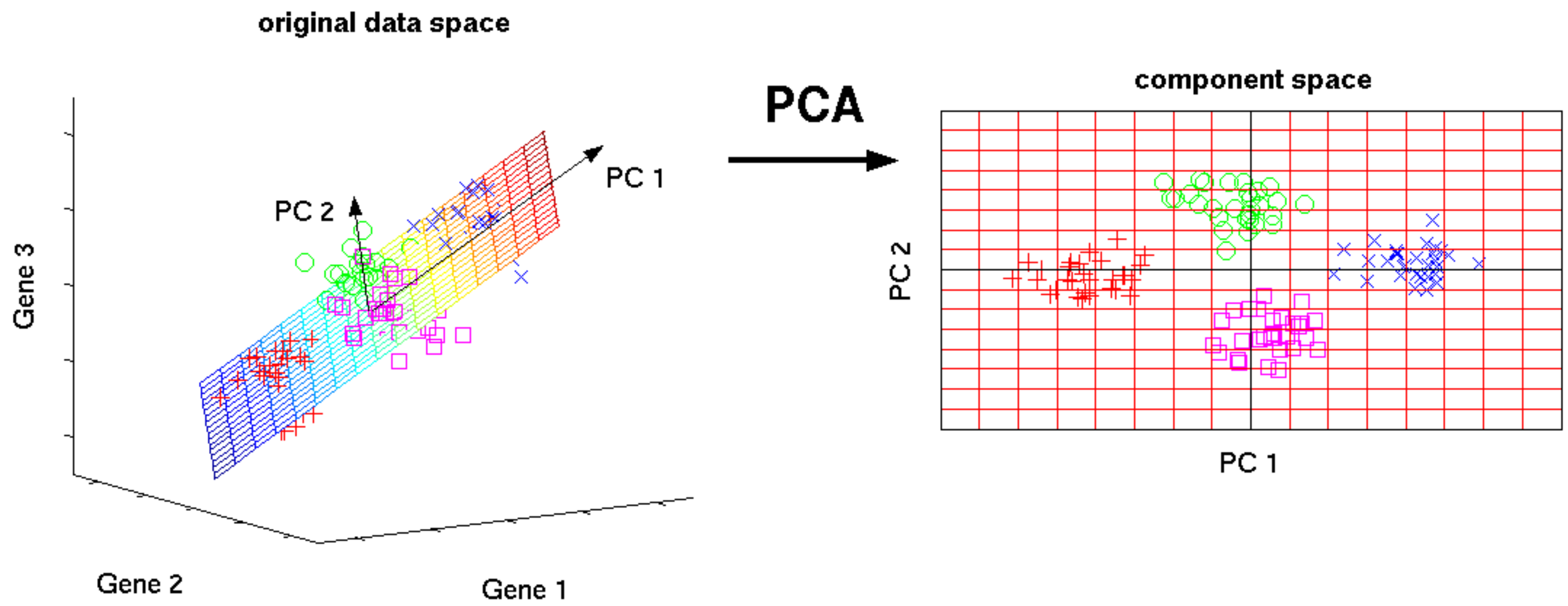


If we're going to only see the data along one dimension, though, it might be better to make that dimension the principal component with most variation. We don't lose much by dropping PC2 since it contributes the least to the variation in the data set.



Dimension reduction: 3D example

The image below shows the transformation of a high dimensional data (3 dimension) to low dimensional data (2 dimension) using PCA. Not to forget, each resultant dimension is a linear combination of p features.

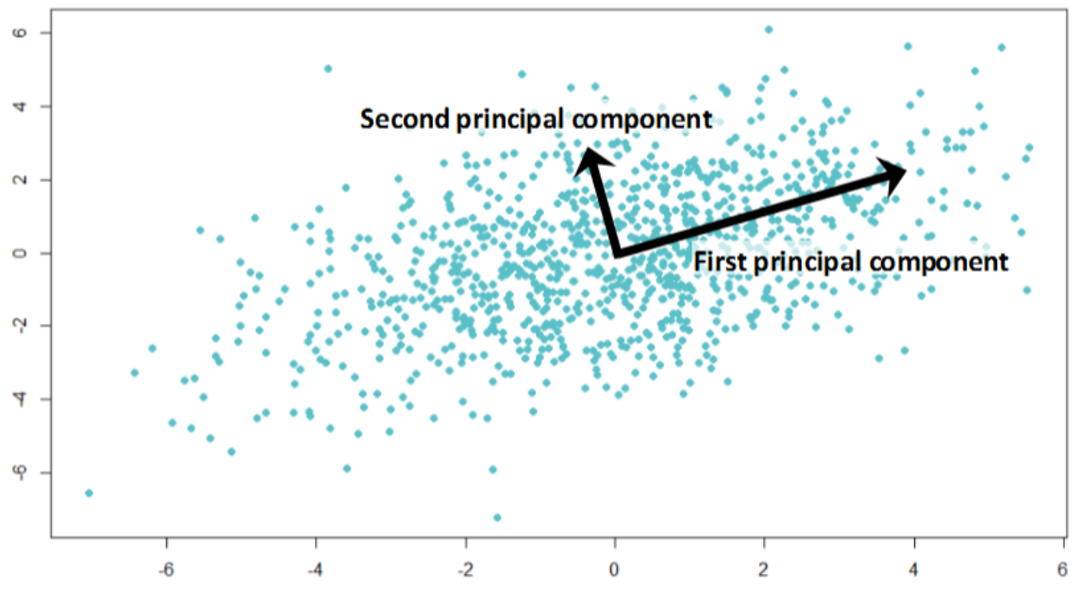


PCA

- The first principal component of a data set is the normalized linear combination of the features that has the largest variance
- The first principal component results in a line which is closest to the data i.e. it minimizes the sum of squared distance between a data point and the line
- The second principal component is the normalized linear combination of the features that accounts for as much of the remaining variation as possible, with the constraint that the correlation between the first and second component is 0
- The second component can be represented as:

$$Z^2 = \Phi^{12}X^1 + \Phi^{22}X^2 + \Phi^{32}X^3 + ... + \Phi p2X^p$$

- If the two components are uncorrelated, their directions should be orthogonal. This image is based on a simulated data with 2 predictors. Notice the direction of the components, as expected they are orthogonal. This suggests the correlation between these components is zero.
- All subsequent principal components have this same property – they are linear combinations that account for as much of the remaining variation as possible and they are not correlated with the other principal components

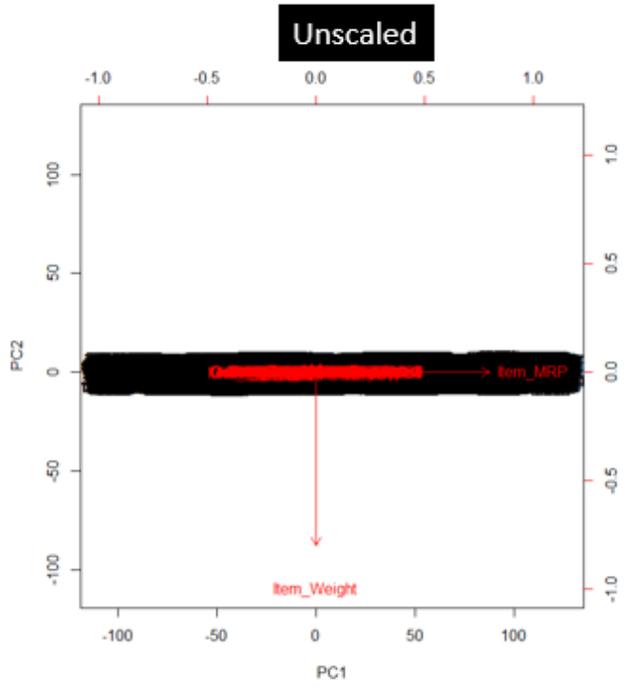


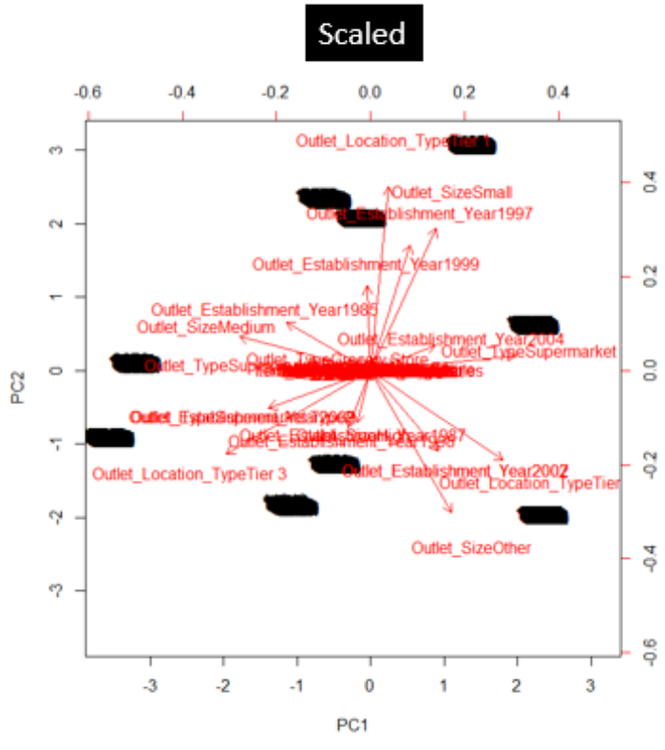
Why is standardization of features necessary ?

- The principal components are supplied with normalized version of original predictors. This is because, the original predictors may have different scales. For example: imagine a data set with variables' measuring units as gallons, kilometers, light years etc. So the scale of variances in these variables will be different.

- Performing PCA on un-normalized variables will lead to insanely large loadings for variables with high variance. In turn, this will lead to dependence of a principal component on the variable with high variance. This is undesirable.

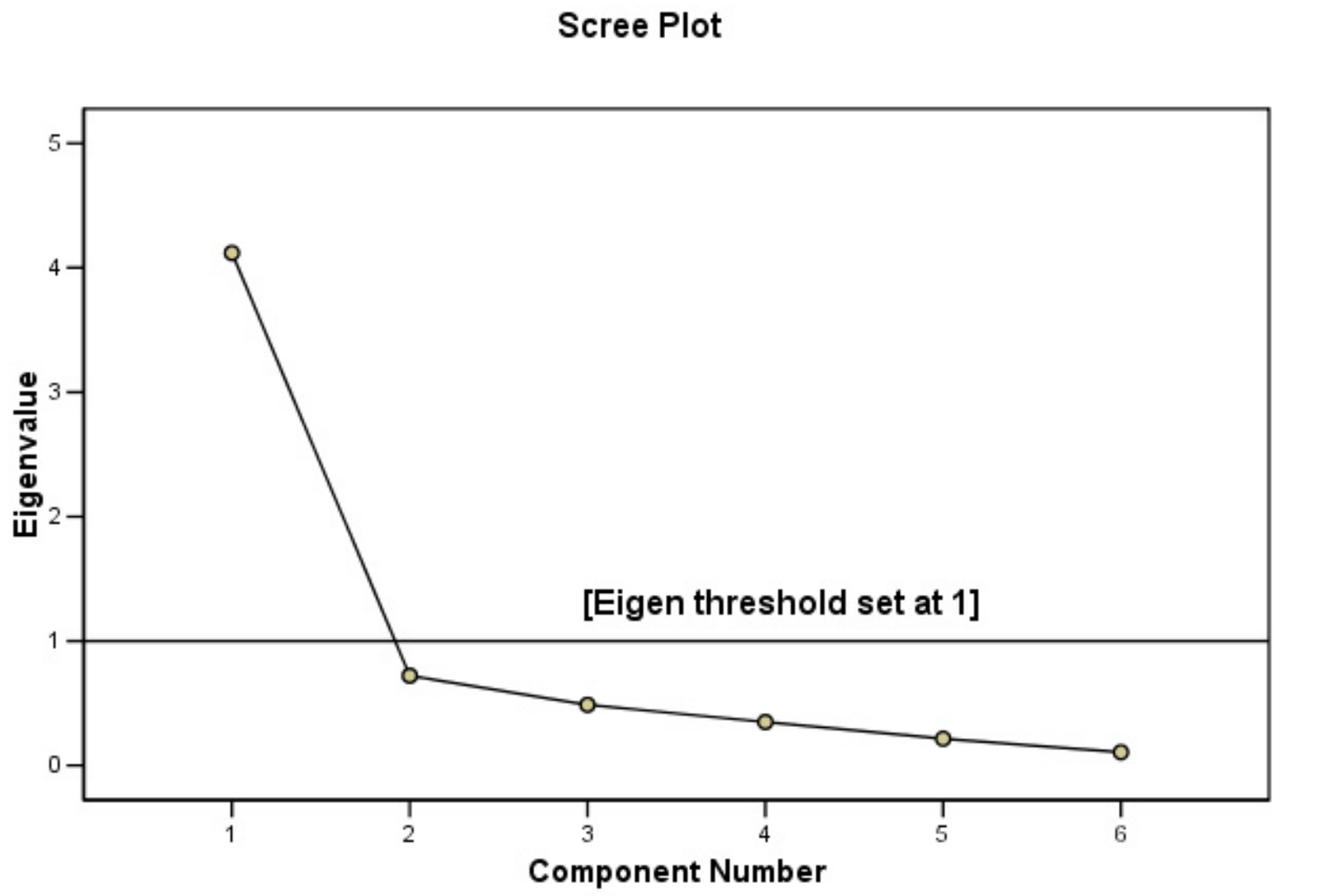
- As shown in image below, PCA was run on a data set twice (with unscaled and scaled predictors). This data set has ~40 variables. You can see, first principal component is dominated by a variable Item_MRP. And, second principal component is dominated by a variable Item_Weight. This domination prevails due to high value of variance associated with a variable. When the variables are scaled, we get a much better representation of variables in 2D space.





How many principal components should we use?

- The “scree plot” can be used as a guide: we look for an “elbow”



Summary

- PCA is used to overcome features redundancy in a data set.
- These features are low dimensional in nature.
- These principal components are a resultant of normalized linear combination of original predictor variables.
- These components aim to capture as much information as possible with high explained variance.
- The first component has the highest variance followed by second, third and so on.
- The components must be uncorrelated.
- Normalizing data becomes extremely important when the predictors are measured in different units.
- PCA works best on data set having 3 or higher dimensions.
- PCA is applied on a data set with numeric variables.
- PCA is a tool which helps to produce better visualizations of high dimensional data.

PCA advantages and disadvantages

PCA's key advantages are its low noise sensitivity, the decreased requirements for capacity and memory, and increased efficiency given the processes taking place in a smaller dimensions.

The complete advantages of PCA for image processing are listed below:

- Reduced complexity in images' grouping with the use of PCA
- Smaller database representation since only the trainee images are stored in the form of their projections on a reduced basis
- Reduction of noise since the maximum variation basis is chosen and so the small variations in the background are ignored automatically

Two key disadvantages of PCA are:

- The covariance matrix is difficult to be evaluated in an accurate manner
- Original data structure is lost