proschool
An ims Initiative

Principal Component Analysis

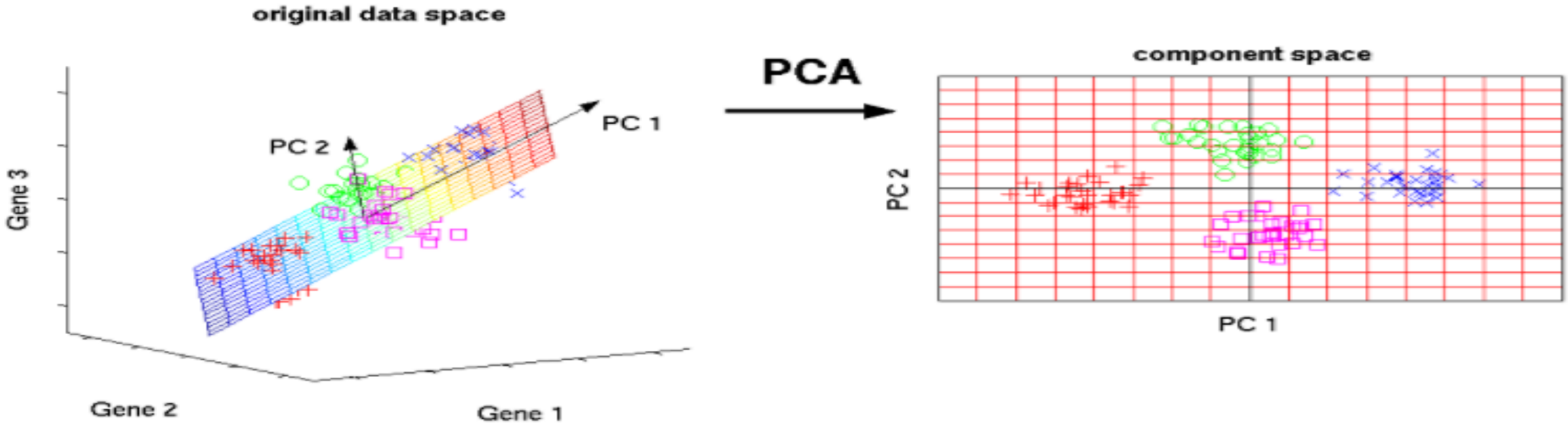
Principal Component Analysis (PCA)

- Principal Component Analysis is a **dimensionality reduction technique**
- PCA aims at reducing a large set of variables to a small set of variables to capture the underlying variance of the data without much loss of information
- Reduced set of variables, which are called principal components . A reduced set is much easier to analyze and interpret
- PCA is a statistical procedure that uses an orthogonal linear transformation to convert a set of observations of **possibly correlated variables** into a set of values of **linearly uncorrelated variables** called principal components to a **new coordinate system**

Example

- **Example :** Transformation of a high dimensional data to low dimensional data (2 dimension) using PCA .

- It does this using a linear combination (weighted average) of a set of variables



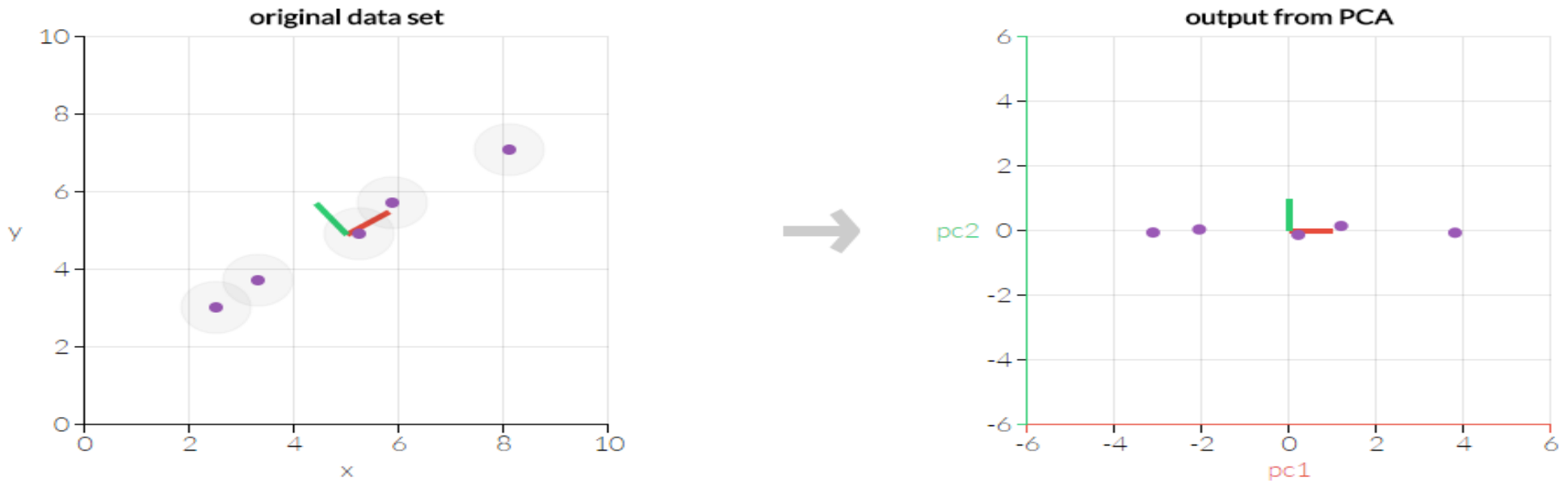
PCA Pointers

- PCA, at its core, doesn't select the most "important" features.
- What it does is a linear transformation of your data to a new coordinate system where the first component direction is the one which has the largest variance ,same for second, third...etc
- This leads to the issue of component interpretability. i.e because your components are these strange combinations of real variables.
- It becomes very difficult to identify a real world quantity with your component.

Principal Components

- 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.



Terms

- Principle components are the eigenvectors of the covariance matrix.
- Before we talk more about PCA there are few important topics we have to know
  - **What is Covariance matrix**
  - **What is Eigenvalues & eigenvectors**

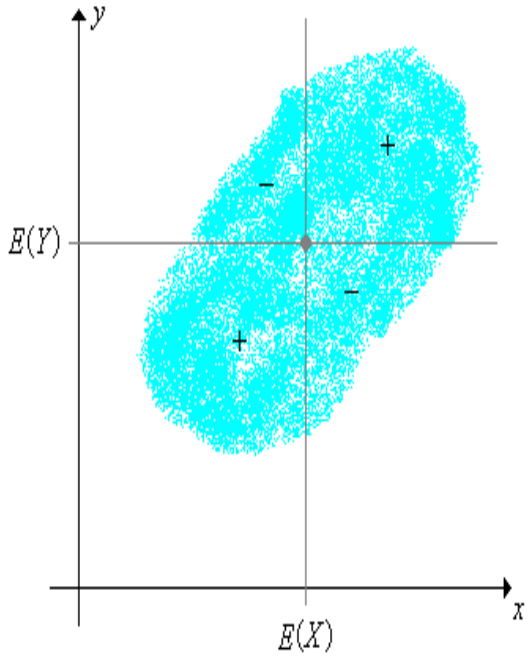
Covariance

Covariance is the measure of how two different variables change together

$$\mathrm{cov}(X,Y)=E([X-E(X)][Y-E(Y)])$$

What is understood by variance in several dimensions ("total variance") is simply a sum of variances in each dimension.

Mathematically, it's a trace of the covariance matrix: trace is simply a sum of all diagonal elements.



Covariance

- Representing Covariance between dimensions as a matrix e.g. for 3 dimensions:

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

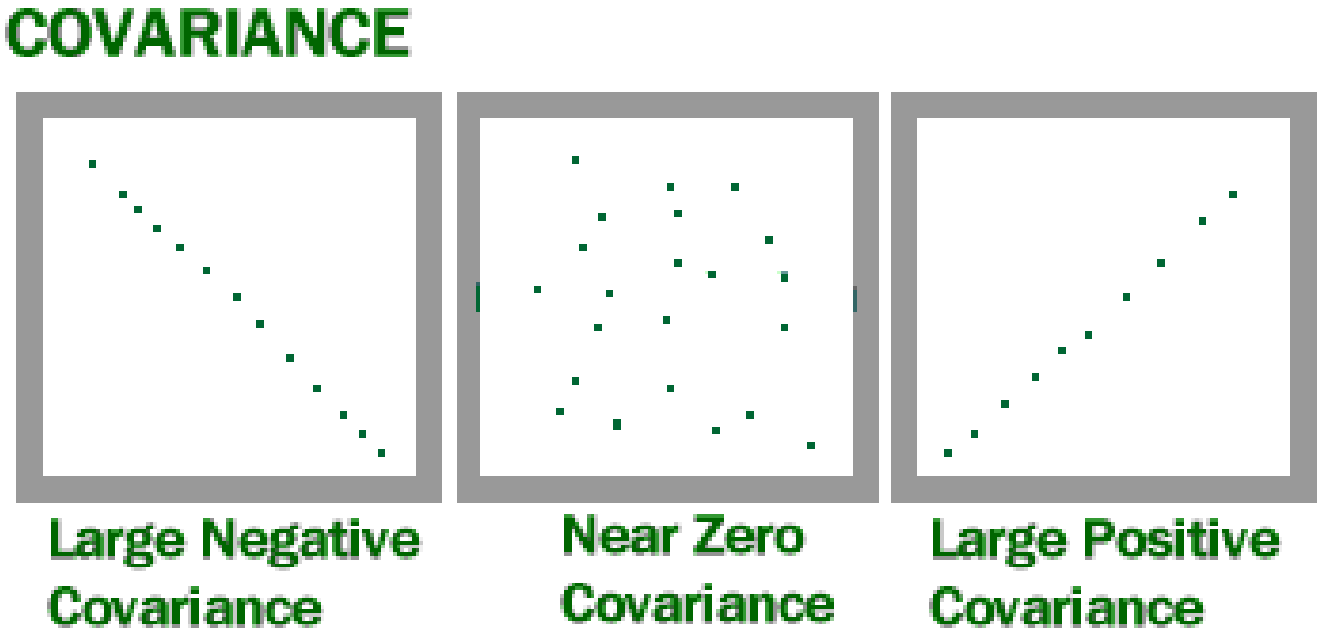
- N-dimensional data will result in **nxn** covariance matrix

- Covariance: measures the correlation between X and Y

  - **Cov(X,Y)=0: independent**
  - **Cov(X,Y)>0: move same direction**
  - **Cov(X,Y)<0: move opposite direction**

Covariance

- If covariance is positive, both dimensions increase together.
- If negative, as one increases, the other decreases.
- Zero: independent of each other.



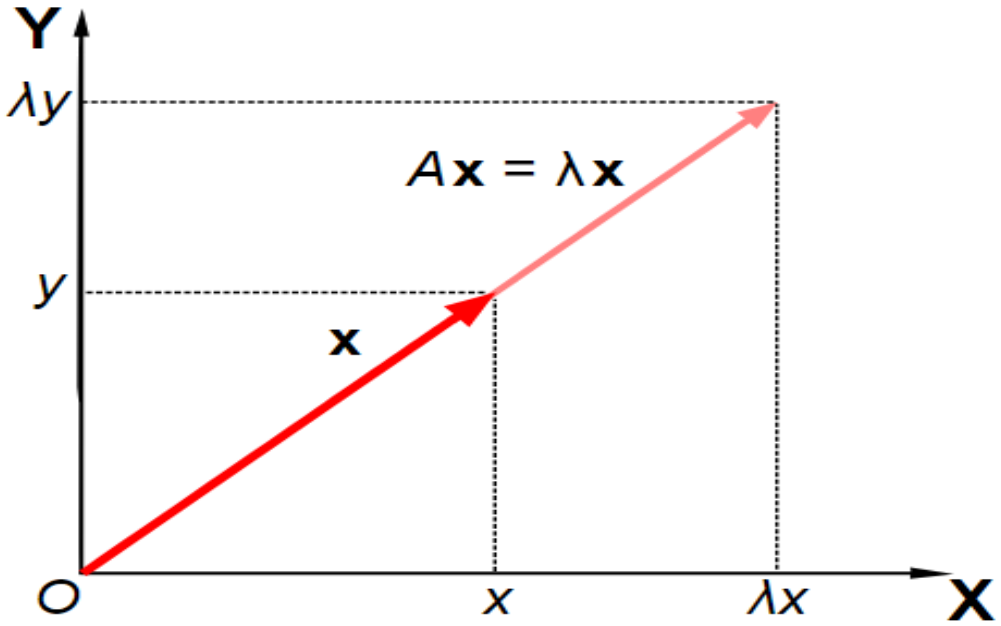
Eigenvalues & Eigenvectors

- An **eigenvector** was the direction of the line (vertical, horizontal, 45 degrees , 30 degrees ..etc.)
- An **eigenvalue** is a number, telling you how much variance there is in the data in that direction
- The eigenvalues explain the variance of the data along the new feature axes.
- Eigenvectors and values exist in pairs: every eigenvector has a corresponding eigenvalue
- The eigenvector with the highest eigenvalue is therefore the **principal component**.

Eigenvalues & Eigenvectors

- For matrix A, vectors x (column vector) having same direction as Ax :

- Eigenvectors of A is **x** such that $Ax=\lambda x$,
  - then x is an eigenvector of the linear transformation A
  - and the **scale factor λ** is the eigenvalue corresponding to that eigenvector



Eigenvalues & Eigenvectors

- λ may be negative, in which case the eigenvector reverses direction as part of the scaling
- In fact the amount of eigenvectors/values that exist equals the number of dimensions the data set has.
- There are 2 variables, it's a 2 dimensional data set, therefore there are 2 eigenvectors/values. There's 3 variables, 3-D data set, so 3 eigenvectors/values.
- The leading eigenvector of covariance matrix gives the direction of maximal variance.
- Second eigenvector gives the direction of maximal variance under an additional constraint that it should be orthogonal to the first eigenvector
- 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.

What Are Principal Components?

- The principal components produces a linear combinations or dimensions of the data that are really high in variance and that are uncorrelated

- When two variables are really correlated with each other, one new variable (ie the first principle component) can really summarize both of those two variables very well

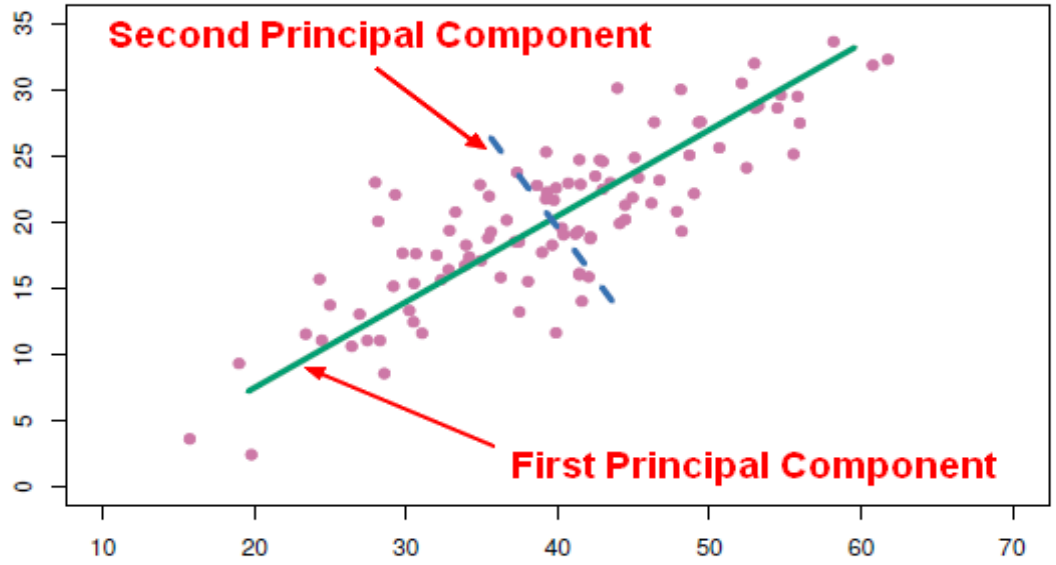
- The variables with unusually large variances will get selected.

PC1 & PC2

- **First Principal component** : Linear combination of original predictor variables which captures the maximum variance in the data set. Larger the variability captured in first component

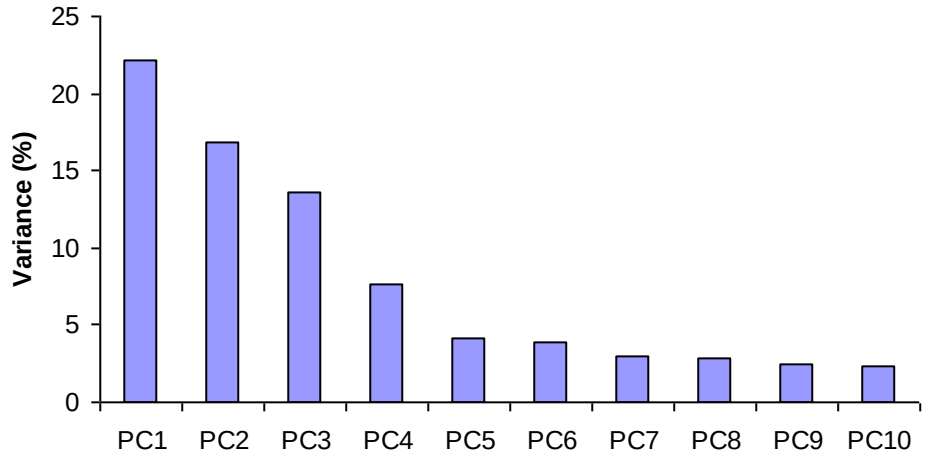
- **Second principal component :** is also a linear combination of **original predictors** which captures the remaining variance in the data set and is **uncorrelated with first PC.**

- No other component can have variability higher than first principal component. The direction along which there is greatest variation



Components

- The direction with maximum variation left in data, orthogonal to the first PC
- In general, only few directions manage to capture most of the variability in the data
- How much each PC explaned variance we can see for each dataset
- We can consider first 4 PC beacuse it explains most of the data



Steps for PCA

- Steps to perform the PCA

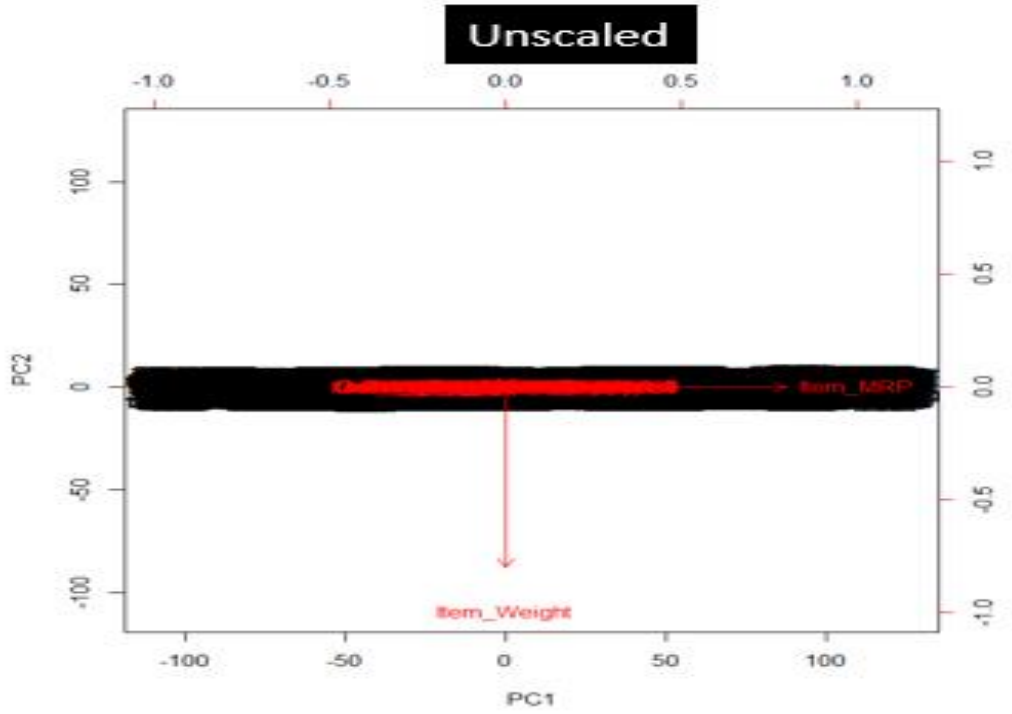
1. **Standardize the data.**

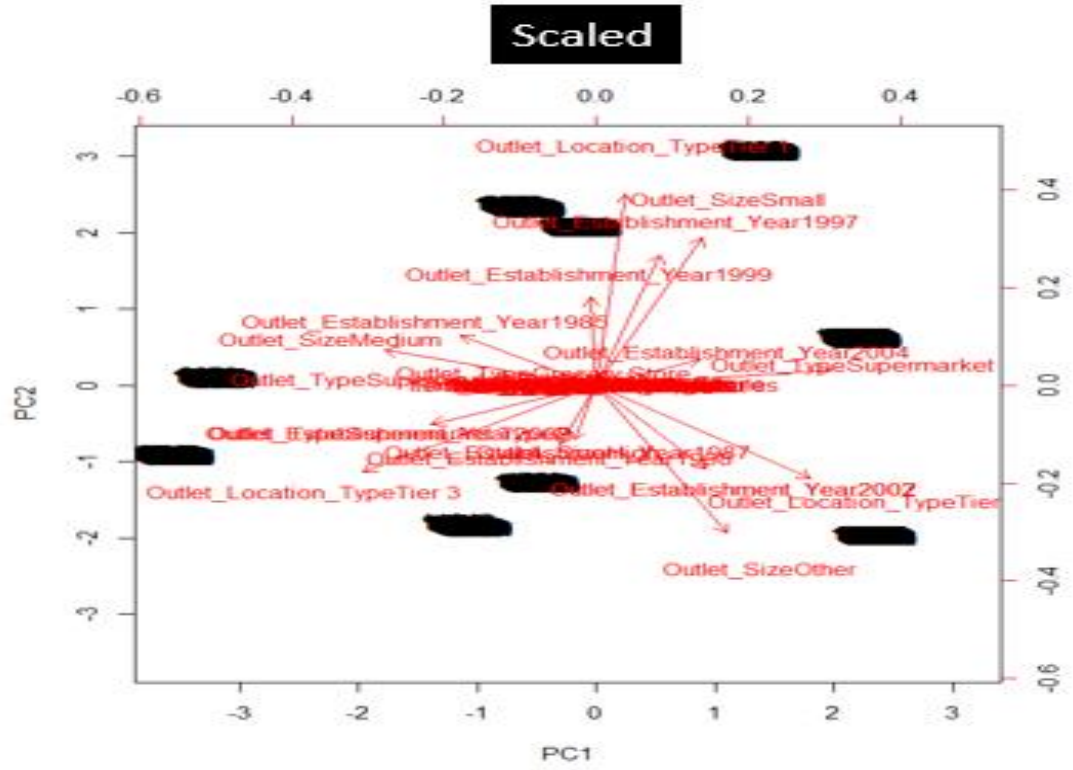
   - Perform PCA after scaling, i.e. z-scoring for each variable. Original predictors may have different scales.
   - Standardization" (or "scaling") within variables will express each observation relative to its position in the distribution for that variable.
   - Performing PCA on un-scaled variables will lead to large loadings for variables with high variance.
   - PCA was run on a data set twice (with unscaled and scaled predictors).

Steps for PCA

- This data set has ~40 variables. You can see, first principal component is dominated by one variable And, second principal component is dominated by another variable

- 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.





Steps for PCA

2. Calculate the covariance matrix.

- A correlation matrix is like a covariance matrix but first the columns have been standardized. This means the matrix should be numeric and have standardized data.

3. Find the eigenvectors of the covariance matrix.

- To convert the data into the new axes, we will multiply the original variable data by eigenvectors, which indicate the direction of the new axes (principal components).
- Each eigenvector will correspond to an eigenvalue, whose magnitude indicates how much of the data's variability is explained by its eigenvector.
- *Covariance Matrix * Eigenvector = eigenvalue * eigenvector*

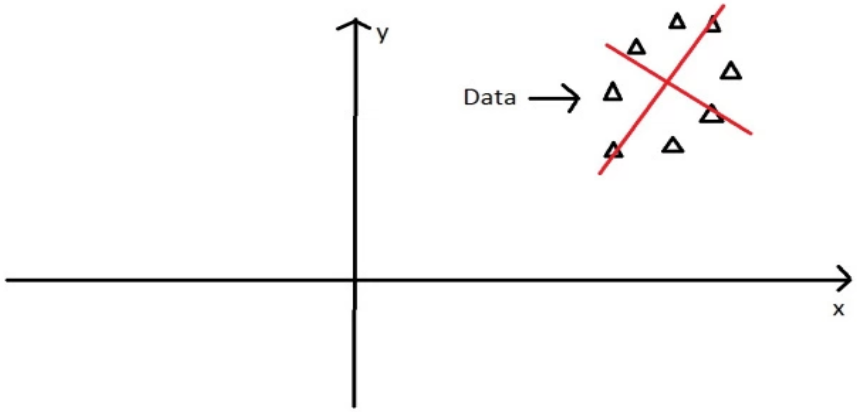
Steps for PCA

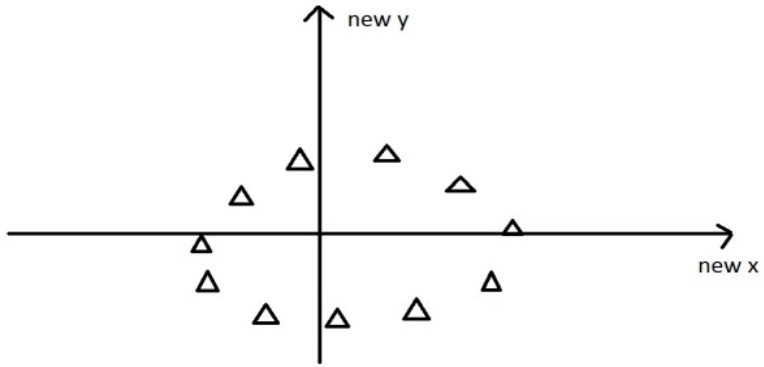
4**. Translate the data to be in terms of the components.**

- Since the eigenvectors indicates the direction of the principal components (new axes), we will multiply the original data by the eigenvectors to re-orient our data onto the new axes. This re-oriented data is called a score.

  *original data * eigenvectors*

Note : 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.





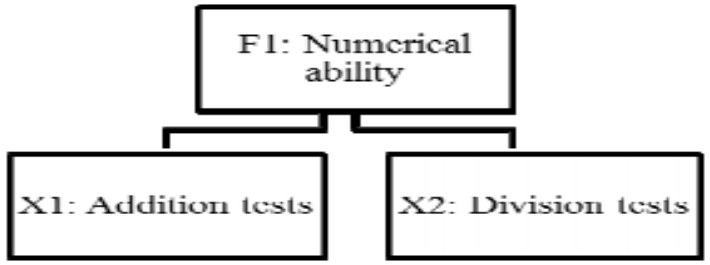
Summary

- For a fair comparison, data in PCA need to be "dimensionally homogeneous", i.e. measured in the same units.
- It is sometimes useful to transform the original variables prior to the Principal Component Analysis to "linearize" these relationships.
- Use standard transformations (logarithm, power, Box-Cox) to linearize
- One can certainly imagine a situation where PCA on correlations between all variables and PCA on correlations between "highly variable" will yield very different results.

Factor Analysis

Introduction

- Factor analysis is a method of dimension reduction.
- Factor analysis summarizes data into a few dimensions by condensing a large number of variables into a smaller set of underlying unobservable **(latent) variables or factors.**
- More specifically, the goal of factor analysis is to reduce “the dimensionality of the original space and to give an interpretation to the new space, spanned by a reduced number of new dimensions to explain the variance in the observed variables in terms of underlying latent factors
- The goal in Factor Analysis is to explain the covariances or correlations between the variables.
- Although they aren’t part of an experiment’s data set they can cause effects in your experimental results.



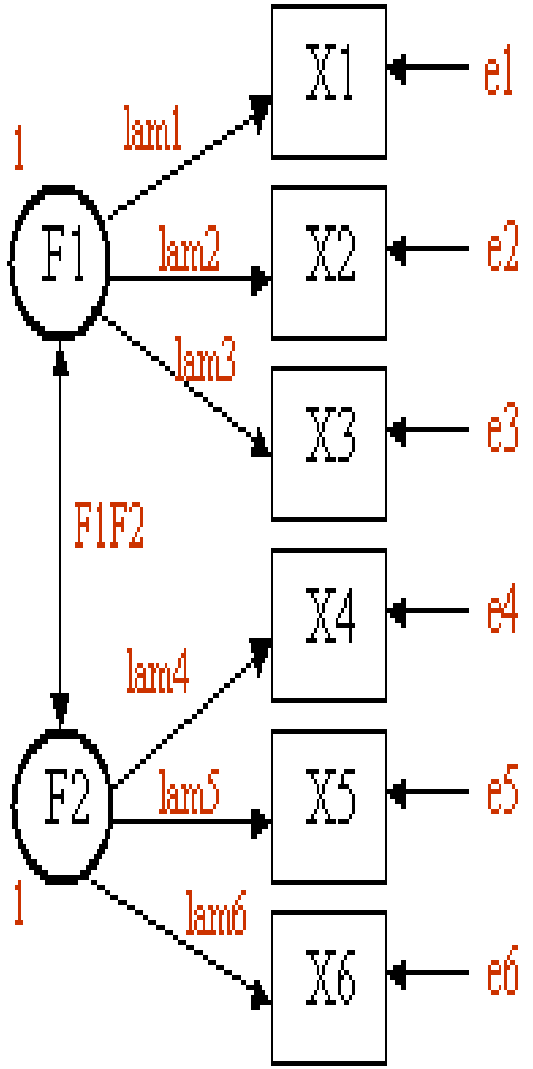
Factors

- **What are factors**
  - A “factor” is a set of observed variables that have similar response patterns because they are associated with a variable that isn’t directly measured.
  - Factors are listed according to factor loadings, or how much variation in the data they can explain.

- The starting point of factor analysis is a correlation matrix, in which the inter-correlations between the studied variables are presented.

- The dimensionality of this matrix can be reduced by “looking for variables that correlate highly with a group of other variables, but correlate very badly with variables outside of that group”

- These variables with high inter-correlations could well measure one underlying variable, **which is called a ‘factor’.**



Types of Factors

- **The are two types FA**
  - Exploratory and
  - Confirmatory.

- **Exploratory factor analysis** :
  - Is if you don’t have any idea about what structure your data is or how many dimensions are in a set of variables.

- **Confirmatory Factor Analysis** :
  - Is used for verification as long as you have a specific idea about what structure your data is or how many dimensions are in a set of variables.

When to use?

- **Use EFA** - If you want to explore patterns
- **Use CFA** - If you want to perform hypothesis testing
- EFA is almost identical to Confirmatory Factor Analysis(CFA). Both techniques can be used to confirm or explore
- PCA is a more basic version of exploratory factor analysis (EFA)

Terms

- Few topics to know while learning about FA ,These will help while interpretation of the FA
  - **Communality**
  - **Uniqueness**
  - **Loadings**
  - **Rotation of Factors**

Loadings

- Loadings are the weights and correlations between each variable and the factor.
- The higher the load the more relevant in defining the factor's dimensionality.
- A negative value indicates an inverse impact on the factor.
- The factor patterns define decreasing amounts of variation in the data.
- Each pattern may involve all or almost all the variables, and the variables may therefore have **moderate or high loadings** for several factor patterns.
- We will scale the loadings by dividing them by the corresponding communality

Communality

- Variance that is shared with other variables
- The communalities for the variable are computed by taking the sum of the squared loadings for that variable.
- For example, to compute the communality for **mpg**
  $0.643^2 + (-0.478^2) + (-0.473^2) = 0.873$

```
Uniquenesses:
  mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
0.135 0.055 0.090 0.127 0.290 0.060 0.051 0.223 0.208 0.125 0.158

Loadings:
      Factor1 Factor2 Factor3
mpg    0.643  -0.478  -0.473
cyl   -0.618   0.703   0.261
disp  -0.719   0.537   0.323
hp    -0.291   0.725   0.513
drat   0.804  -0.241
wt    -0.778   0.248   0.524
qsec  -0.177  -0.946  -0.151
vs     0.295  -0.805  -0.204
am     0.880
gear   0.908          0.224
carb   0.114   0.559   0.719
```

- The communality for a given variable can be interpreted as the proportion of variation in that variable explained by the three factors.
- If we perform multiple regression of mpg against the three common factors, we obtain an R2 = 0.873, indicating that about 87% of the variation in climate is explained by the factor model

Uniqueness

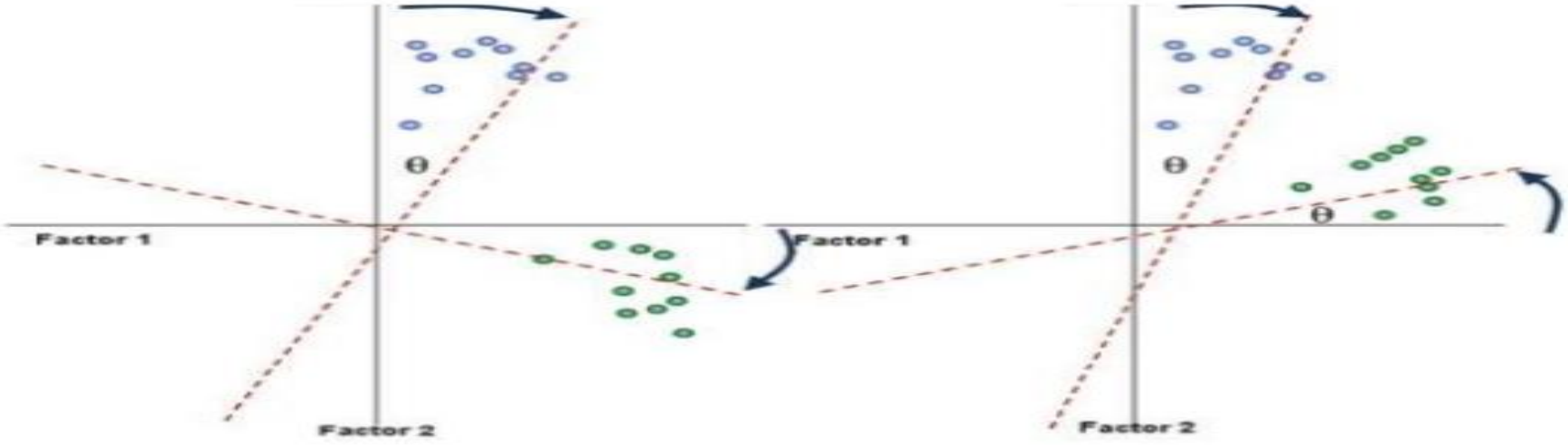
- Uniqueness is the variance that is ‘unique’ to the variable and not shared with other variables.

Uniqueness = 1 – communality

- Note :The greater ‘uniqueness’ the lower the relevance of the variable in the factor model.

Rotation of Factors

- Rotations are done for the sake **of interpretation** of the extracted factors in factor analysis
- Un-Rotated factors are not very much interpretable
- Rotations that allow for **correlation** are called **oblique rotations**
- These are less frequently used because their results are more difficult to interpret (**Promax** is example)



Rotation of Factors

- Rotations that assume the factors are **not correlated** are called **orthogonal rotations.**
- Orthogonal rotations - More suitable if our aim is data reduction & Axes maintain 90 degree
- Types of **orthogonal rotations**
  - **varimax** - which simplifies the factors (that minimizes the number of variables that have high loadings on each factor)
  - **quartimax** – which simplifies the variables (rotates the factors in order to minimize the number of factors needed to explain each variable)
  - **Equimax** -- is a combination of the Varimax method, which simplifies the factors, and the Quartimax method, which simplifies the variables
- By default the rotation is varimax which produces orthogonal factors.
- Varimax rotation is the most common of the rotations that are available

Requirements for FA

- To perform a factor analysis, there has to be univariate and multivariate normality within the data
  It is also important that there is an absence of univariate and multivariate outliers
- Factor is based on the assumption that there is a linear relationship between the factors and the variables when computing the correlations
- For something to be labeled as a factor it should have at least 3 variables
- A factor with 2 variables is only considered reliable when the variables are highly correlated with each another ($r > .70$) but fairly uncorrelated with other variables.
- The recommended sample size is at least 300 participants, and the variables that are subjected to factor analysis each should have at least 5 to 10 observations
- A factor loading for a variable is a measure of how much the variable contributes to the factor thus, high factor loading scores indicate that the dimensions of the factors are better accounted for by the variables.

Steps for FA

- Correlation matrix for all variables
- Determine the number of factors(Based on Scree Plots)
- Rotate the factors
- Interpret the factor

Approach

- The **traditional approach** to naming factors is as follows
  - Examine the variables that load heavily on the factor
  - Try do decide what construct is common to these variables
  - Name the factor after that construct

- **Business use case is shown in the next slide using the traditional approach**

Example

- The use case here explains the how factor analysis can be used to identify latent factors for the cars dataset having below observed variables

- we have reduced the variable set down to three variable categories or factors (factor1, factor2,factor3) which can be used as input variables to a model

- Using the factor loading scores within each factor, we then identify those groups of variables with the largest factor loading score from each factor

- Name the factor after that construct
  - Reliability
  - Style
  - Economy

Rotated Component Matrix

	1	2	3
Price			.823
Mileage (fuel efficiency)			.719
Low Maintenance	.553		.433
Resale value	.459		.316
Seating capacity	.588		
After-sales support	.425	.309	
Brand		.723	
Eco-friendliness	.773		
Looks		.789	
Engine power & performance	.423	.494	
Safety features	.687		

Variable loading to Factors

Factor 1	Factor 2	Factor 3
Safety features	Brand	Mileage (fuel efficiency)
Seating capacity	Engine power & performance	Price
Low Maintenance	Looks	
Resale value		
After-sales support		
Eco Friendliness		

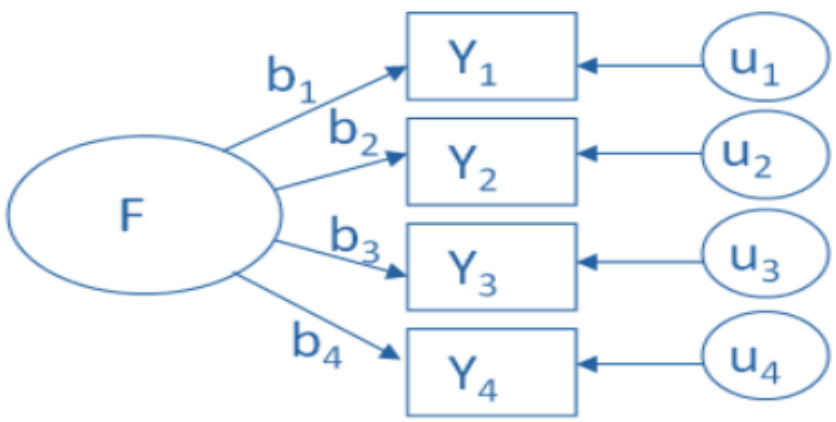
Pic credits : http://marketing-yogi.blogspot.in

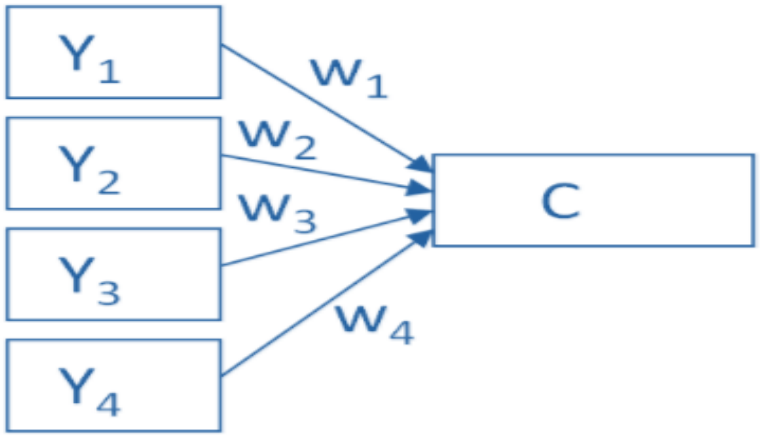
Other Example - FA in marketing is important because it reflects the perception of the buyer of the product

Conclusion

- PCA & FA Both are data reduction techniques they allow capture the variance in variables in a smaller set.
- If Your main aim is only to reduce observed data - use PCA.
- If Your aim is to reason about latent factors - use factor analysis.
- FA analyzes only the variance shared among the variables (common variance without error or unique variance) PCA analyzes all of the variance
- In Principal Components Analysis, the components are calculated as linear combinations of the original variables.
- In Factor Analysis, the original variables are defined as linear combinations of the factors.

Conclusion





FA : F is the latent Factor, is causing the responses on the four measured Y variables.

PCA : The direction of the arrows that the Y variables contribute to the component variable C

So the arrows go in the opposite direction from PCA

Thank You.