Principal Component Analysis (PCA)

- Designed to reduce the dimensionality of our data in order to capture as much of the variance as possible with the least amount of data.
- Works to optimize data storage and analysis speed and identify the key elements contributing to our output.
- Creates new variables "components" that are essentially linear combinations of the original variables

PCA Overview

- Given the essential assumption that a few key components capture the variability in the data as well as the relationship to the response variable, PCA can work to limit overfitting of the data.
- Overfitting – When a model is complex enough to start picking up on random errors or "noise" in the data set aside from actually measuring the relationship between the variables.

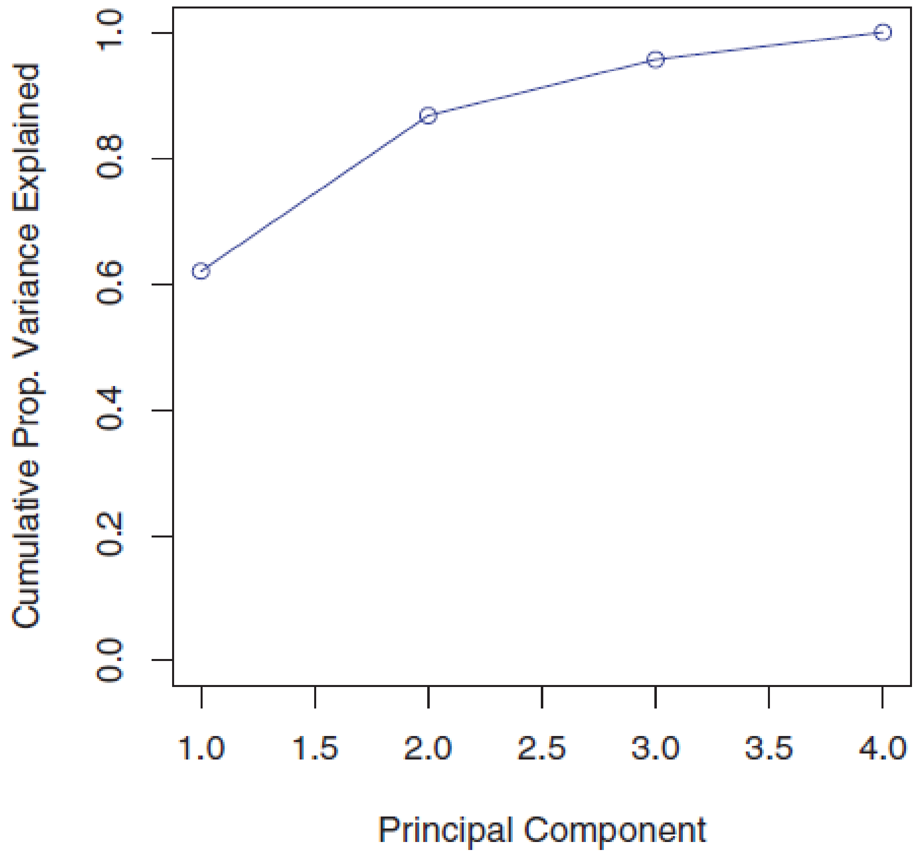
PCA Overview

- We should first scale our data to z scores to standardize the individual contributions of each component and to limit the effect of potentially high variance predictor variables. However if the units are the same, you may not want to standardize the variables.
- Proportion of Variance Explained (PVE) – We can calculate a cumulative summary of the amount of variance our PCA is capturing and plot the output, look for the elbow

PCA Overview

- We wish to explain/summarize the underlying variance-covariance structure of a large set of variables through a few linear combinations of these variables.

PCA Overview



PCA Overview

- PCA works by using eigenvalues and eigenvectors of the a matrix = A
- Eigenvalue of A is a number (scalar) r when subtractor from the diagonal entries of A converts A into a **singular matrix**
- Singular Matrix is not invertible or essentially its determinate is zero
- This is the same as saying when we multiply the identity matrix by r and subtract that result from A we get a singular matrix or
- r is a eigenvalue of A if A-rI is a singular matrix

Identity Matrix

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

- Identity Matrix – Just a square matrix in which all the diagonal elements are 1s and the rest are zeros

PCA Overview

- Quick example, sometimes we can just eyeball the eigenvalue, what number would we subtract from the diagonal to make the matrix singular?

$$A = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \end{pmatrix}$$

$A =$

2

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

PCA Overview

- Let's review determinate:

$$det \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$$

- 2*3 – 1*1 = 5

PCA Overview

- So we can apply this same idea to our function:
- In order to asses whether we have a singular matrix the determinate must equal zero or det(*A*-*rI*) = 0

$$det(A\text{-}rI) \text{ of } \begin{pmatrix} -1 & 3 \\ 2 & 0 \end{pmatrix}$$

- Find the polynomial:

$$det(A\text{-}rI) = \det \begin{pmatrix} -1-r & 3 \\ 2 & 0-r \end{pmatrix}$$

  - = r(r+1)-6=$r^2 + r - 6$=$(r+3)(r-2)$

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

PCA Overview

- $= r(r+1)-6=r^2 + r - 6=(r+3)(r-2)$
- The roots of this polynomial:-3 and 2 are the eigenvalues
- Next we need to find the eigenvectors that will simplify our matrix, we do this by plugging eigenvalues into: $(A-rI)v=0$
- Matrix A is singular if and only if Ax = 0 has a nonzero solution so applying this to the above means that $(A-rI)$ in order for $(A-rI)$ to singular when r is a eigenvalue of A the system of equations $(A-rI)v=0$ has to have a solution other than v=0

PCA Overview

- $= r(r+1)-6 = r^2 + r - 6 = (r+3)(r-2)$
- The roots of this polynomial:-3 and 2 are the eigenvalues
- $(A-(-3)I)v =$ $det \begin{pmatrix} 2 & 3 \\ 3 & 3 \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} = 0$
- $= \begin{pmatrix} 3 \\ -2 \end{pmatrix}$ along with others $\begin{pmatrix} -3 \\ 2 \end{pmatrix}$ $\begin{pmatrix} 1 \\ -2/3 \end{pmatrix}$
- Do the same for 2, $(A-(-3)I)v =$ $det \begin{pmatrix} -3 & 3 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} = 0$ $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$

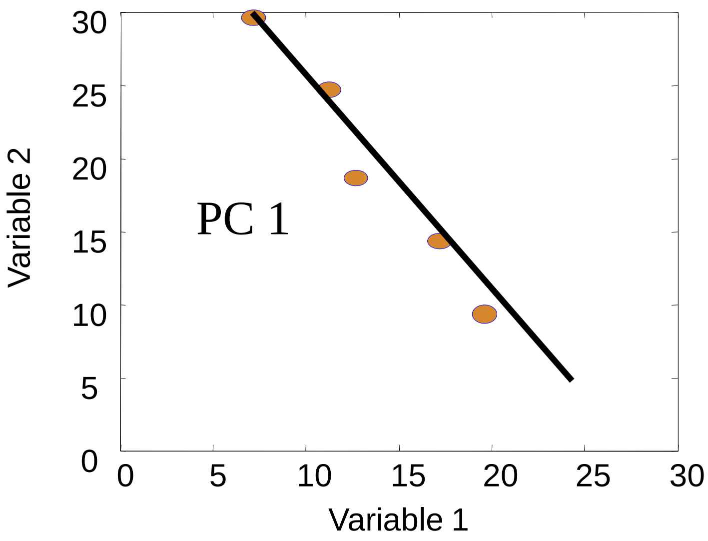
PCA Overview

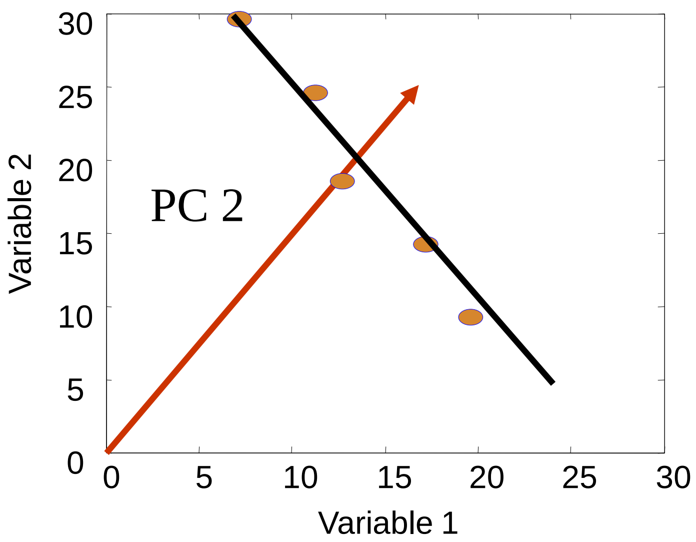
- This process is used on the covariance matrix of our components, that shows the variance and co-variance of the variables with the co-variance (diagonal) of the variables to themselves as being the variance.

- Cov(x,y) = $\frac{1}{n}\sum_{i=1}^{n}(xi - x)(yi - y)$ – Covariance calculation

PCA Overview

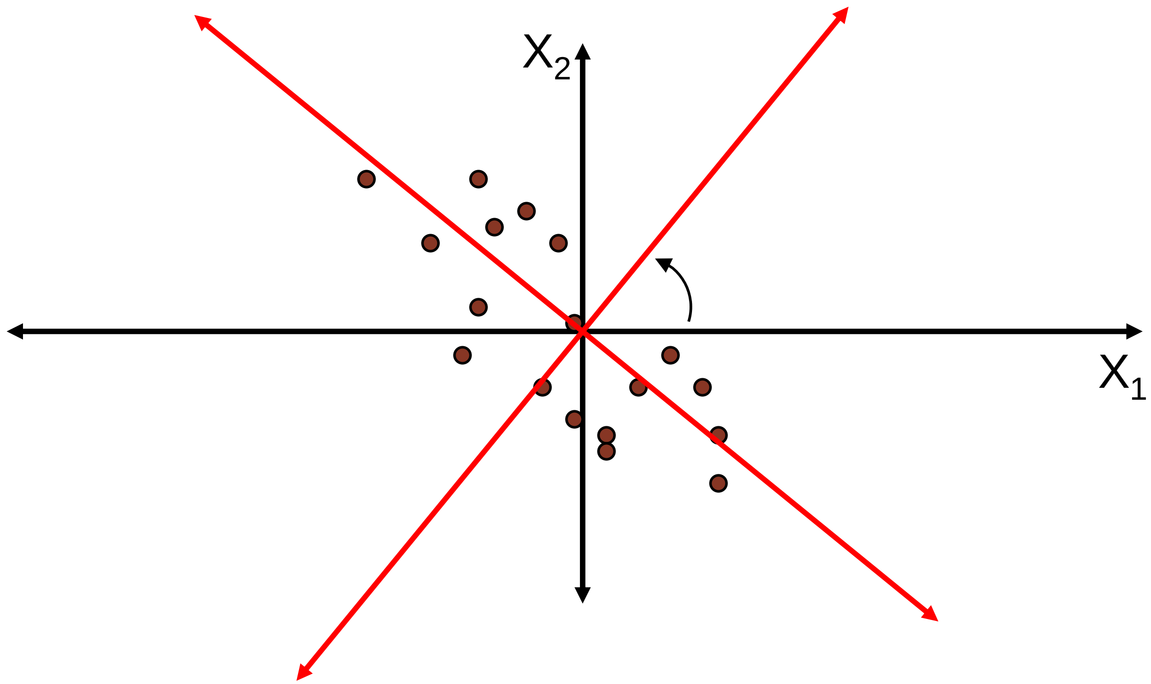
- First PC is direction of maximum variance from origin
- Subsequent PCs are orthogonal to 1st PC and describe maximum residual variance



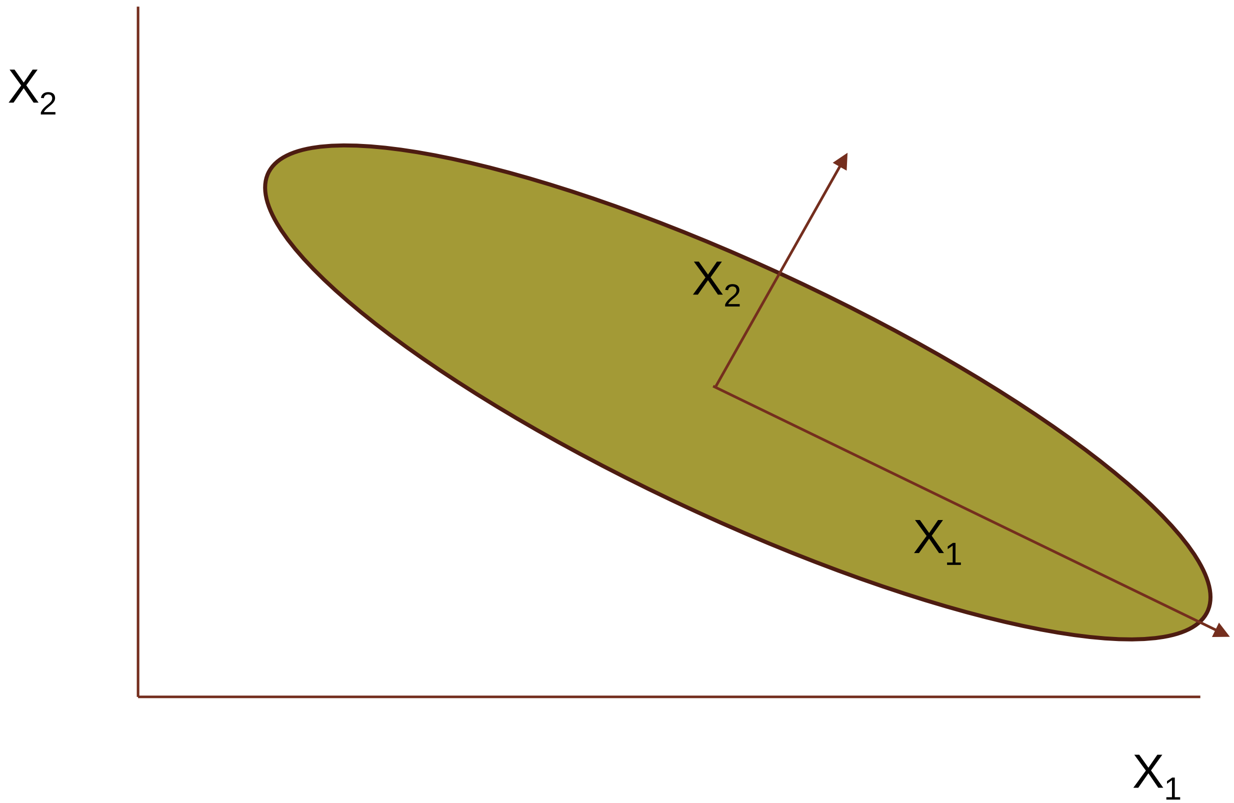


PCA Overview

- Suppose we have a population measured on p random variables X1,…,Xp. Note that these random variables represent the p-axes of the Cartesian coordinate system in which the population resides. Our goal is to develop a new set of p axes (linear combinations of the original p axes) in the directions of greatest variability:



PCA Overview



Example in R