Lecture 12

Principle Components Analysis

Reading: JWHT Chapter 10

DSA 8020 Statistical Methods II
March 29-April 2, 2021

Whitney Huang
Clemson University

Agenda

An Overview of Multivariate Analysis

- In many studies, observations are collected on several variables on each experimental/observational unit
- Multivariate analysis is a collection of statistical methods for analyzing these multivariate data sets
- **Common Objectives**
  - Dimensionality reduction
  - Classification
  - Grouping (Clustering)

You will learn more on Multivariate Analysis in DSA 8070

Multivariate Data

We display a multivariate data that contains n units on p variables using a matrix

$$\boldsymbol{X} = \begin{pmatrix} X_{1,1} & X_{2,1} & \cdots & X_{p,1} \\ X_{1,2} & X_{2,2} & \cdots & X_{p,2} \\ \vdots & \cdots & \ddots & \vdots \\ X_{1,n} & X_{2,n} & \cdots & X_{p,n} \end{pmatrix}$$

Summary Statistics

- Mean Vector: $\bar{\boldsymbol{X}} = (\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_p)^T$
- Covariance Matrix: $\Sigma = \{\sigma_{ij}\}_{i,j=1}^{p}$, where $\sigma_{ii} = \mathrm{Var}(X_i), \quad i = 1, \cdots, p$ and $\sigma_{ij} = \mathrm{Cov}(X_i, X_j), i \neq j$

Multivariate Data

We display a multivariate data that contains n units on p variables using a matrix

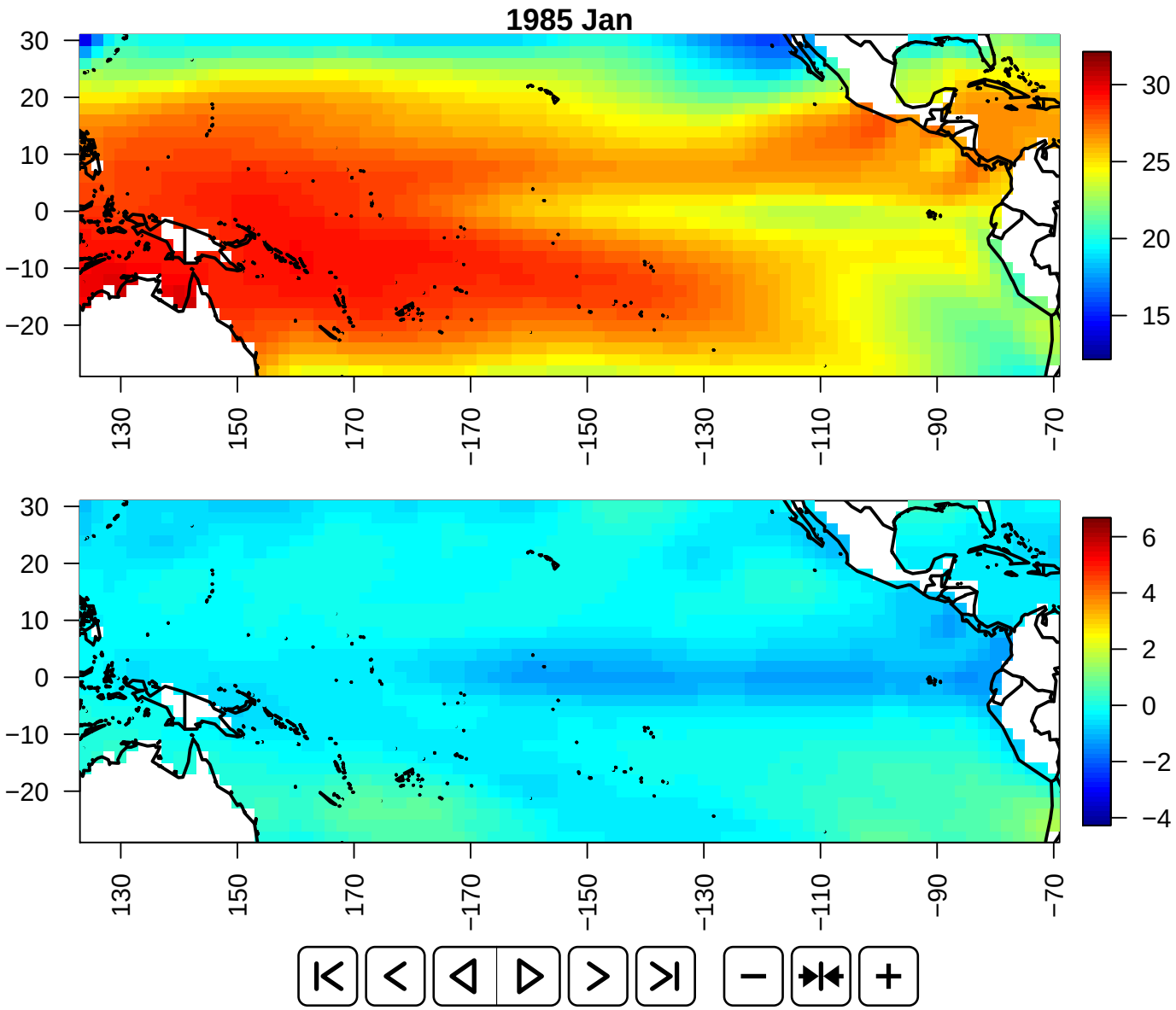
$$\boldsymbol{X} = \begin{pmatrix} X_{1,1} & X_{2,1} & \cdots & X_{p,1} \\ X_{1,2} & X_{2,2} & \cdots & X_{p,2} \\ \vdots & \cdots & \ddots & \vdots \\ X_{1,n} & X_{2,n} & \cdots & X_{p,n} \end{pmatrix}$$

Summary Statistics

- Mean Vector: $\bar{\boldsymbol{X}} = (\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_p)^T$
- Covariance Matrix: $\Sigma = \{\sigma_{ij}\}_{i,j=1}^{p}$, where $\sigma_{ii} = \mathrm{Var}(X_i), \quad i = 1, \cdots, p$ and $\sigma_{ij} = \mathrm{Cov}(X_i, X_j), i \neq j$

Next, we are going to introduce **Principal Component Analysis (PCA)**, a useful tool for conducting dimension reduction

Example: Monthly Sea Surface Temperatures



Sea Surface Temperatures and Anomalies

- The "data" are gridded at a $2°$ by $2°$ resolution from $124°E - 70°W$ and $30°S - 30°N$. The dimension of this SST data set is 2303 (number of grid points in space) $\times$ 552 (monthly time series from 1970 Jan. to 2015 Dec.)

- Sea-surface temperature anomalies are the temperature differences from the climatology (i.e. long-term monthly mean temperatures)

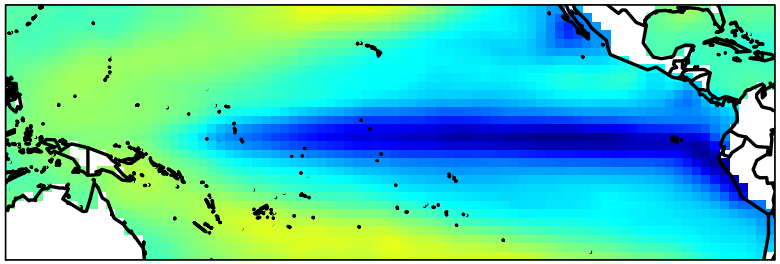
- We will demonstrate the use of Empirical Orthogonal Function (EOF) analysis to uncover the low-dimensional spatial structure of this spatio-temporal data set

The Emipirical Orthogonal Function (EOF) Decomposition

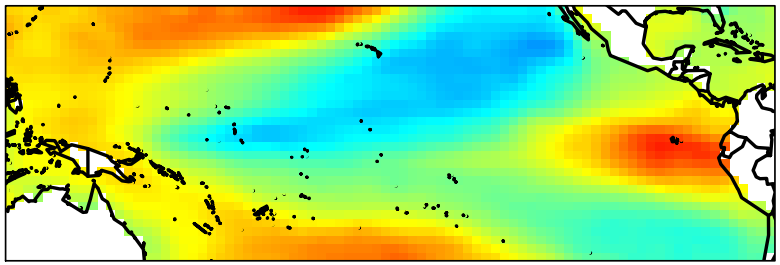
Empirical orthogonal functions (EOFs) are the geophysicist's terminology for the eigenvectors in the eigen-decomposition of an empirical covariance matrix. In its discrete formulation, EOF analysis is simply Principal Component Analysis (PCA). EOFs are usually used

- To find principal spatial structures
- To reduce the dimension (spatially or temporally) in large spatio-temporal datasets

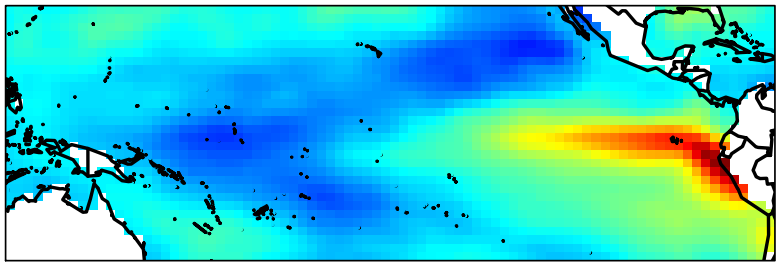
Perform EOF Decomposition and Plot the First Three Modes



EOF1: The classic ENSO pattern

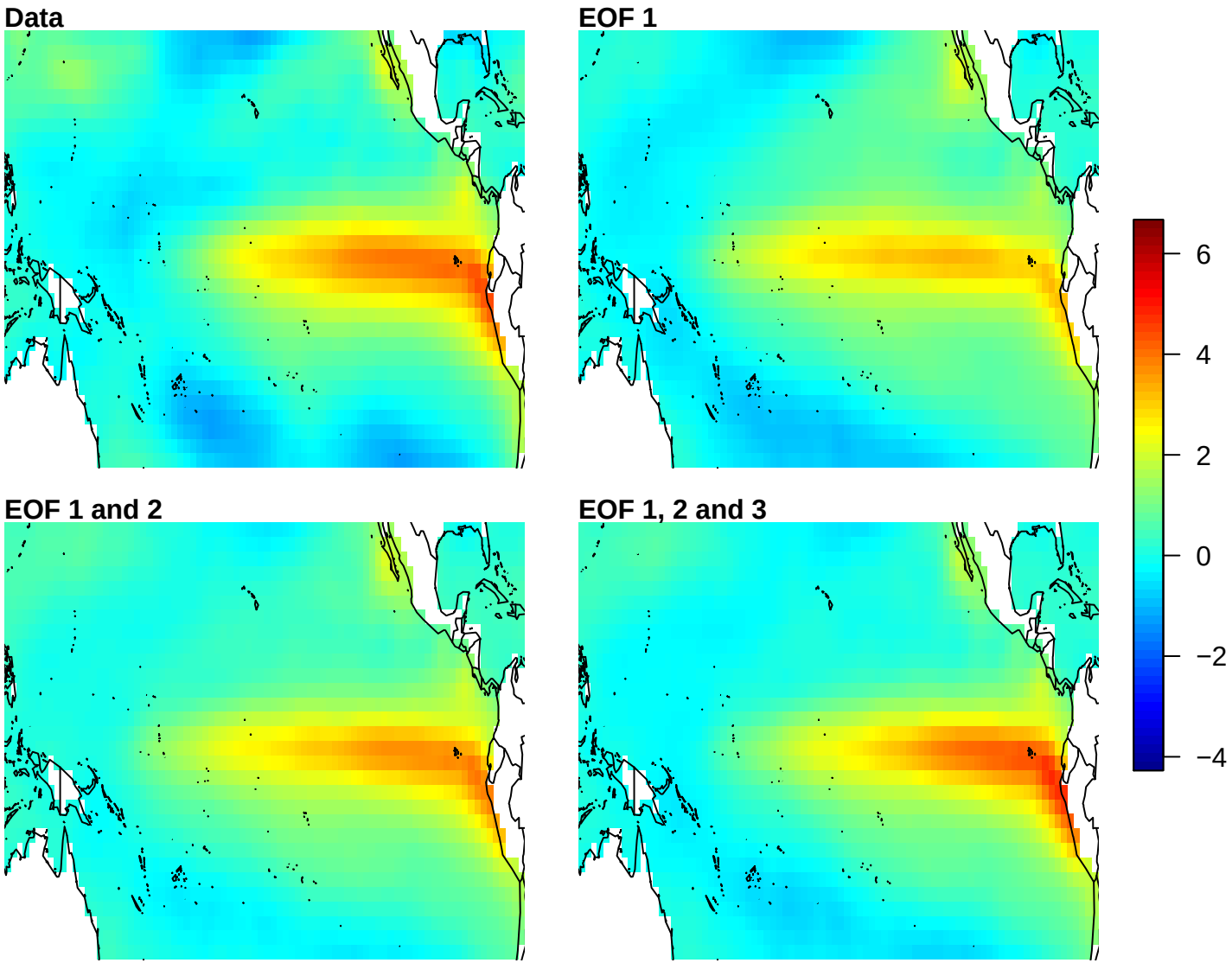


EOF2: A modulation of the center



EOF3: Messing with the coast of SA and the Northern Pacific.

1998 Jan El Niño Event



Principal Component Analysis

Given a random sample from a p-dimensional random vector
$\boldsymbol{X}_i = \{X_{1,i}, X_{2,i}, \cdots, X_{p,i}\}, \quad i = 1, \cdots, n$

- Dimension reduction technique
  - Large number of variables (p)
  - Number of variables (p) may be greater than number of observations (n)
- Create new, uncorrelated variables (principal components) for the follow up analysis
  - Principal Component Regression
  - Interpretation of principal components can be difficult in some situations

Finding Principal Components

Principal Components (PCs) are uncorrelated **linear combinations** $\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_p$ determined sequentially, as follows:

1. The first PC is the linear combination $\tilde{X}_1 = \boldsymbol{c}_1^T \boldsymbol{X} = \sum_{i=1}^p c_{1i} X_i$ that maximize $\mathrm{Var}(\tilde{X}_1)$ subject to $\boldsymbol{c}_1^T \boldsymbol{c}_1 = 1$

2. The second PC is the linear combination $\tilde{X}_2 = \boldsymbol{c}_2^T \boldsymbol{X} = \sum_{i=1}^p c_{2i} X_i$ that maximize $\mathrm{Var}(\tilde{X}_2)$ subject to $\boldsymbol{c}_2^T \boldsymbol{c}_2 = 1$ and $\boldsymbol{c}_2^T \boldsymbol{c}_1 = 0$

$\vdots$

3. The j_{th} PC is the linear combination $\tilde{X}_j = \boldsymbol{c}_j^T \boldsymbol{X} = \sum_{i=1}^p c_{ji} X_i$ that maximize $\mathrm{Var}(\tilde{X}_j)$ subject to $\boldsymbol{c}_j^T \boldsymbol{c}_j = 1$ and $\boldsymbol{c}_j^T \boldsymbol{c}_k = 0, \ \forall k < j$

Finding Principal Components by Decomposing Covariance Matrix

- Let Σ, the covariance matrix of $\boldsymbol{X}$, have eigenvalue-eigenvector pairs $(\lambda_i, \boldsymbol{e}_i)_{i=1}^p$ with with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ Then, the k_{th} principal component is given by

$$\tilde{X}_k = \boldsymbol{e}_k^T \boldsymbol{X} = e_{k1}X_1 + e_{k2}X_2 + \cdots e_{kp}X_p$$

- Then,

$$\text{Var}(\tilde{X}_i) = \lambda_i, \quad i = 1, \cdots, p$$

$$\text{Cov}(\tilde{X}_j, \tilde{X}_k) = 0, \quad \forall j \neq k$$

PCA and Proportion of Variance Explained

- It can be shown that

$$\sum_{i=1}^{p} \mathrm{Var}(\tilde{X}_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \mathrm{Var}(X_i)$$

- The proportion of the total variance associated with the k_{th} principal component is given by

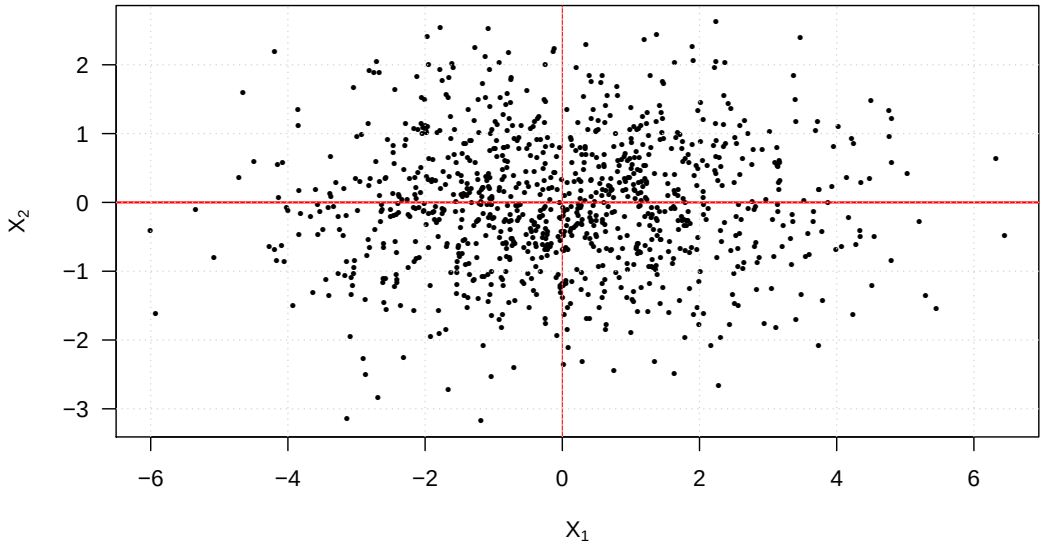
$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

- If a large proportion of the total population variance (say 80% or 90%) is explained by the first k PCs, then we can restrict attention to the first k PCs without much loss of information

Toy Example 1

Suppose we have $\boldsymbol{X} = (X_1, X_2)^T$ where $X_1 \sim \mathrm{N}(0, 4)$, $X_2 \sim \mathrm{N}(0, 1)$ are independent

- Total variation = $\mathrm{Var}(X_1) + \mathrm{Var}(X_2) = 5$
- X_1 axis explains 80% of total variation
- X_2 axis explains the remaining 20% of total variation



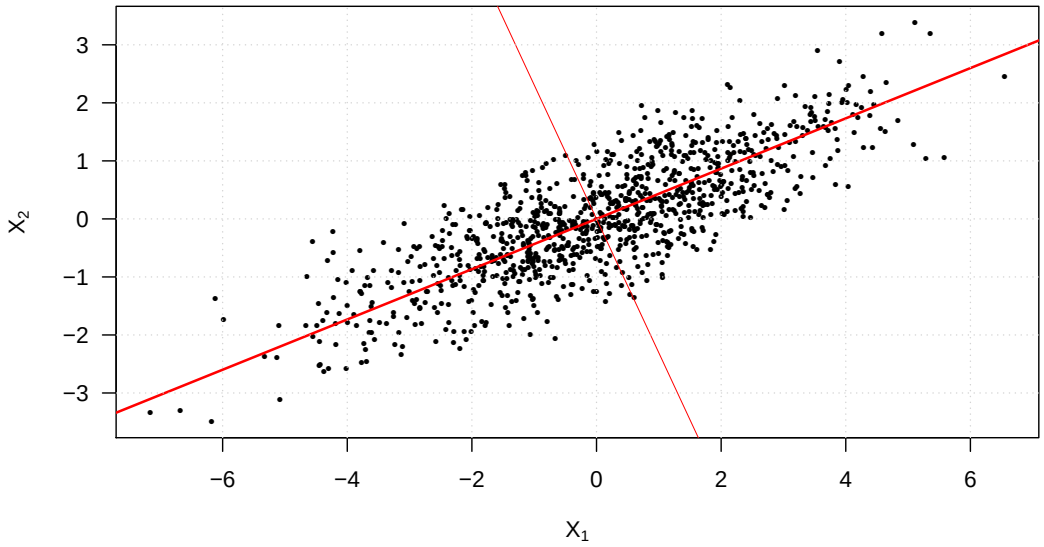
Toy Example 2

Suppose we have $\boldsymbol{X} = (X_1, X_2)^T$ where $X_1 \sim \mathrm{N}(0, 4)$, $X_2 \sim \mathrm{N}(0, 1)$ and $\mathrm{Cor}(X_1, X_2) = 0.8$

- Total variation
  $= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) = \mathrm{Var}(\tilde{X}_1) + \mathrm{Var}(\tilde{X}_2) = 5$

- $\tilde{X}_1 = .9175X_1 + .3975X_2$ explains 93.9% of total variation

- $\tilde{X}_2 = .3975X_1 - .9176X_2$ explains the remaining 6.1% of total variation



Principal Component Regression

We are going to use Longley's data set, which provides a well-known example of multicollinearity, to illustrate Principal Component Regression

Correlation matrix

```
round(cor(longley[, -7]), 3)
```

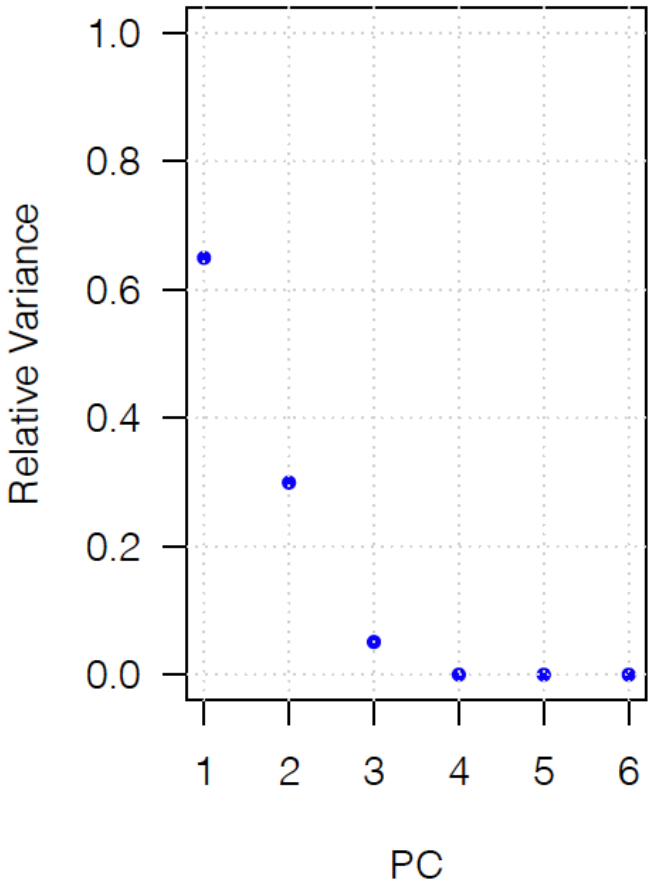
```
##              GNP.deflator   GNP Unemployed Armed.Forces Population  Year
## GNP.deflator        1.000 0.992      0.621        0.465      0.979 0.991
## GNP                 0.992 1.000      0.604        0.446      0.991 0.995
## Unemployed          0.621 0.604      1.000       -0.177      0.687 0.668
## Armed.Forces        0.465 0.446     -0.177        1.000      0.364 0.417
## Population          0.979 0.991      0.687        0.364      1.000 0.994
## Year                0.991 0.995      0.668        0.417      0.994 1.000
```

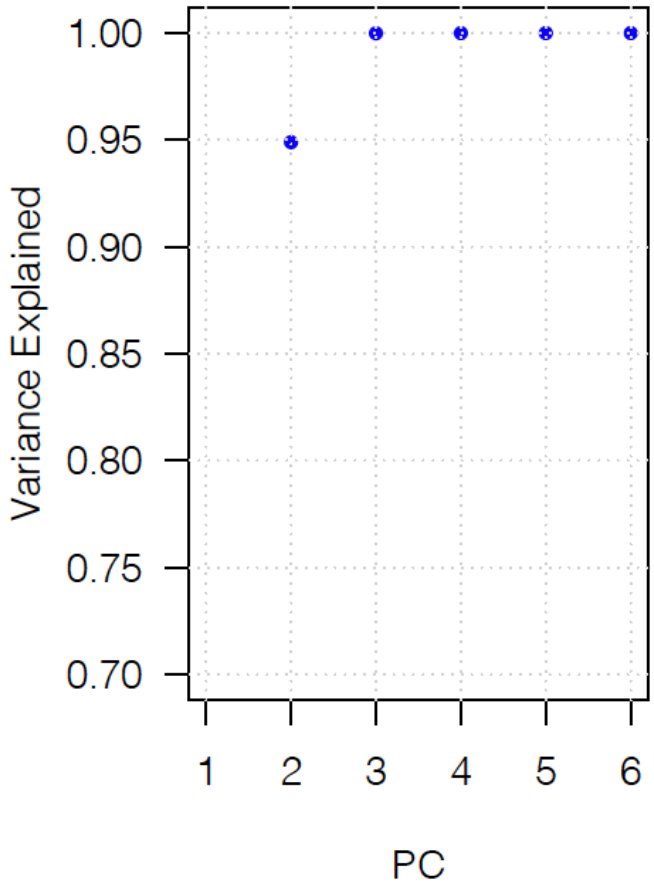
Variance inflation factor

```
vif(longley[, -7])
```

```
## GNP.deflator          GNP   Unemployed Armed.Forces   Population         Year
##    135.53244   1788.51348     33.61889      3.58893    399.15102    758.98060
```

How Many PCs to Use?





Principal Component Regression (PCR)

In PCR, instead of regressing the y on $\boldsymbol{x} = (x_1, \cdots, x_p)^T$ directly, $\tilde{x}$ are used as regressors.

```
library(pls)
pcrFit <- pcr(Employed ~ ., data = longley, valdiation = "cv")
summary(pcrFit)
```

```
## Data:    X dimension: 16 6
##  Y dimension: 16 1
## Fit method: svdpc
## Number of components considered: 6
## TRAINING: % variance explained
##           1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## X           64.96    94.90    99.99   100.00   100.00   100.00
## Employed    78.42    89.73    98.51    98.56    98.83    99.55
```