

# Lecture 37

## Principal component analysis (PCA)

*STAT 8020 Statistical Methods II*

November 25, 2019

Sea Surface  
Temperatures Example

Principal component  
analysis (PCA)

Principal Component  
Regression

Whitney Huang  
Clemson University

Sea Surface  
Temperatures Example

Principal component  
analysis (PCA)

Principal Component  
Regression

## 1 Sea Surface Temperatures Example

## 2 Principal component analysis (PCA)

## 3 Principal Component Regression

# Example: Monthly Sea Surface Temperatures

Principal component  
analysis (PCA)



Sea Surface  
Temperatures Example

Principal component  
analysis (PCA)

Principal Component  
Regression

# Sea Surface Temperatures and Anomalies

Principal component analysis (PCA)



Sea Surface Temperatures Example

Principal component analysis (PCA)

Principal Component Regression

- The “data” are gridded at a  $2^{\circ}$  by  $2^{\circ}$  resolution from  $124^{\circ}E - 70^{\circ}W$  and  $30^{\circ}S - 30^{\circ}N$ . 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 structure of this spatio-temporal data set

# The Empirical Orthogonal Function (EOF) Decomposition

Principal component analysis (PCA)



Sea Surface Temperatures Example

Principal component analysis (PCA)

Principal Component Regression

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

# Screen Plot for EOFs

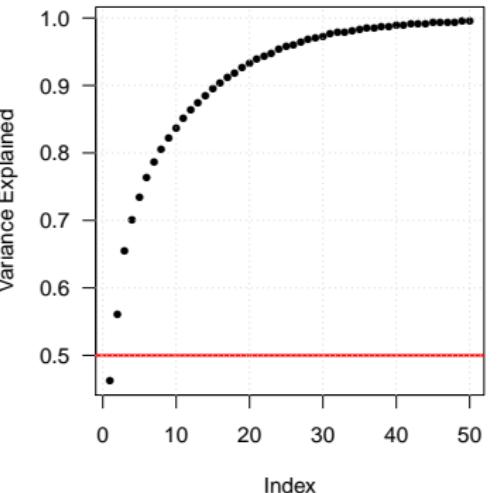
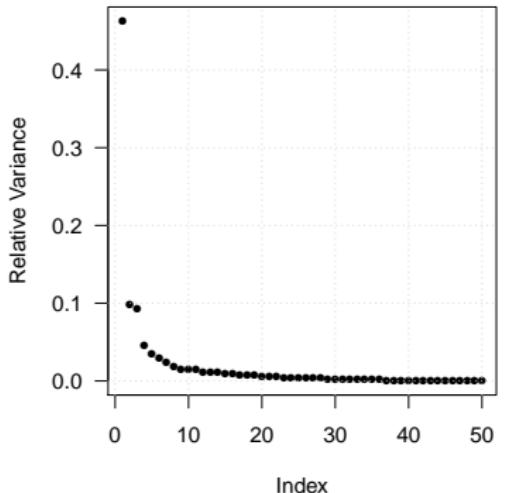
Principal component analysis (PCA)



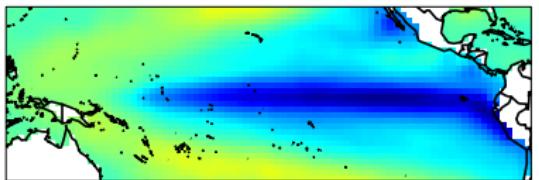
Sea Surface Temperatures Example

Principal component analysis (PCA)

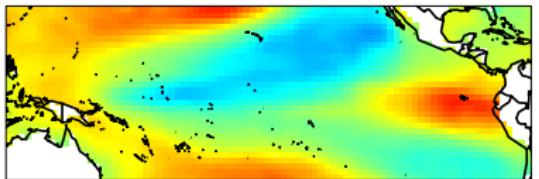
Principal Component Regression



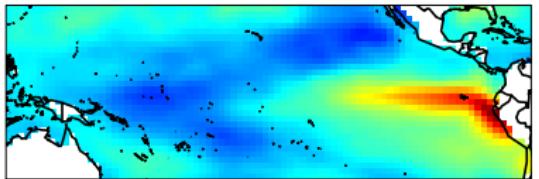
# Perform EOF Decomposition and Plot the First Three Modes



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

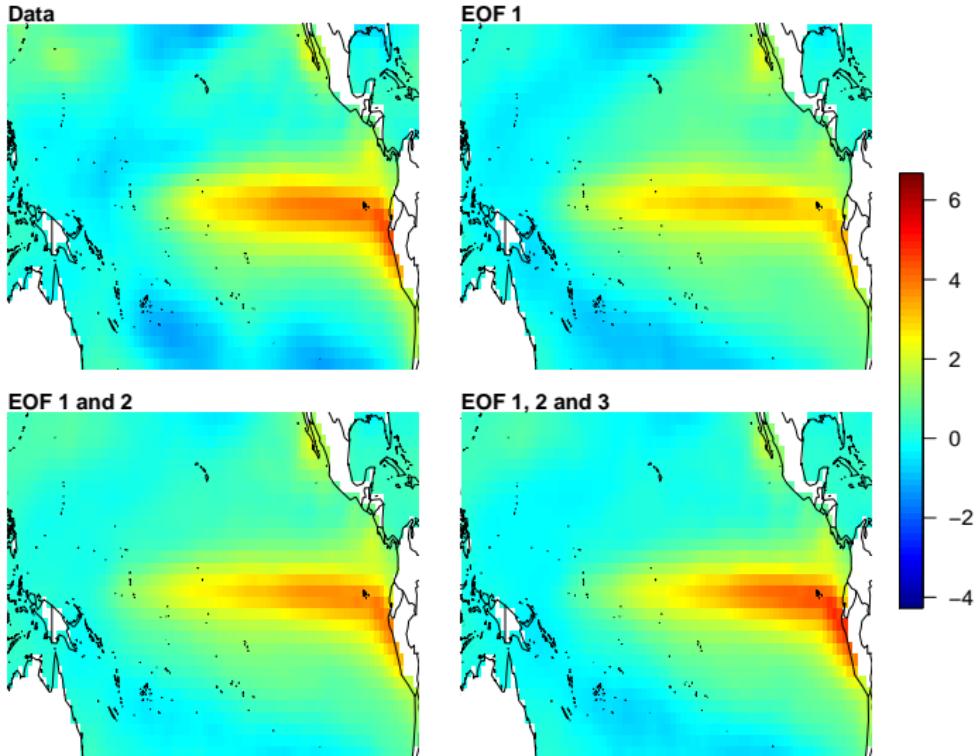
Principal component analysis (PCA)



Sea Surface Temperatures Example

Principal component analysis (PCA)

Principal Component Regression



# Principal Component Analysis

Principal component analysis (PCA)

Given a random sample from a  $p$ -dimensional random vector

$$\mathbf{X}_i = \{X_{1,i}, X_{2,i}, \dots, X_{p,i}\}, \quad i = 1, \dots, n$$



Sea Surface Temperatures Example

Principal component analysis (PCA)

Principal Component Regression

- 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 (PC) are uncorrelated **linear combinations**  $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p$  determined sequentially, as follows:

- ① The first PC is the linear combination

$$\tilde{X}_1 = \mathbf{c}_1^T \mathbf{X} = \sum_{i=1}^p c_{1i} X_i \text{ that maximize } \text{Var}(\tilde{X}_1) \text{ subject to } \mathbf{c}_1^T \mathbf{c}_1 = 1$$

- ② The second PC is the linear combination

$$\tilde{X}_2 = \mathbf{c}_2^T \mathbf{X} = \sum_{i=1}^p c_{2i} X_i \text{ that maximize } \text{Var}(\tilde{X}_2) \text{ subject to } \mathbf{c}_2^T \mathbf{c}_2 = 1 \text{ and } \mathbf{c}_2^T \mathbf{c}_1 = 0$$

⋮

- ③ The  $j_{th}$  PC is the linear combination

$$\tilde{X}_j = \mathbf{c}_j^T \mathbf{X} = \sum_{i=1}^p c_{ji} X_i \text{ that maximize } \text{Var}(\tilde{X}_j) \text{ subject to } \mathbf{c}_j^T \mathbf{c}_j = 1 \text{ and } \mathbf{c}_j^T \mathbf{c}_k = 0 \forall k < j$$

# Principal Components

Principal component analysis (PCA)



Sea Surface Temperatures Example

Principal component analysis (PCA)

Principal Component Regression

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

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

- Then,

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

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

## PCA and Proportion of Variance Explained

Principal component analysis (PCA)



- It can be shown that

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

- 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

Sea Surface Temperatures Example

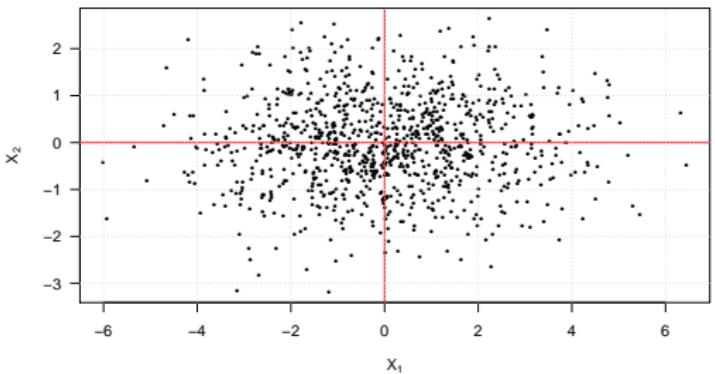
Principal component analysis (PCA)

Principal Component Regression

## Toy Example 1

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

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



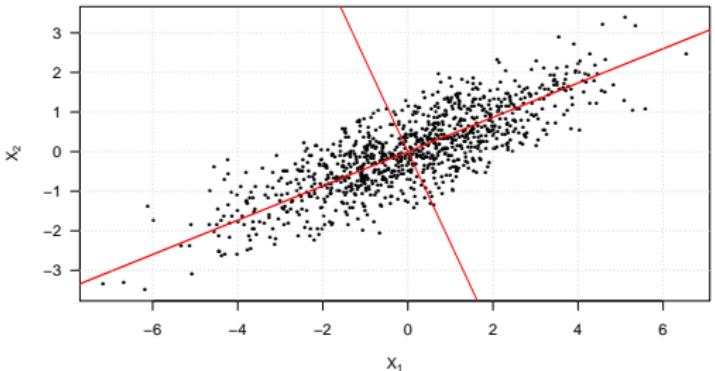
## Toy Example 2

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

- Total variation

$$= \text{Var}(X_1) + \text{Var}(X_2) = \text{Var}(\tilde{X}_1) + \text{Var}(\tilde{X}_2) = 5$$

- $\tilde{X}_1 = .9175X_1 + .3975X_2$  explains 93.9% of total variation
- $\tilde{X}_2 = .3975X_1 - .9176X_2$  explains the remaining 6.1% of total variation



# Longley's Economic Regression Data

Principal component analysis (PCA)



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

Sea Surface Temperatures Example

Principal component analysis (PCA)

Principal Component Regression

	GNP	Unemployed	Armed.Forces	Population	Year	Employed
GNP	1.00	0.60	0.45	0.99	1.00	0.98
Unemployed	0.60	1.00	-0.18	0.69	0.67	0.50
Armed.Forces	0.45	-0.18	1.00	0.36	0.42	0.46
Population	0.99	0.69	0.36	1.00	0.99	0.96
Year	1.00	0.67	0.42	0.99	1.00	0.97
Employed	0.98	0.50	0.46	0.96	0.97	1.00

GNP	Unemployed	Armed.Forces	Population	Year
14350.70398	601.69137	98.18754	558.11084	22897.44840
Employed				
1064.78369				

# How Many PCs to Use?

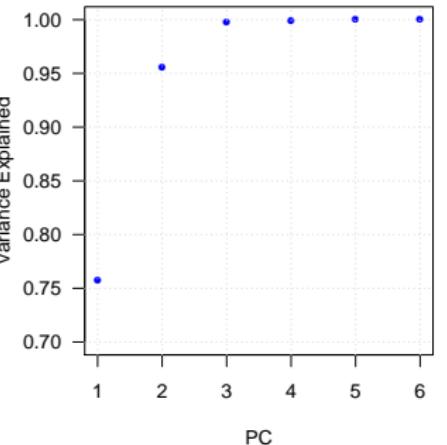
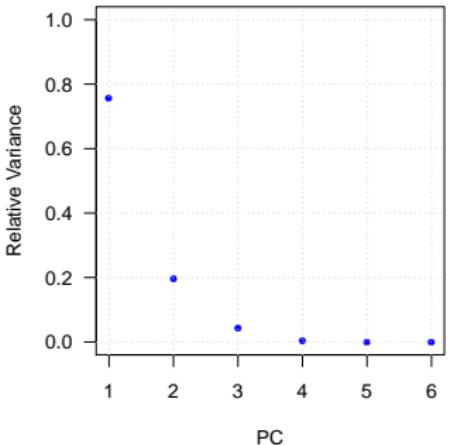
Principal component analysis (PCA)



Sea Surface Temperatures Example

Principal component analysis (PCA)

Principal Component Regression



# Principal Component Regression

Principal component analysis (PCA)



- Ordinary multiple linear regression:

$$Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Sea Surface Temperatures Example

- Principal Component Regression:

$$Y = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\varepsilon}}$$

Principal component analysis (PCA)

Principal Component Regression

An example of R output:

```
# Fit model
pcrFit<-pcr(Employed ~ ., data = longley, validation = "cv")

# Summarize the fit
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
```