Lecture 37

Principal component analysis (PCA)

STAT 8020 Statistical Methods II
November 25, 2019

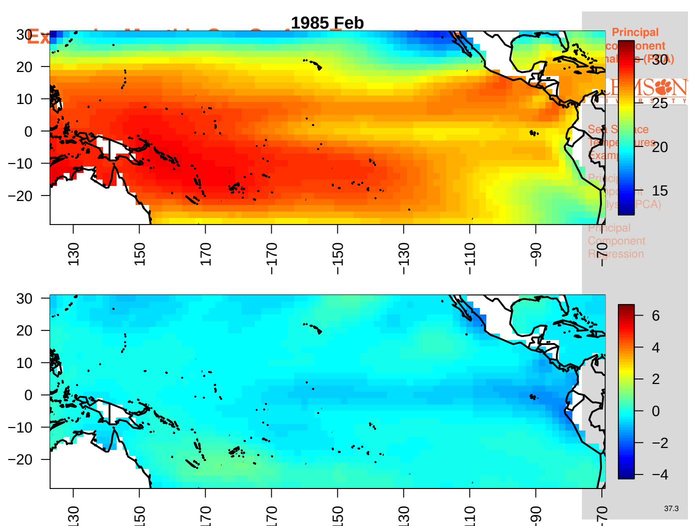
Whitney Huang
Clemson University

Notes

Agenda

1. Sea Surface Temperatures Example
2. Principal component analysis (PCA)
3. Principal Component Regression

Notes



Notes

Sea Surface Temperatures and Anomalies

- The "data" are gridded at a 2° by 2° 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) ×
  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

Notes

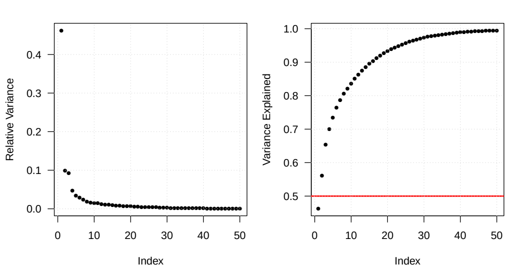
The Emipirical Orthogonal Function (EOF) Decomposition

Two parts:

- Basis functions – spatial fields in the SST example
  - Basis functions determined in an optimal way (in terms of least square) from the same data ⇒ "empirical"
- Coefficients – monthly time series in the SST example
  - Coefficients found by least squares

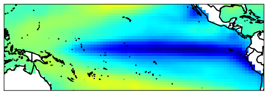
Notes

Screen Plot for EOFs

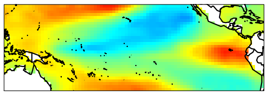


Notes

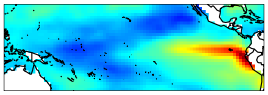
The First 3 EOFs



EOF1: The classic ENSO pattern



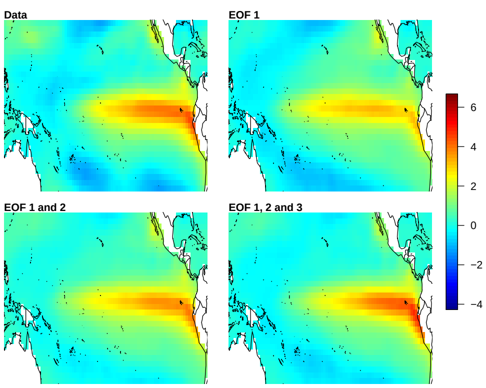
EOF2: A modulation of the center



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

Notes

1998 Jan El Niño Event



Notes

Principal Component Analysis

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

- Dimension reduction technique
  - Large number of variables (p)
  - Number of variables (p) may be greater than number of observations (n)
- Create new, uncorrelated variables for the follow up analysis
  - Principal Component Regression
  - Interpretation of new variables?

Notes

Finding Principal Components

Principal Components (PC) 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 $\tilde{X}_1 = c_1^T X = \sum_{i=1}^p c_{1i} X_i$ that maximize $\mathrm{Var}(\tilde{X}_1)$ subject to $c^T c = 1$
2. The second PC is the linear combination $\tilde{X}_2 = c_2^T X = \sum_{i=1}^p c_{2i} X_i$ that maximize $\mathrm{Var}(\tilde{X}_2)$ subject to $c_1^T c_1 = 1$ and $c_2^T c_1 = 0$

$$\vdots$$

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

Notes

Principal Components

- Let Σ have eigenvalue-eigenvector pairs $(\lambda_i, e_i)_{i=1}^p$ with with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ Then, the i_{th} principal component is given by

$$\tilde{X}_i = e_i^T X = e_{i1} X_1 + e_{i2} X_2 + \cdots e_{ip} X_p$$

- Then,

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

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

Notes

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

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

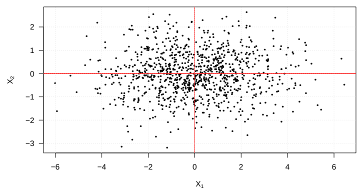
- If a large proportion of the total population variance (say 80% or 90%) is explained by the first k PCs, then we can restrict attention to the first k PCs without much loss of information

Notes

Toy Example 1

Suppose we have $X = (X_1, X_2)^T$ 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

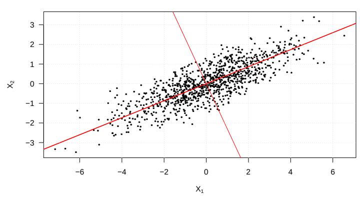


Notes

Toy Example 2

Suppose we have $X = (X_1, X_2)^T$ 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



Notes

Longley's Economic Regression Data

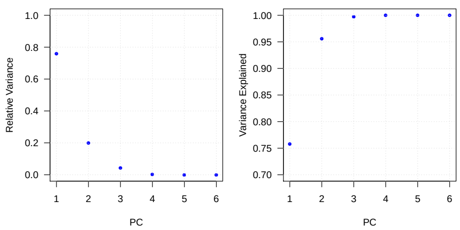
We are going to use Longley's data set, which provides a well-known example of multicollinearity, to illustrate Principal Component Regression

	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
```


Notes

How Many PCs to Use?



Notes

Principal Component Regression in R

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


Notes

Notes