

Lecture 8

Principle Component Analysis

Reading: Zelterman Chapter 8.1-8.4; Izenman Chapter 7.1-7.2

DSA 8070 Multivariate Analysis

Background

Finding Principal
Components

Principal Component
Analysis in Practice

Whitney Huang
Clemson University

Background

Finding Principal
Components

Principal Component
Analysis in Practice

1 Background

2 Finding Principal Components

3 Principal Component Analysis in Practice

History

- Karl Pearson (1901): a procedure for finding lines and planes which best fit a set of points in p -dimensional space
- Harold Hotelling (1933): to find a smaller “fundamental set of independent variables” that determines the values of the original set of p variables

LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space.* By KARL PEARSON, F.R.S., University College, London*.

(1) In many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the "best-fitting" straight line or plane. Analytically this consists in taking

$$\begin{aligned}y &= a_0 + a_1x_1 \quad \text{or} \quad z = a_0 + a_1x + b_1y, \\ \text{or} \quad z &= a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \dots + a_nx_n,\end{aligned}$$

where $y, x_1, x_2, x_3, \dots, x_n$ are variables, and determining the "best" values for the constants $a_0, a_1, b_1, a_2, a_3, a_1, a_2, a_3, \dots, a_n$ in relation to the observed corresponding values of the variables. In nearly all the cases dealt with in the text-books of least squares, the variables on the right of our equations are treated as the independent, those on the left as the dependent variables. The result of this treatment is that we get one straight line or plane if we treat some one variable as independent, and a quite different one if we treat another variable as the independent variable. There is no paradox

ANALYSIS OF A COMPLEX OF STATISTICAL VARIABLES INTO PRINCIPAL COMPONENTS¹

HAROLD HOTELLING

Columbia University

1. INTRODUCTION

Consider n variables attaching to each individual of a population. These statistical variables x_1, x_2, \dots, x_n might for example be scores made by school children in tests of speed and skill in solving arithmetical problems or in reading; or they might be various physical properties of telephone poles, or the rates of exchange among various currencies. The x 's will ordinarily be correlated. It is natural to ask whether some more fundamental set of independent variables exists, perhaps fewer in number than the x 's, which determine the values the x 's will take. If $\gamma_1, \gamma_2, \dots$ are such variables, we shall then have a set of relations of the form

$$x_i = f_i(\gamma_1, \gamma_2, \dots) \quad (i = 1, 2, \dots, n) \quad (1)$$

Quantities such as the γ 's have been called mental factors in recent psychological literature. However in view of the prospect of application of these ideas outside of psychology, and the conflicting usage attaching to the word "factor" in mathematics, it will be better simply to call the γ 's components of the complex depicted by the tests.

Background

Finding Principal Components

Principal Component Analysis in Practice

Background

Finding Principal
ComponentsPrincipal Component
Analysis in Practice

Basic Idea

Reduce the **dimensionality** of a data set in which there is a large number (i.e., p is “large”) of inter-related variables while retaining as much as possible the **variation** in the original set of variables

- The reduction is achieved by transforming the original variables to a new set of variables, “**principal components**”, that are **uncorrelated**
- These principal components are **ordered** such that **the first few retains most of the variation present in the data**
- Goals/Objectives
 - Reduction and summary
 - Study the structure of **covariance/correlation matrix**

Background

Finding Principal
Components

Principal Component
Analysis in Practice

- Interpretation (by studying the structure of covariance/correlation matrix)
- Select a sub-set of the original variables, that are uncorrelated to each other, to be used in other multivariate procedures (e.g., multiple regression, classification)
- Detect outliers or clusters of multivariate observations

Background

Finding Principal
ComponentsPrincipal Component
Analysis in Practice

Multivariate Data

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

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

Summary Statistics

- **Mean Vector:** $\bar{\mathbf{X}} = (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p)^T$, where

$$\bar{X}_j = \frac{\sum_{i=1}^n X_{j,i}}{n}, \quad j = 1, \dots, p$$

- **Covariance Matrix:** $\Sigma = \{\sigma_{ij}\}_{i,j=1}^p$, where
 $\sigma_{ii} = \text{Var}(X_i)$, $i = 1, \dots, p$ and $\sigma_{ij} = \text{Cov}(X_i, X_j)$, $i \neq j$

Background

Finding Principal
ComponentsPrincipal Component
Analysis in Practice

Multivariate Data

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

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

Summary Statistics

- **Mean Vector:** $\bar{\mathbf{X}} = (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p)^T$, where

$$\bar{X}_j = \frac{\sum_{i=1}^n X_{j,i}}{n}, \quad j = 1, \dots, p$$

- **Covariance Matrix:** $\Sigma = \{\sigma_{ij}\}_{i,j=1}^p$, where
 $\sigma_{ii} = \text{Var}(X_i)$, $i = 1, \dots, p$ and $\sigma_{ij} = \text{Cov}(X_i, X_j)$, $i \neq j$

Next, we are going to discuss how to find **principal components**

Background

Finding Principal
ComponentsPrincipal Component
Analysis in Practice

Finding Principal Components

Principal Components (PCs) 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

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

- ② The second PC is the linear combination

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

- ③ The p_{th} PC is the linear combination

$$\tilde{X}_p = \mathbf{c}_p^T \mathbf{X} = \sum_{i=1}^p c_{pi} X_i \text{ that maximizes } \text{Var}(\tilde{X}_p) \text{ subject to}$$
$$\mathbf{c}_p^T \mathbf{c}_p = 1 \text{ and } \mathbf{c}_p^T \mathbf{c}_k = 0, \forall k < p$$

Finding Principal Components by Decomposing Covariance Matrix

- Let Σ , 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

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

⇒ we can perform a single matrix operation to get the coefficients to form all the PCs!

- Then,

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

Moreover $\text{Var}(\tilde{X}_1) \geq \text{Var}(\tilde{X}_2) \geq \dots \geq \text{Var}(\tilde{X}_p) \geq 0$

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

⇒ different PCs are uncorrelated with each other

Background

Finding Principal Components

Principal Component Analysis in Practice

Background

Finding Principal Components

Principal Component Analysis in Practice

PCA and Proportion of Variance Explained

- 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 \Rightarrow we achieve dimension reduction by considering $k < p$ uncorrelated components rather than the original p correlated variables

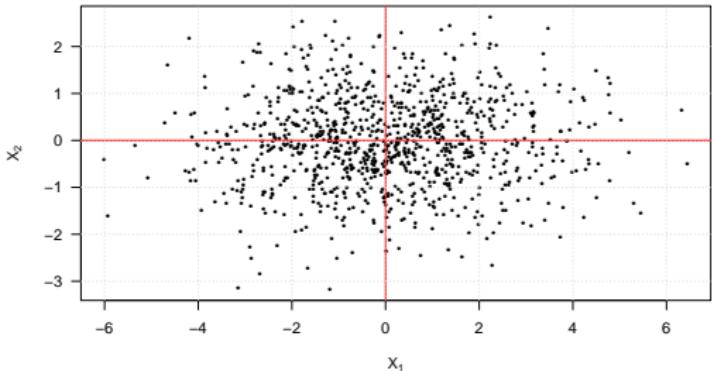
Background

Finding Principal
ComponentsPrincipal Component
Analysis in Practice

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



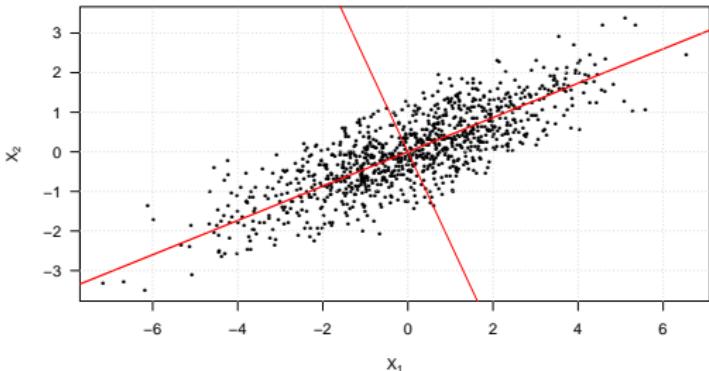
Toy Example 2

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



Background

Finding Principal Components

Principal Component Analysis in Practice

Background

Finding Principal Components

Principal Component Analysis in Practice

PCs of Standardized versus Original Variables

If we use standardized variables, i.e., $Z_j = \frac{X_j - \mu_j}{\sqrt{\sigma_{jj}}}$ $j = 1, \dots, p$ ("z-scores"). Then we are going to work with the **correlation matrix** instead of the **covariance matrix** of $(X_1, \dots, X_p)^T$

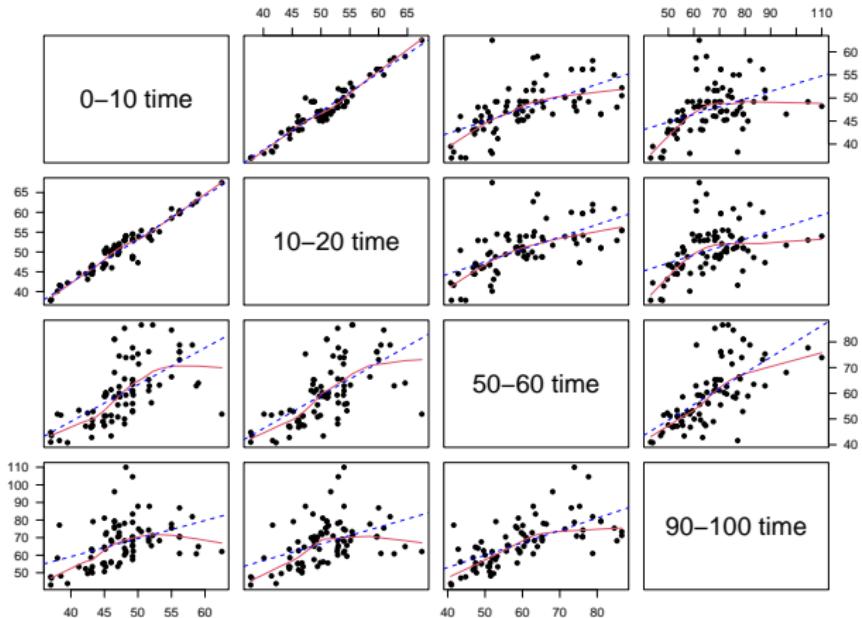
- We can obtain PCs of standardized variables by applying spectral decomposition of the correlation matrix
- However, the PCs (and the proportion of variance explained) are, in general, different than those from original variables
- If units of p variables are comparable, covariance PCA may be more informative, if units of p variables are incomparable, correlation PCA may be more appropriate

Background

Finding Principal
ComponentsPrincipal Component
Analysis in Practice

Example: Men's 100k Road Race

The data consists of the times (in minutes) to complete successive 10k segments ($p = 10$) of the race. There are 80 racers in total ($n = 80$)



Eigenvalues of $\hat{\Sigma}$

Background

Finding Principal Components

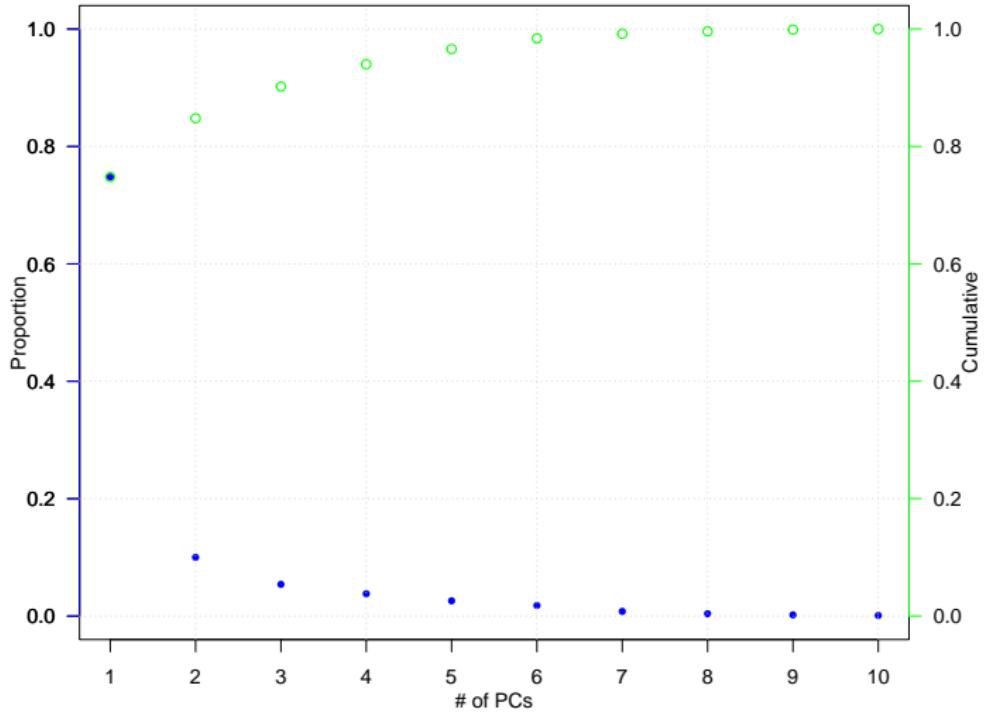
Principal Component Analysis in Practice

	Eigenvalue	Proportion	Cumulative
PC1	735.77	0.75	0.75
PC2	98.47	0.10	0.85
PC3	53.27	0.05	0.90
PC4	37.30	0.04	0.94
PC5	26.04	0.03	0.97
PC6	17.25	0.02	0.98
PC7	8.03	0.01	0.99
PC8	4.25	0.00	1.00
PC9	2.40	0.00	1.00
PC10	1.29	0.00	1.00

Much of the total variance can be explained by the first three PCs

How Many Components to Retain?

A **scree plot** displays the variance explained by each component



Background

Finding Principal Components

Principal Component Analysis in Practice

Men's 100k Road Race Component Weights

Background

Finding Principal
Components

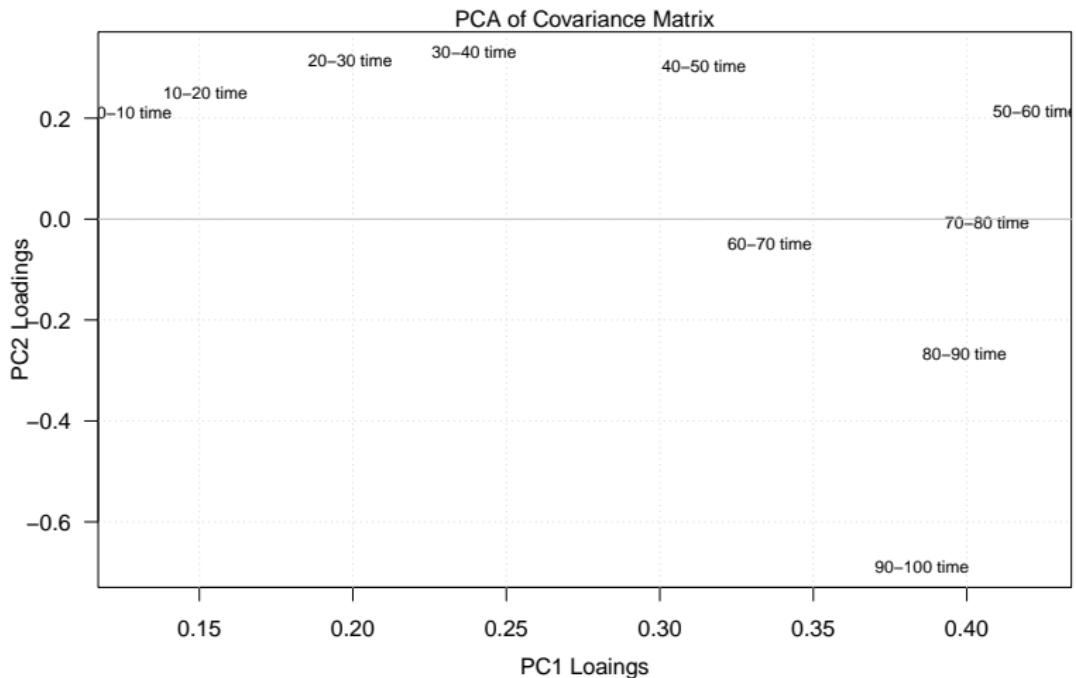
Principal Component
Analysis in Practice

	Comp.1	Comp.2	Comp.3
0-10 time	0.13	0.21	0.36
10-20 time	0.15	0.25	0.42
20-30 time	0.20	0.31	0.34
30-40 time	0.24	0.33	0.20
40-50 time	0.31	0.30	-0.13
50-60 time	0.42	0.21	-0.22
60-70 time	0.34	-0.05	-0.19
70-80 time	0.41	-0.01	-0.54
80-90 time	0.40	-0.27	0.15
90-100 time	0.39	-0.69	0.35

What these numbers mean?

[Background](#)[Finding Principal Components](#)[Principal Component Analysis in Practice](#)

Visualizing the Weights to Gain Insight



First component: overall speed

Second component: contrast long and short races

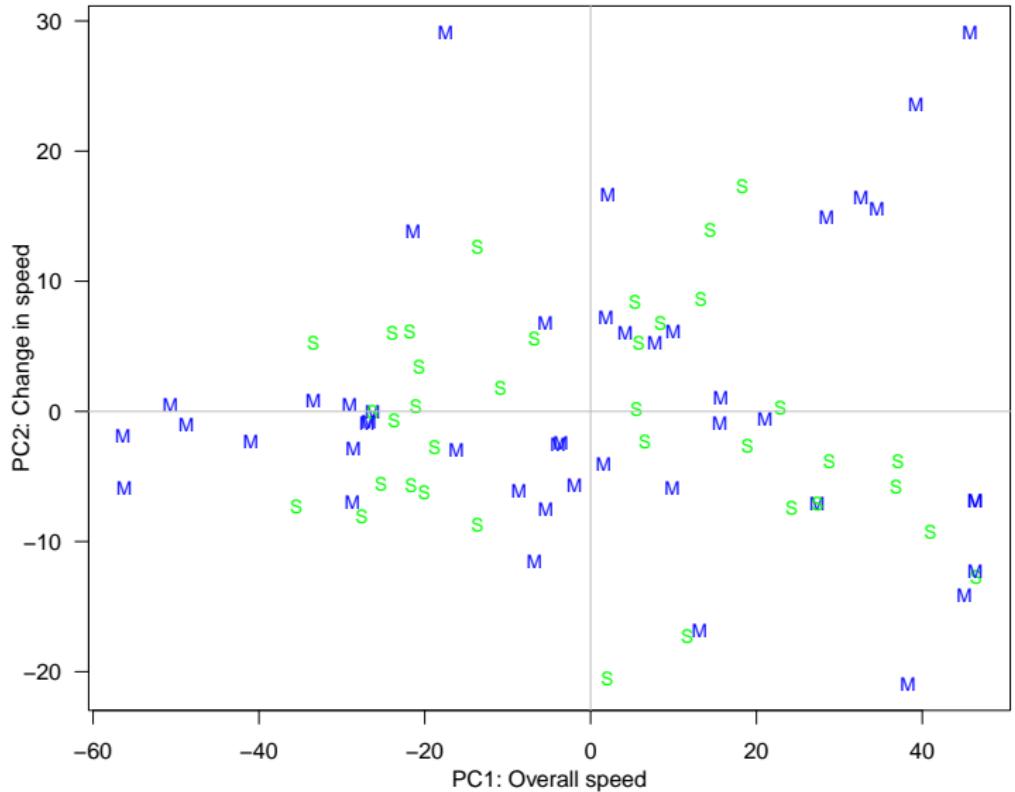
Background

Finding Principal Components

Principal Component Analysis in Practice

Looking for Patterns

Mature runners: Age < 40 (M); Senior runners: Age ≥ 40 (S)

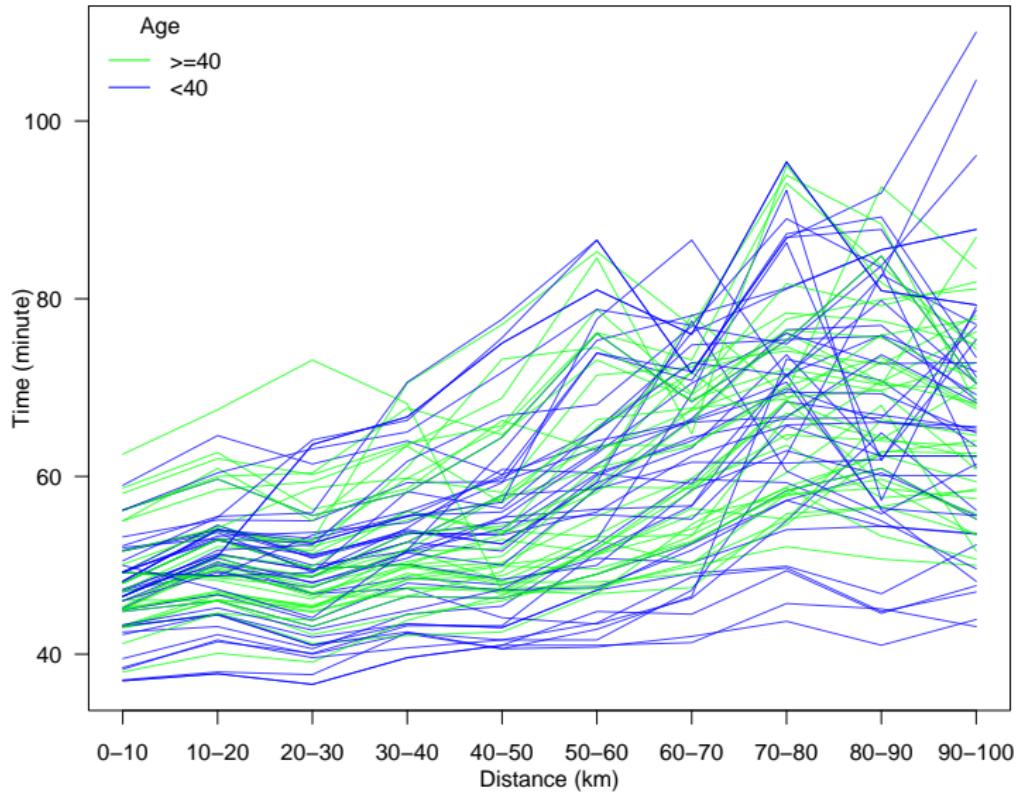


Relating to Original Data: Profile Plot

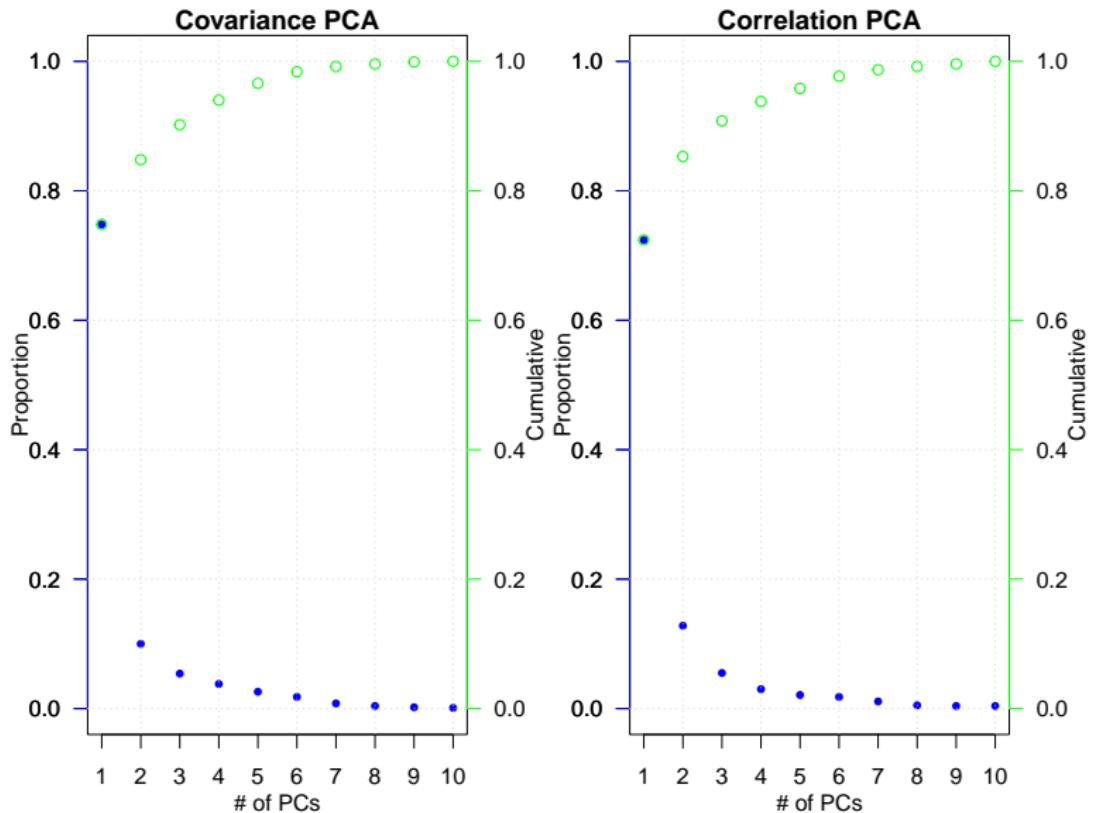
Background

Finding Principal Components

Principal Component Analysis in Practice



Correlation PCA versus Covariance PCA



Background

Finding Principal Components

Principal Component Analysis in Practice

Example: Monthly Sea Surface Temperatures

Background

Finding Principal
Components

Principal Component
Analysis in Practice

- 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) \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

Background

Finding Principal Components

Principal Component Analysis in Practice

The Empirical Orthogonal Function (EOF) Decomposition

Background

Finding Principal Components

Principal Component Analysis in Practice

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

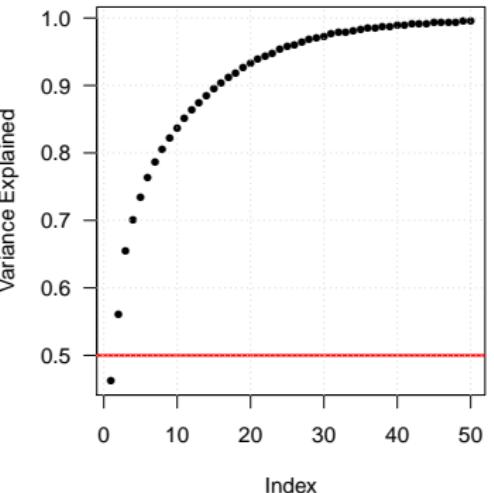
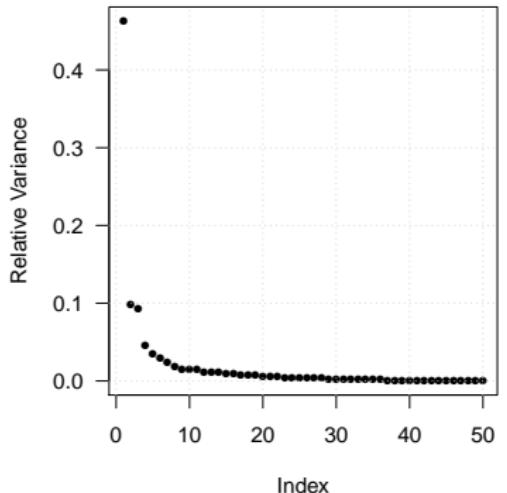
- To find principal spatial structures
- To reduce the dimension (spatially or temporally) in large spatio-temporal datasets

Screen Plot for EOFs

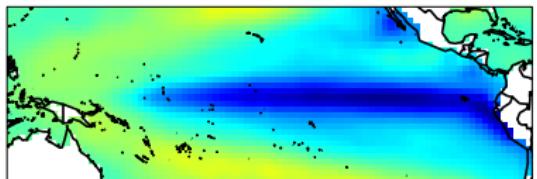
Background

Finding Principal Components

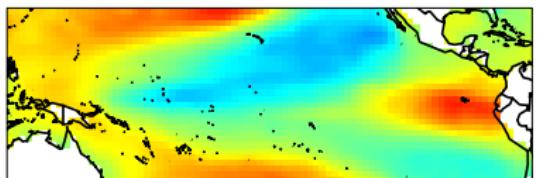
Principal Component Analysis in Practice



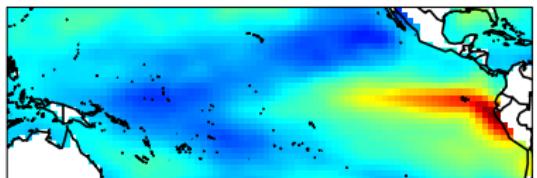
Perform EOF Decomposition and Plot the First Three Modes



EOF1: The classic
ENSO pattern



EOF2: A modulation
of the center



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

Background

Finding Principal
Components

Principal Component
Analysis in Practice

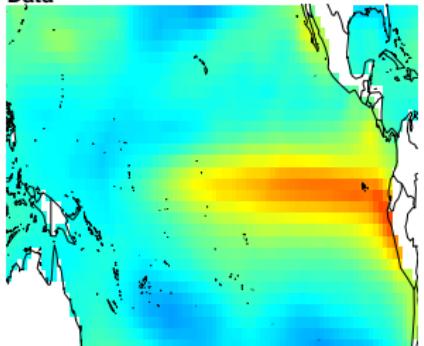
1998 Jan El Niño Event

Background

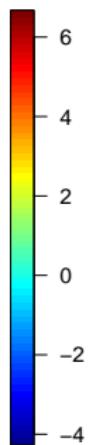
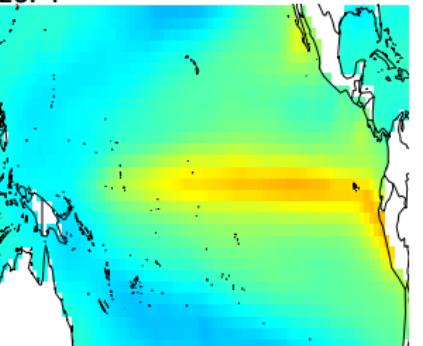
Finding Principal
Components

Principal Component
Analysis in Practice

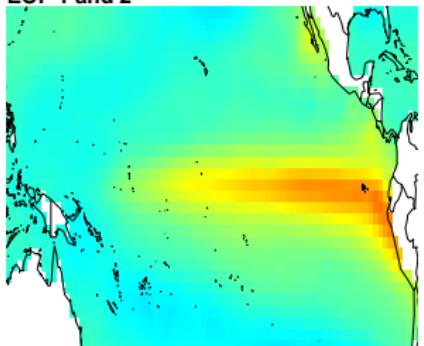
Data



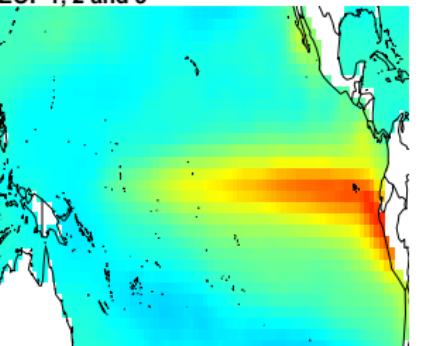
EOF 1



EOF 1 and 2



EOF 1, 2 and 3



In this lecture, we learned about:

- What Principal Components Analysis is
- How to find Principal Components
- How to use Principal Components to achieve dimensionality reduction

In the next lecture, we will learn about **Factor Analysis**

Background

Finding Principal
Components

Principal Component
Analysis in Practice