PCA Revealed

Part 4: Summarizing Information Approach

Gaston **S**anchez

August 2014

Readme

License:

Presentation

Algebraic Approach

In these slides we'll cover PCA from a Summarizing Information perspective

About

This approach is perhaps the most common way to present PCA in the multivariate literature

Reminder

PCA

Principal Components Analysis (PCA) allows us to study and explore a set of quantitative variables measured on a set of objects

Core Idea

With PCA we seek to reduce the dimensionality (reduce the number of variables) of a data set while retaining as much as possible of the variation present in the data

Summarizing Information Perspective

Data

The data structure for PCA is in tabular format, which can be mathematically handled as a **matrix** $\mathbf{X}$:

$$\mathbf{X}_{n,p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

- n objects in the rows
- p quantitative variables in the columns

Data Considerations

Variables

We will denote the p variables in $\mathbf{X}$ by $X_1, X_2, \dots, X_p$

Mean centered

For convenience (to make computations easier and notation simpler) we will assume that the data is centered

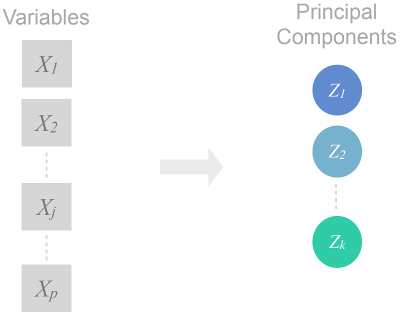
$$\bar{X}_j = \sum_{i=1}^{n} x_{ij} = 0$$

(i.e. *centered*: variables with mean $= 0$)

Looking for PCs

Overall Idea

Given a set of p variables $X_1, X_2, \dots, X_p$, we want to obtain new k variables $Z_1, Z_2, \dots, Z_k$, called the **Principal Components** (PCs)



Looking for PCs

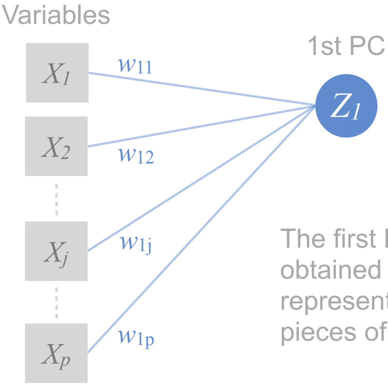
PC as linear combinations

We want to compute the **PCs as linear combinations** of the original variables.

$$
\begin{array}{rl}
\mathsf{PC}_1 \longrightarrow & Z_1 = w_{11}X_1 + w_{12}X_2 + \cdots + w_{1p}X_p \\
\mathsf{PC}_2 \longrightarrow & Z_2 = w_{21}X_1 + w_{22}X_2 + \cdots + w_{2p}X_p \\
\vdots \quad & \qquad\qquad\qquad \vdots \\
\mathsf{PC}_\mathsf{k} \longrightarrow & Z_k = w_{k1}X_1 + w_{k2}X_2 + \cdots + w_{kp}X_p
\end{array}
$$

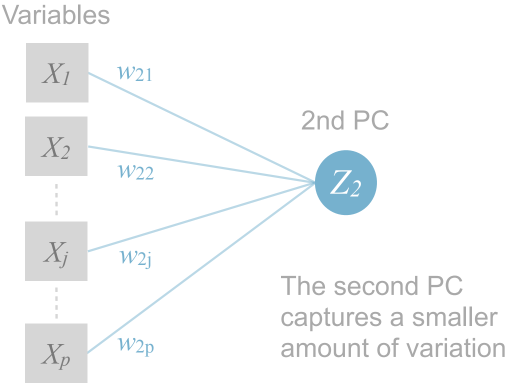
(i.e. linear combination = weighted sum)

1st PC

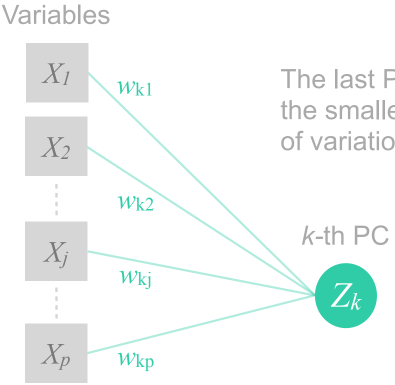


The first PC is obtained in order to represent the main pieces of variation

2nd PC

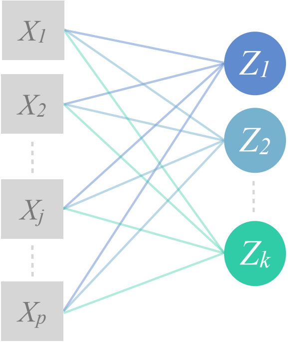

Variables
X_1
X_2
X_j
X_p
w_{21}
w_{22}
w_{2j}
w_{2p}
2nd PC
Z_2
The second PC captures a smaller amount of variation

k-th PC



The last PC captures the smallest amount of variation

PCs as linear combinations



$$= w_{11} X_1 + w_{12} X_2 + \ldots + w_{1p} X_p$$

$$= w_{21} X_1 + w_{22} X_2 + \ldots + w_{2p} X_p$$

$$= w_{k1} X_1 + w_{k2} X_2 + \ldots + w_{kp} X_p$$

Introductory Recap

Summarize Variation

We look to transform the original variables into a smaller set of new variables, the Principal Components, that summarize the variation in data.

PCs

The PCs are obtained as linear combinations (i.e. weighted sums) of the original variables. We look for PCs having maximum variance, and being mutually uncorrelated.

PC Requirements

Main Requirement

The main requirement for the Principal Components $Z_1, Z_2, \dots, Z_k$, is that they need to **capture most of the variation** in the data $\mathbf{X}$.

Convenient Requirement

To avoid a PC capturing the same variation as other PCs (i.e. avoiding redundant information), we also require them to be mutually orthogonal so they are uncorrelated with each other.

Algebraic Formulation

Capturing variation?

Variation

Looking for PCs that *capture most of the variation in the data* implies —in statistical terms— that we want to obtain **PCs with maximum variance**

In other words

We look for vectors of weights $\mathbf{w_j} = \{w_{j1}, w_{j2}, \dots, w_{jp}\}$ such that each component $Z_j = \mathbf{X}\mathbf{w_j}$ has maximum variance (for $j = 1, \dots, k$)

Algebraic Formulation

More formally

We want to find a vector $\mathbf{w_j}$ such that

$$\max_{\mathbf{w_j}} \ var(Z_j = \mathbf{X}\mathbf{w_j})$$

that is

$$\max_{\mathbf{w_j}} \ \mathbf{w_j'}\mathbf{X}'\mathbf{X}\mathbf{w_j}$$

Note that

- $\mathbf{X}'\mathbf{X}$ is the variance-covariance matrix
- Without constraints, the previous expression is unbounded

Maximization Constraints

Usefull Restriction

To get a feasible solution we need to impose the restriction that $\mathbf{w_j}$ is of unit norm: $\|\mathbf{w_j}\| = 1 \Rightarrow \mathbf{w_j'w_j} = 1$

Criterion to be maximized

If we denote $\mathbf{S} = \mathbf{X'X}$ then the criterion to be maximized is:

$$\max_{\mathbf{w_j}} \mathbf{w_j'Sw_j}$$

subject to $\quad \mathbf{w_j'w_j} = 1 \quad$ and $\quad Z_j' Z_h = 0 \quad (j \neq h)$

Pay Attention!

$$\mathbf{w}_j' \mathbf{S} \mathbf{w}_j$$

This expression is of extreme importance. Why?

- It is a quadratic form
- $\mathbf{S}$ is a semi-positive definite matrix
- $\mathbf{S}$ has non-negative real eigenvalues

Finding PCs

Finding 1st PC

How to find the 1st PC?

In order to get the first principal component $Z_1 = \mathbf{X}\mathbf{w_1}$, we need to find $\mathbf{w_1}$ such that

$$\max_{\mathbf{w_1}} \; \mathbf{w_1'}\mathbf{S}\mathbf{w_1}$$

subject to $\mathbf{w_j'}\mathbf{w_j} = 1$

Maximization... how?

Being a maximization problem, the typical procedure to find the solution is by using the Lagrangian multiplier method.

Lagrangian Multiplier

Finding 1st PC

Using Lagrange multipliers we get:

$$\mathbf{w_1'Sw_1} - \lambda(\mathbf{w_1'w_1} - 1)$$

Differentiation with respect to $\mathbf{w_1}$ gives:

$$\mathbf{Sw_1} - \lambda_1 \mathbf{w_1} = \mathbf{0}$$

Rearranging some terms we get:

$$\mathbf{Sw_1} = \lambda_1 \mathbf{w_1}$$

Lagrangian Multiplier Solution

What does this mean?

$$\mathbf{S}\mathbf{w_1} = \lambda_1 \mathbf{w_1}$$

It means that

- λ_1 is an eigenvalue of $\mathbf{S}$
- and $\mathbf{w_1}$ is the corresponding eigenvector

Finding 2nd PC

How to find the 2nd PC?

In order to find the second principal component $Z_2 = \mathbf{X}\mathbf{w_2}$, we need to find $\mathbf{w_2}$ such that

$$\max_{\mathbf{w_2}} \; \mathbf{w_2'}\mathbf{S}\mathbf{w_2}$$

subject to $\quad \|\mathbf{w_2}\| = 1 \quad$ and $\quad Z_1' Z_2 = 0$

(remember that Z_2 must be uncorrelated to Z_1)

Finding 2nd PC

Another eigenvalue-eigenvector pair

Applying the Lagrange multipliers, it can be shown that the desired $\mathbf{w_2}$ is such that

$$\mathbf{S}\mathbf{w_2} = \lambda_2 \mathbf{w_2}$$

In other words

- λ_2 is an eigenvalue of $\mathbf{S}$
- and $\mathbf{w_2}$ is the corresponding eigenvector

Finding all PCs

All PCs can be found simultaneously by **diagonalizing** $\mathbf{S}$

Diagonalizing $\mathbf{S}$ involves expressing it as the product:

$$\mathbf{S} = \mathbf{W}\mathbf{D}\mathbf{W}'$$

- $\mathbf{D}$ is a diagonal matrix
- the elements in the diagonal of $\mathbf{D}$ are the eigenvalues of $\mathbf{S}$
- the columns of $\mathbf{W}$ are orthonormal: $\mathbf{W}'\mathbf{W} = \mathbf{I}$
- the columns of $\mathbf{W}$ are the eigenvectors of $\mathbf{S}$
- $\mathbf{W}' = \mathbf{W}^{-1}$

Diagonalization of $\mathbf{S}$

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1p} \\ w_{21} & w_{22} & \cdots & w_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ w_{p1} & w_{p2} & \cdots & w_{pp} \end{pmatrix}$$

$$\mathbf{D} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}$$

$$\mathbf{S} = \mathbf{W}\mathbf{D}\mathbf{W}'$$

Diagonalization of $\mathbf{S}$

$$\mathbf{S} = \mathbf{WDW'} =$$

$$\begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1p} \\ w_{21} & w_{22} & \cdots & w_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ w_{p1} & w_{p2} & \cdots & w_{pp} \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix} \begin{pmatrix} w_{11} & w_{21} & \cdots & w_{p1} \\ w_{12} & w_{22} & \cdots & w_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1p} & w_{2p} & \cdots & w_{pp} \end{pmatrix}$$

Diagonalization and Eigenvalue Decomposition

Diagonalization?

A symmetric matrix $\mathbf{S}$ is **orthogonally diagonalizable** if and only if there is an orthogonal matrix $\mathbf{W}$ and a diagonal matrix $\mathbf{D}$ such that

$$\mathbf{S} = \mathbf{W}\mathbf{D}\mathbf{W}'$$

$$\mathbf{D} = \mathbf{W}'\mathbf{S}\mathbf{W}$$

Note that

- $\mathbf{W}' = \mathbf{W}^{-1}$
- the columns of $\mathbf{W}$ are orthonormal: $\mathbf{W}'\mathbf{W} = \mathbf{I}$
- the elements in the diagonal of $\mathbf{D}$ are the eigenvalues of $\mathbf{S}$
- the columns of $\mathbf{W}$ are the eigenvectors of $\mathbf{S}$

Spectral Theorem and Diagonalization

Eigenvalue Decomposition (EVD)

Diagonalizing a symmetric matrix is nothing more than obtaining its **eigenvalue decomposition** (a.k.a. spectral decomposition)

Spectral Decomposition of Symmetric Matrices

A $p \times p$ symmetric matrix $\mathbf{S}$ has the following properties:

- $\mathbf{S}$ has p real eigenvalues (counting multiplicites)
- the eigenvectors corresponding to different eigenvalues are orthogonal
- $\mathbf{S}$ is orthogonally diagonalizable $(\mathbf{S} = \mathbf{WDW}')$
- the set of eigenvalues of $\mathbf{S}$ is called the **spectrum** of $\mathbf{S}$

EVD Geometric Interpretation

PCA and its Geometrical Standpoint

PCA and EVD

The PCA solution can be obtained with an Eigenvalue Decomposition of the matrix $\mathbf{S} = \mathbf{X}'\mathbf{X}$

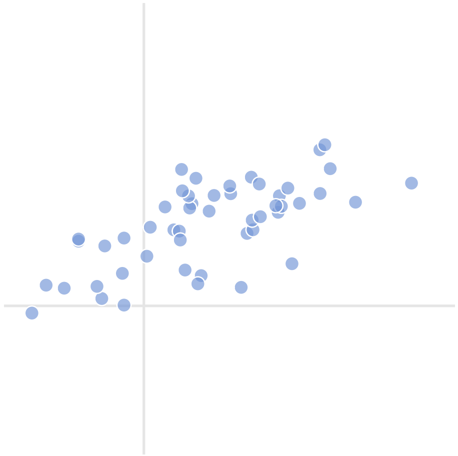
PCA Model

Let's talk about how can we give a geometric interpretation of the EVD and change of variable idea.

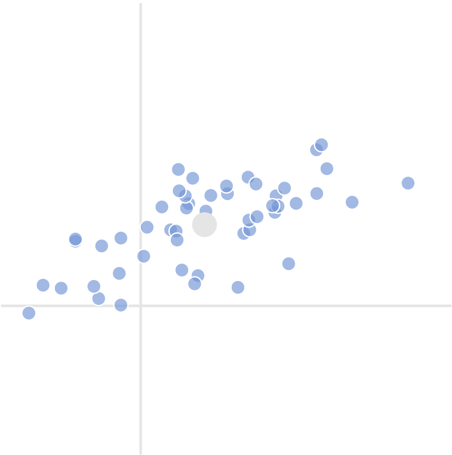
PCA and EVD

The main idea is that the variables in $\mathbf{X}$ are changed into PCs $\mathbf{Z}$. Let's see a toy example for illustration purposes.

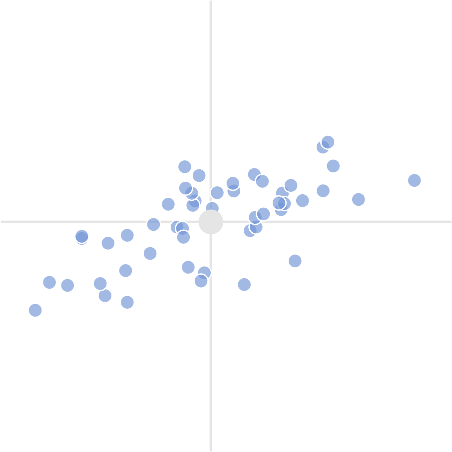
Toy Data (in 2-dimensions)



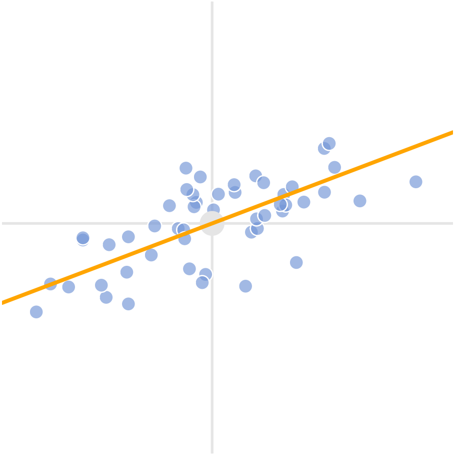
Mean Point (center)



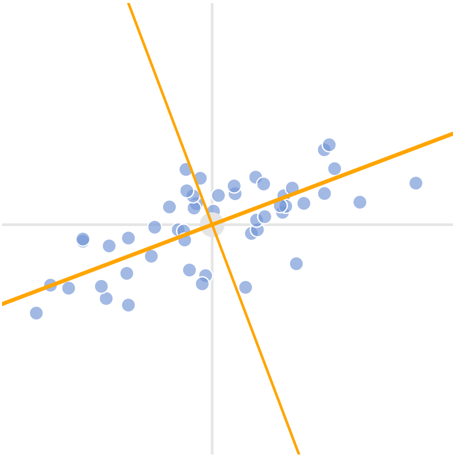
Mean-centering Data



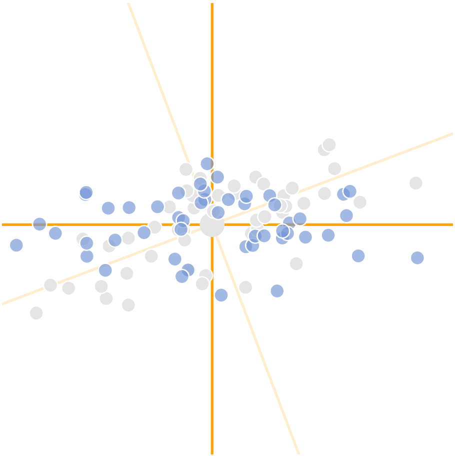
First PC (view as a change of variable)



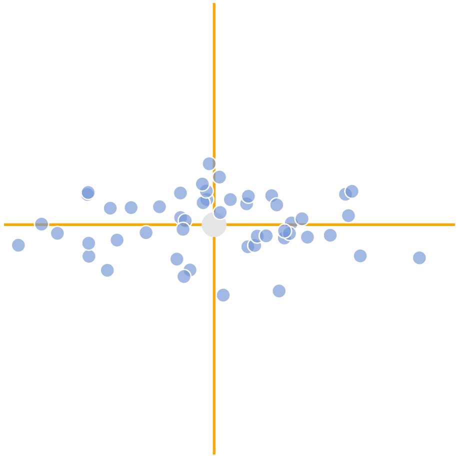
Second PC (view as a change of variable)



Before-and-After Change Comparison



Changed Variables (Rotated Data)



So far we've seen that ...

PCs in matrix format

$$\mathbf{Z}_{n,k} = \mathbf{X}_{n,p}\mathbf{W}_{p,k}$$

PCs are just linear combinations of $\mathbf{X}$ in which the weights $\mathbf{W}$ are obtained by diagonalizing $\mathbf{S} = \mathbf{X}'\mathbf{X}$

PCA Model

PCA model

We can obtain as many different eigenvalues as the rank of $\mathbf{S}$

$$\mathbf{X}_{n,p} = \mathbf{Z}_{n,k}\mathbf{P}'_{k,p}$$

where $\mathbf{Z}$ is the matrix of PCs or *scores*, and $\mathbf{P}$ is the matrix of *loadings*

PCA Model

PCA model

Formally, PCA involves finding scores and loadings such that the data can be expressed as a product of two matrices:

$$\mathbf{X}_{n,p} = \mathbf{Z}_{n,k}\mathbf{P'}_{k,p}$$

where $\mathbf{Z}$ is the matrix of PCs or *scores*, and $\mathbf{P}$ is the matrix of *loadings*

Ideally

We expect k to be much more smaller than p so we get a data reduction without losing too much information.

PCA and Data Decomposition

Computation of all PCs

We can obtain as many PCs as the rank of $\mathbf{X}$ (i.e. $k = rank(\mathbf{X})$)

$$\mathbf{X}_{n,p} = \mathbf{Z}_{n,k}\mathbf{P'}_{k,p}$$

Keeping just a few PCs

But usually we will only retain just a few PCs (i.e. $k \ll p$)

$$\mathbf{X}_{n,p} \approx \mathbf{Z}_{n,k}\mathbf{P'}_{k,p} + Residual$$

(just a few PCs will *optimally* summarize the main structure of the data)