Principal Components Analysis (part III)

Predictive Modeling & Statistical Learning

Gaston Sanchez

More PCA

Presentation

About

In these slides we will talk about PCA in the way that is usually introduced in most multivariate statistics books.

Data Matrix

Data

The analyzed data can be expressed in matrix format $\mathbf{X}$:

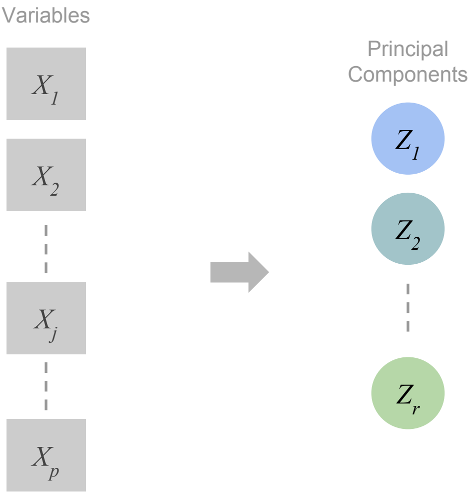
$$\underset{n \times p}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

- n objects in the rows
- p variables in the columns
- We'll assume standardized variables (mean $=$ 0, var $=$ 1)

Looking for PCs

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

Looking for PCs



Looking for PCs

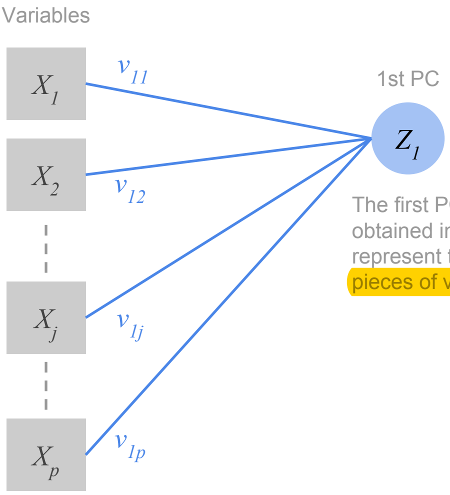
PC as linear combinations

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

$$\begin{aligned}
\text{PC}_1 \longrightarrow &\quad Z_1 = v_{11}X_1 + v_{12}X_2 + \cdots + v_{1p}X_p \\
\text{PC}_2 \longrightarrow &\quad Z_2 = v_{21}X_1 + v_{22}X_2 + \cdots + v_{2p}X_p \\
\vdots \qquad &\qquad\qquad\qquad \vdots \\
\text{PC}_\text{r} \longrightarrow &\quad Z_r = v_{r1}X_1 + v_{r2}X_2 + \cdots + v_{rp}X_p
\end{aligned}$$

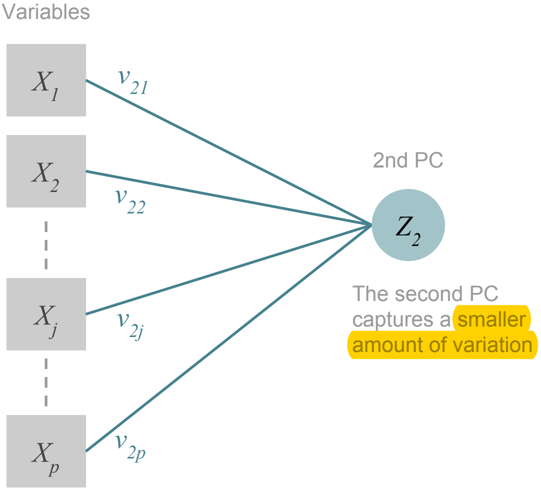
(i.e. linear combination = weighted sum)

1st PC



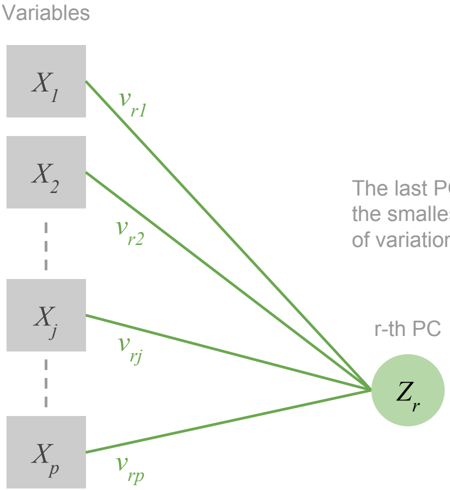
The first PC is obtained in order to represent the main pieces of variation

2nd PC



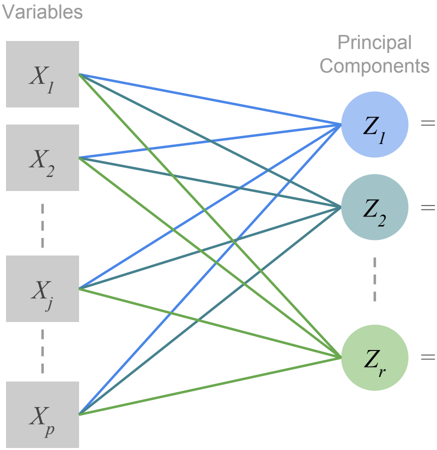
The second PC captures a smaller amount of variation

k-th PC



The last PC captures the smallest amount of variation

PCs as linear combinations



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

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

$$= v_{r1} X_1 + v_{r2} X_2 + \ldots + v_{rp} X_p$$

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.

Finding PCs

Capturing variation with PCs

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 $\mathbf{v_k} = \{v_{k1}, v_{k2}, \dots, v_{kp}\}$ such that each component $\mathbf{z_k} = \mathbf{X}\mathbf{v_k}$ has maximum variance (for $k = 1, \dots, r$)

Algebraic Formulation

More formally

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

$$\max_{\mathbf{v_k}} \; var(\mathbf{z_k} = \mathbf{X}\mathbf{v_k})$$

that is

$$\max_{\mathbf{v_k}} \left(\frac{1}{n-1}\right)\mathbf{v_k}^\top\mathbf{X}^\top\mathbf{X}\mathbf{v_k}$$

Note that:

- $\frac{1}{n-1}\mathbf{X}^\top\mathbf{X}$ is the correlation matrix
- Without constraints, the previous expression is unbounded

Maximization Constraints

Usefull Restriction

To get a feasible solution we need to impose the restriction that $\mathbf{v_k}$ is of unit norm: $\|\mathbf{v_k}\| = 1 \;\Rightarrow\; \mathbf{v_k}^\mathsf{T}\mathbf{v_k} = 1$

Criterion to be maximized

If we denote $\mathbf{S} = \frac{1}{n-1}\mathbf{X}^\mathsf{T}\mathbf{X}$, the criterion to be maximized is:

$$\max_{\mathbf{v_k}} \; \mathbf{v_k}^\mathsf{T}\mathbf{S}\mathbf{v_k}$$

subject to $\quad \mathbf{v_k}^\mathsf{T}\mathbf{v_k} = 1 \quad$ and $\quad \mathbf{z_k}^\mathsf{T}\mathbf{z_h} = 0 \quad (k \neq h)$

Pay Attention!

$$\mathbf{v_k}^\top \mathbf{S} \mathbf{v_k}$$

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 1st PC

Finding 1st PC

In order to find the first principal component $\mathbf{z_1} = \mathbf{X}\mathbf{v_1}$, we need to find $\mathbf{v_1}$ such that

$$\max_{\mathbf{v_1}} \left(\frac{1}{n-1}\right) \mathbf{v_1}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{v_1}$$

subject to $\mathbf{v_1}^\mathsf{T}\mathbf{v_1} = 1$

Note that $\frac{1}{n-1}$ is just a scalar, so the criterion is sometimes reexpresed as:

$$\max_{\mathbf{v_1}} \mathbf{v_1}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{v_1}$$

Finding 1st PC

$$\max_{\mathbf{v_1}} \mathbf{v_1}^\top \mathbf{X}^\top \mathbf{X} \mathbf{v_1}$$

What to do?

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:

$$\mathbf{v_1}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{v_1} - \lambda(\mathbf{v_1}^\mathsf{T}\mathbf{v_1} - 1)$$

Differentiation with respect to $\mathbf{v_1}$ gives:

$$\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{v_1} - \lambda_1\mathbf{v_1} = \mathbf{0}$$

Rearranging some terms we get:

$$\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{v_1} = \lambda_1\mathbf{v_1}$$

Lagrangian Multiplier Solution

What does this mean?

$$\left(\frac{1}{n-1}\right)\mathbf{X}^{\top}\mathbf{X}\mathbf{v_1} = \lambda_1\mathbf{v_1}$$

Lagrangian Multiplier Solution

What does this mean?

$$\left(\frac{1}{n-1}\right)\mathbf{X}^{\top}\mathbf{X}\mathbf{v_1} = \lambda_1\mathbf{v_1}$$

It means that

- λ_1 is an eigenvalue of $\frac{1}{n-1}\mathbf{X}^{\top}\mathbf{X}$
- and $\mathbf{v_1}$ is the corresponding eigenvector

Finding 2nd PC

Finding 2nd PC

How to find the 2nd PC

In order to find the second principal component $\mathbf{z_2} = \mathbf{X}\mathbf{v_2}$, we need to find $\mathbf{v_2}$ such that

$$\max_{\mathbf{v_2}} \left(\frac{1}{n-1}\right) \mathbf{v_2}^\mathsf{T} \mathbf{X}^\mathsf{T} \mathbf{X} \mathbf{v_2}$$

subject to $\quad \|\mathbf{v_2}\| = 1 \quad$ and $\quad \mathbf{z_1}^\mathsf{T} \mathbf{z_2} = 0$

Finding 2nd PC

Another eigenvalue-eigenvector pair

Applying the Lagrange multipliers, it can be shown that the desired $\mathbf{v_2}$ is such that

$$\left(\frac{1}{n-1}\right)\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{v_2} = \lambda_2\mathbf{v_2}$$

In other words

- λ_2 is an eigenvalue of $\frac{1}{n-1}\mathbf{X}^\mathsf{T}\mathbf{X}$
- and $\mathbf{v_2}$ is the corresponding eigenvector

Matrix Decompositions

Finding all PCs

Diagonalization

All Principal Components can be found simultaneously by **diagonalizing** $\frac{1}{n-1}\mathbf{X}^\mathsf{T}\mathbf{X}$

Eigenvalue Decomposition (EVD)

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

Data Decomposition

Algebraically

PCA involves an **Eigen-Value Decomposition** (EVD) of the data matrix $\frac{1}{n-1}\mathbf{X}^\top\mathbf{X}$, that is:

$$\frac{1}{n-1}\mathbf{X}^\top\mathbf{X} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$$

- $\mathbf{V}$ is orthonormal matrix of eigenvectors (i.e. $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$)
- $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues

EVD Approach

PCs

Principal components $\mathbf{Z} = [Z_1|Z_2|\ldots|Z_k]$ are obtained as:

$$\mathbf{Z} = \mathbf{X}\mathbf{V}$$

Note that the variance of each component turns out to be equal to its associated eigenvalue:

$$var(\mathbf{z_k}) = \frac{1}{\sqrt{n-1}}\mathbf{z_k}^{\mathsf{T}}\mathbf{z_k} = \lambda_k$$

PCs for approximating Data

PCA and Data Decomposition

Interestingly, we can also express $\mathbf{X}$ in terms of the PCs $\mathbf{Z}$ and the loadings (eigenvectors) $\mathbf{V}$ as:

$$\mathbf{X} = \mathbf{Z}\mathbf{V}^\top$$

PCA and Data Decomposition

Assuming that $\mathbf{X}$ is of full-column rank (i.e. $p = rank(\mathbf{X})$)

$$\underset{(n,p)}{\mathbf{X}} = \underset{(n,p)}{\mathbf{Z}}\,\underset{(p,p)}{\mathbf{V}^\mathsf{T}}$$

PCA and Data Decomposition

Assuming that $\mathbf{X}$ is of full-column rank (i.e. $p = rank(\mathbf{X})$)

$$\underset{(n,p)}{\mathbf{X}} = \underset{(n,p)}{\mathbf{Z}} \underset{(p,p)}{\mathbf{V}}^{\mathsf{T}}$$

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

$$\underset{(n,p)}{\mathbf{X}} \approx \underset{(n,k)}{\mathbf{Z}} \underset{(k,p)}{\mathbf{V}}^{\mathsf{T}} = \underset{(n,p)}{\hat{\mathbf{X}}}$$

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

SVD Decomposition

SVD

Recall that any matrix $\mathbf{M}$ or rank p can be decomposed as a product of three simpler matrices $\mathbf{U}$, $\mathbf{D}$ and $\mathbf{V}$

$$\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathsf{T}}$$

- $\mathbf{U}_{n,p}$ (left singular vectors)
- $\mathbf{D}_{p,p}$ (singular values)
- $\mathbf{V}_{p,p}$ (right singular vectors)

SVD Approach

PCA via SVD

It can be shown that PCA involves a **SVD** based on the data matrix $\mathbf{X}$

$$\left(\frac{1}{\sqrt{n-1}}\right)\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathsf{T}} = \mathbf{Z}\mathbf{V}^{\mathsf{T}}$$

where:

- $\mathbf{Z} = \mathbf{U}\mathbf{D}$ is the matrix of PCs (or scores)
- $\mathbf{V}$ is the matrix of Loadings

Changing Scales

You can actually play with the scale of the principal components and the variable factors:

$$
\begin{aligned}
\left(\frac{1}{\sqrt{n-1}}\right)\mathbf{X} &= \mathbf{UDV}^{\mathsf{T}} \\
&= (\mathbf{UD})\mathbf{V}^{\mathsf{T}} \\
&= \mathbf{U}(\mathbf{DV}^{\mathsf{T}}) \\
&= (\mathbf{UD}^{1/2})(\mathbf{D}^{1/2}\mathbf{V}^{\mathsf{T}}) \\
&= (\mathbf{UD}^{\alpha})(\mathbf{D}^{1-\alpha}\mathbf{V}^{\mathsf{T}})
\end{aligned}
$$

with $0 \leq \alpha \leq 1$

What does PCA look like?

PCA and its Geometrical Standpoint

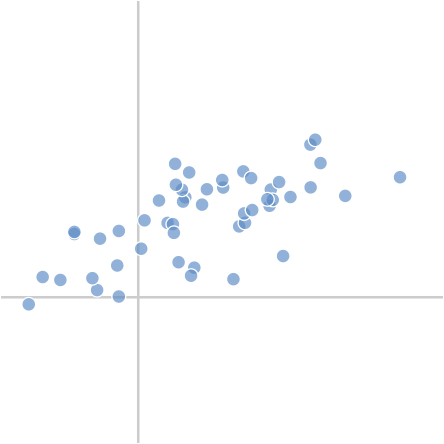
PCA and EVD

We've seen that the PCA solution can be obtained with an Eigenvalue Decomposition of the matrix $\mathbf{S} = \frac{1}{n-1}\mathbf{X}^{\mathsf{T}}\mathbf{X}$

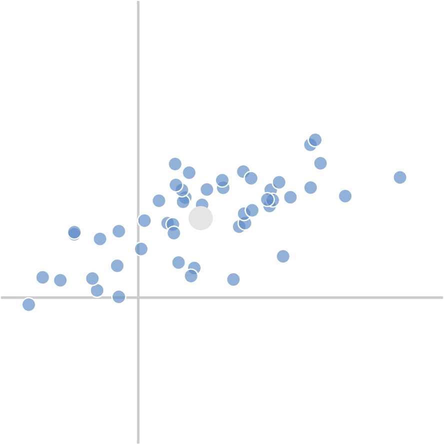
Now let's briefly talk about how can we give a geometric interpretation of the EVD and the concept of *change of variable*.

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

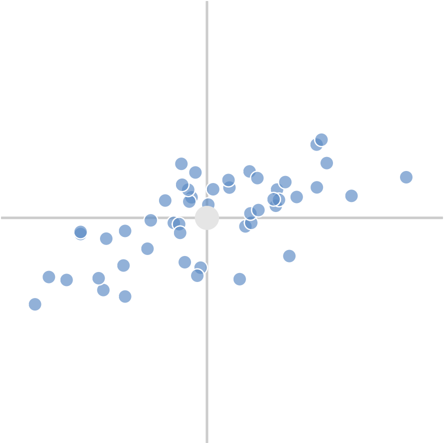
Toy Data (in 2-dimensions)



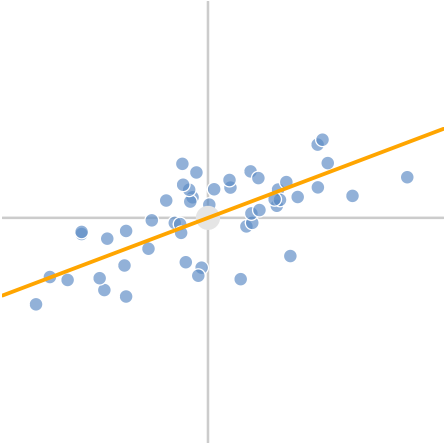
Mean Point (center)



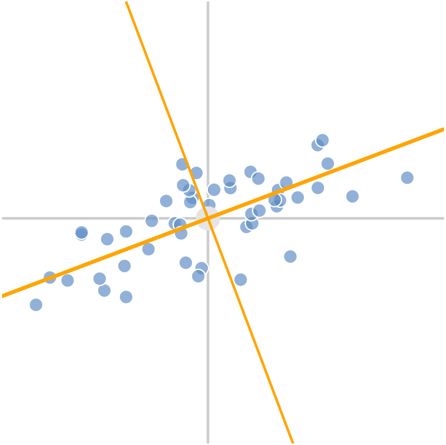
Mean-centering Data



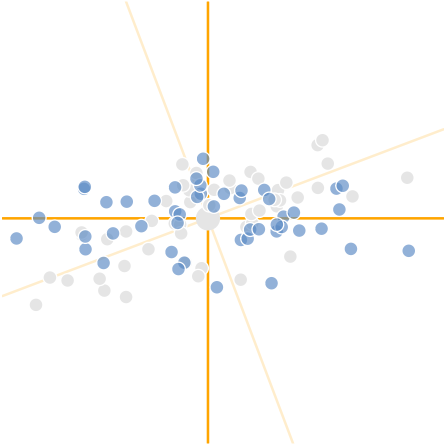
First PC (view as a change of variable)



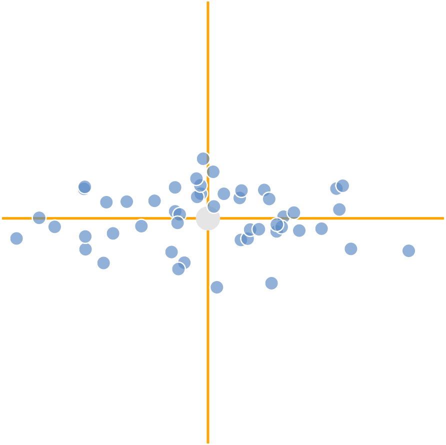
Second PC (view as a change of variable)



Before-and-After Change Comparison



Changed Variables (Rotated Data)



References

- **Principal Component Analysis** by Herve Abdi and Lynne Williams (2010). *Wiley Interdisciplinary Reviews: Computational Statistics. Volume 2(4), 433-459*.
- **An R and S-Plus Companion to Multivariate Analysis** by Brian Everitt (2004). *Chapter 3: Principal Components Analysis*. Springer.
- **Principal Component Analysis** by Ian jolliffe (2002). Springer.
- **Data Mining and Statistics for Decision Making** by Stephane Tuffery (2011). *Chapter 7: Factor Analysis*. Editions Technip, Paris.
- **Exploratory Multivariate Analysis by Example Using R** by Husson, Le and Pages (2010). *Chapter 1: Principal Component Analysis (PCA)*. CRC Press.

References (French Literature)

- **Statistique Exploratoire Multidimensionnelle** by Lebart et al (2004). *Chapter 3, section 3: Analyse factorielle discriminante*. Dunod, Paris.
- **Probabilites, analyse des donnees et statistique** by Gilbert Saporta (2011). *Chapter 6: Analyse en Composantes Principaux*. Editions Technip, Paris.
- **Statistique: Methodes pour decrire, expliquer et prevoir** by Michel Tenenhaus (2008). *Chapter 10: L'analyse discriminante*. Dunod, Paris.
- **Analyses factorielles simples et multiples** by Brigitte Escofier et Jerome Pages (2016, 5th edition). *Chapter 2: L'analyse discriminante*. Dunod, Paris.