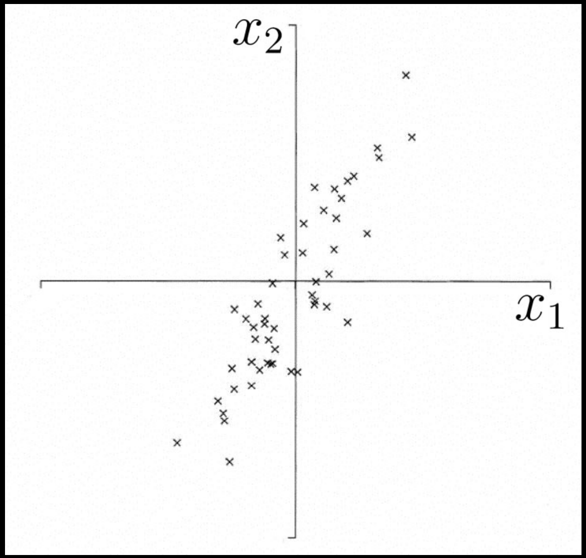
an introduction to Principal Component Analysis (PCA)

abstract

Principal component analysis (PCA) is a technique that is useful for the compression and classification of data. The purpose is to reduce the dimensionality of a data set (sample) by finding a new set of variables, smaller than the original set of variables, that nonetheless retains most of the sample's information.

By information we mean the variation present in the sample, given by the correlations between the original variables. The new variables, called principal components (PCs), are uncorrelated, and are ordered by the fraction of the total information each retains.

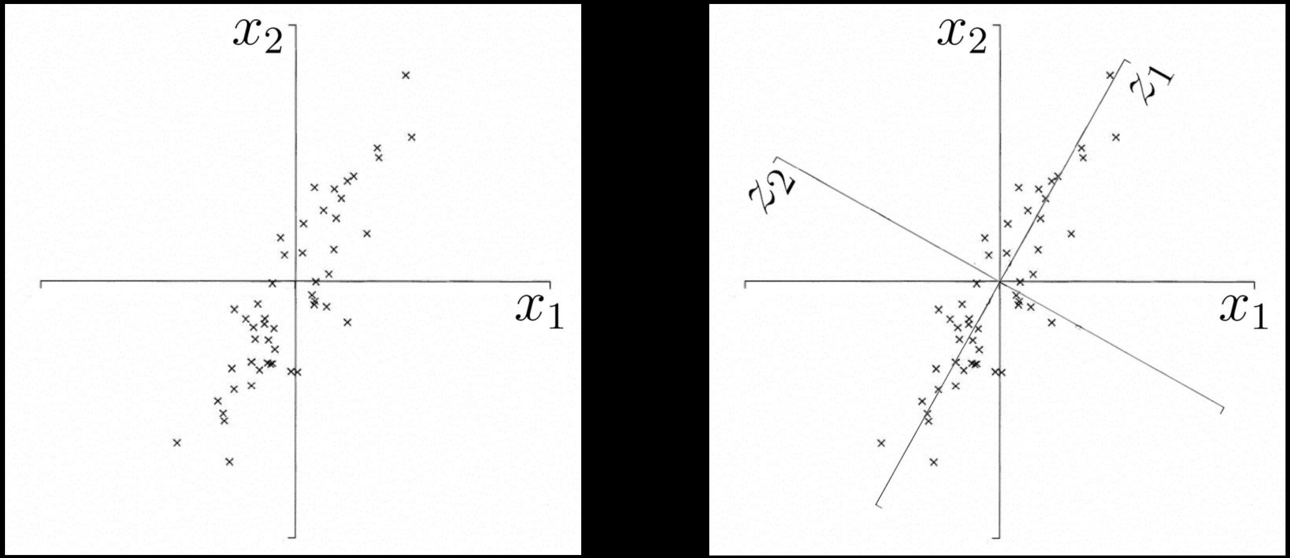
Geometric picture of principal components (PCs)



A sample of *n* observations in the 2-D space $\mathbf{x} = (x_1, x_2)$

Goal: to account for the variation in a sample
in as few variables as possible, to some accuracy

Geometric picture of principal components (PCs)



- the 1st PC z_1 is a minimum distance fit to a line in $\mathbf{x}$ space
- the 2nd PC z_2 is a minimum distance fit to a line in the plane perpendicular to the 1st PC

PCs are a series of linear least squares fits to a sample, each orthogonal to all the previous.

Algebraic definition of PCs

Given a sample of *n* observations on a vector of *p* variables

$$\mathbf{x} = (x_1, x_2, \ldots, x_p)$$

define the first principal component of the sample by the linear transformation

$$z_1 \equiv \mathbf{a}_1^{\mathrm{T}}\mathbf{x} = \sum_{i=1}^{p} \mathrm{a}_{i1} x_i$$

where the vector

$$\mathbf{a}_1 = (\mathrm{a}_{11}, \mathrm{a}_{21}, \ldots, \mathrm{a}_{p1})$$

is chosen such that

$$\mathrm{var}[z_1]$$ is maximum

Algebraic definition of PCs

Likewise, define the k^{th} PC of the sample by the linear transformation

$$z_k \equiv \mathbf{a}_k^{\mathrm{T}} \mathbf{x} \qquad k = 1, \ldots, p$$

where the vector $\mathbf{a}_k = (\mathrm{a}_{1k}, \mathrm{a}_{2k}, \ldots, \mathrm{a}_{pk})$

is chosen such that $\mathrm{var}[z_k]$ is maximum

subject to $\mathrm{cov}[z_k, z_l] = 0$ for $k > l \geq 1$

and to $\mathbf{a}_k^{\mathrm{T}} \mathbf{a}_k = 1$

Algebraic derivation of coefficient vectors $\mathbf{a}_k$

To find $\mathbf{a}_1$ first note that

$$\mathrm{var}[z_1] = \langle z_1^2 \rangle - \langle z_1 \rangle^2$$

$$= \sum_{i,j=1}^{p} \mathrm{a}_{i1}\mathrm{a}_{j1}\langle x_i x_j \rangle - \sum_{i,j=1}^{p} \mathrm{a}_{i1}\mathrm{a}_{j1}\langle x_i \rangle \langle x_j \rangle$$

$$= \sum_{i,j=1}^{p} \mathrm{a}_{i1}\mathrm{a}_{j1} S_{ij} \quad \text{where} \quad S_{ij} \equiv \sigma_{x_i x_j} = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle$$

$$= \mathbf{a}_1^{\mathrm{T}} \mathbf{S} \mathbf{a}_1$$

$\mathbf{S}$ is the covariance matrix for the variables $\mathbf{x} = (x_1, x_2, \ldots, x_p)$

Algebraic derivation of coefficient vectors $\mathbf{a}_k$

To find $\mathbf{a}_1$ maximize $\mathrm{var}[z_1]$ subject to $\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 = 1$

Let λ be a Lagrange multiplier

then maximize $\mathbf{a}_1^{\mathrm{T}}\mathbf{S}\mathbf{a}_1 - \lambda(\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 - 1)$

by differentiating… $\mathbf{S}\mathbf{a}_1 - \lambda\mathbf{a}_1 = 0$

$$\Rightarrow (\mathbf{S} - \lambda\mathbf{I}_p)\mathbf{a}_1 = 0$$

therefore

$\mathbf{a}_1$ is an eigenvector of $\mathbf{S}$

corresponding to eigenvalue $\lambda \equiv \lambda_1$

Algebraic derivation of $\mathbf{a}_k$

We have maximized

$$\mathrm{var}[z_1] = \mathbf{a}_1^{\mathrm{T}}\mathbf{S}\mathbf{a}_1 = \mathbf{a}_1^{\mathrm{T}}\lambda_1\mathbf{a}_1 = \lambda_1$$

So λ_1 is the largest eigenvalue of $\mathbf{S}$

The first PC z_1 retains the greatest amount of variation in the sample.

Algebraic derivation of coefficient vectors $\mathbf{a}_k$

To find the next coefficient vector $\mathbf{a}_2$ maximize $\mathrm{var}[z_2]$

subject to $\mathrm{cov}[z_2, z_1] = 0$

and to $\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 = 1$

First note that

$$\mathrm{cov}[z_2, z_1] = \mathbf{a}_1^{\mathrm{T}}\mathbf{S}\mathbf{a}_2 = \lambda_1\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_2$$

then let λ and φ be Lagrange multipliers, and maximize

$$\mathbf{a}_2^{\mathrm{T}}\mathbf{S}\mathbf{a}_2 - \lambda(\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 - 1) - \phi\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_1$$

Algebraic derivation of coefficient vectors $\mathbf{a}_k$

We find that $\mathbf{a}_2$ is also an eigenvector of $\mathbf{S}$ whose eigenvalue $\lambda \equiv \lambda_2$ is the second largest.

In general

$$\mathrm{var}[z_k] = \mathbf{a}_k^{\mathrm{T}} \mathbf{S} \mathbf{a}_k = \lambda_k$$

- The k^{th} largest eigenvalue of $\mathbf{S}$ is the variance of the k^{th} PC.
- The k^{th} PC z_k retains the k^{th} greatest fraction of the variation in the sample.

Algebraic formulation of PCA

Given a sample of n observations on a vector of p variables $\mathbf{x} = (x_1, x_2, \ldots, x_p)$

define a vector of p PCs $\mathbf{z} = (z_1, z_2, \ldots, z_p)$

according to $\mathbf{z} = \mathbf{A}^{\mathrm{T}}\mathbf{x}$

where $\mathbf{A}$ is an orthogonal p x p matrix whose k^{th} column is the k^{th} eigenvector $\mathbf{a}_k$ of $\mathbf{S}$

Then $\mathbf{\Lambda} = \mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{A}$ is the covariance matrix of the PCs,

being diagonal with elements $\Lambda_{ij} = \lambda_i \delta_{ij}$