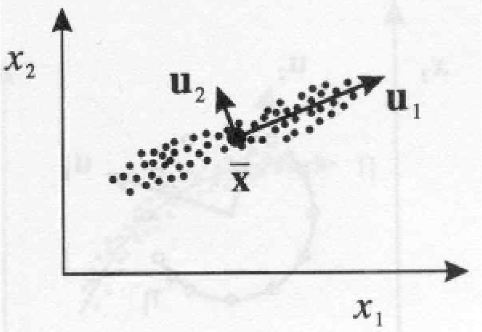
MAC 0459 / 5865

Data Science and Engineering

R. Hirata Jr. (hirata@ime.usp.br)

Class 16 (2020)

Idea



Idea - Variance is related to dispersion of the variable.
Problem - Find a linear transform such that the coordinates are associated to a system whose axis correspond to the larger dispersion of the data.

PCA

PCA: Method that tries to explain the variance-covariance structure in terms of linear combinations of the original variables.

Objetive (PCA):

- interpretation
- dimension reduction

$$\mathrm{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_d \end{bmatrix} \Longrightarrow \mathrm{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d'} \end{bmatrix} \qquad d' \ll d$$

Linear combination

Let a be a vector such that:

$$\mathrm{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

then,

$$\mathrm{a}^t\mathrm{x} = (a_1\ a_2\ a_3) \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = a_1x_1 + a_2x_2 + a_3x_3$$

($\mathrm{a}^t\mathrm{x}$ is a linear combination of the x_i variables.)

Linear combination

Let a be a vector such that:

$$\mathrm{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

then,

$$\mathrm{a}^t\mathrm{x} = (a_1\; a_2\; a_3) \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = a_1x_1 + a_2x_2 + a_3x_3$$

($\mathrm{a}^t\mathrm{x}$ is a linear combination of the x_i variables.)

Fact

$E(\mathrm{a}^t\mathrm{x}) = \mathrm{a}^t E(\mathrm{x}) = \mathrm{a}^t\mu$

$Var(\mathrm{a}^t\mathrm{x}) = \mathrm{a}^t\Sigma\mathrm{a}$

A set of linear combinations

Let y_1, y_2, ..., y_d be d linear combinations of the original variables $x_1, x_2, \ldots, x_d$.

For all $i = 1, 2, \ldots, d$, let

$$y_i = \mathrm{a}_i^t \mathrm{x}$$

Fact

$$Var(y_i) = \mathrm{a}_i^t \, \Sigma \, \mathrm{a}_i$$

$$Cov(y_i, y_j) = \mathrm{a}_i^t \, \Sigma \, \mathrm{a}_j$$

Covariance matrix

Covariance matrix: $\Sigma = Cov(\mathrm{x})$

$$\begin{bmatrix} E(x_1-\mu_1)^2 & E(x_1-\mu_1)(x_2-\mu_2) & \dots & E(x_1-\mu_1)(x_d-\mu_d) \\ E(x_2-\mu_2)(x_1-\mu_1) & E(x_2-\mu_2)^2 & \dots & E(x_2-\mu_2)(x_d-\mu_d) \\ \vdots & \vdots & \dots & \vdots \\ E(x_d-\mu_d)(x_1-\mu_1) & E(x_d-\mu_d)(x_2-\mu_2) & \dots & E(x_d-\mu_d)^2 \end{bmatrix}$$

For $d = 3$:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{bmatrix}$$

Correlation matrix

Correlation (between components x_i and x_j)

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}$$

Notation: R (correlation matrix)

Correlation matrix

Correlation (between components x_i and x_j)

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}$$

Notation: R (correlation matrix)

OBS.: the diagonal of R is composed of 1s !

For $d = 3$:

$$\mathrm{R} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix}$$

Important property

The covariance/correlation matrix Σ is symmetrical and positive semi-definite ($x^t \Sigma x > 0, \forall x \neq 0$)

PCA

The linear combinations are guided to maximize the variance of the resulting variables y_i and, at the same time, be linear independent (orthogonal).

PCA - algorithm

- **Principal component**: linear combination $\mathsf{a}_1^t \mathsf{x}$ that maximizes the variance of $y_1 = \mathsf{a}_1^t \mathsf{x}$, subject to $\mathsf{a}_1^t \mathsf{a}_1 = 1$

PCA - algorithm

- **Principal component**: linear combination $\mathrm{a}_1^t\mathrm{x}$ that maximizes the variance of $y_1 = \mathrm{a}_1^t\mathrm{x}$, subject to $\mathrm{a}_1^t\mathrm{a}_1 = 1$

- **Second principal component**: linear combination $\mathrm{a}_2^t\mathrm{x}$ that maximezes the variance of y_2, subject to $\mathrm{a}_2^t\mathrm{a}_2 = 1$ and null covariance in relation to the principal component.

PCA - algorithm

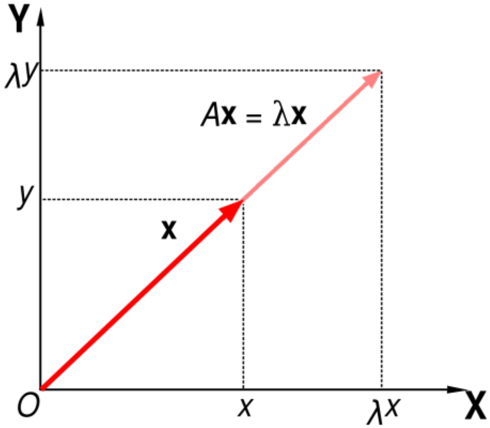
- **Principal component**: linear combination $\mathrm{a}_1^t\mathrm{x}$ that maximizes the variance of $y_1 = \mathrm{a}_1^t\mathrm{x}$, subject to $\mathrm{a}_1^t\mathrm{a}_1 = 1$
- **Second principal component**: linear combination $\mathrm{a}_2^t\mathrm{x}$ that maximezes the variance of y_2, subject to $\mathrm{a}_2^t\mathrm{a}_2 = 1$ and null covariance in relation to the principal component.
- ***i*-th principal component**: linear combinatin $\mathrm{a}_i^t\mathrm{x}$ that maximezes the variance of y_i, subject to $\mathrm{a}_i^t\mathrm{a}_i = 1$ and null covariance in relation to all previous principal components.

Eigenvalues and eigenvectors

Let A be a squared matrix that represents a linear transformation T. λ is an **eigenvalue** of T with the respective **eigenvector** $x \neq 0$ if

$$Ax = \lambda x$$

Geometrical interpretation



Eigenvectors x are vectors such that x and $T(\mathrm{x})$ have the same direction.

The effect of T on the eigenvectors is by only a scalar factor (there is no rotation).

How to computer eigenvalues and eigenvectors?

From

$$A\mathrm{x} = \lambda \mathrm{x}$$

follows that

$$(A - \lambda I)\mathrm{x} = 0$$

Because $\mathrm{x} \neq 0$**, we have that** $A - \lambda I$ **is not invertible.**
(Why? Because if $A - \lambda I$ **were invertible, if we multiply both sides by the inverse, we would have** $\mathrm{x} = 0$**).**

A matrix is invertible if and only if its determinant is not null.
Therefore $det(A - \lambda I) = 0$**.**

How to computer eigenvalues and eigenvectors?

From Linear Algebra, the equation $det(A - \lambda I) = 0$ is the characteristic polynomial of A. Solving the equation we obtain the eigenvalues of A.

To compute the eigenvector associated to each eigenvalue λ, it is enough to find a x that satisfies

$$A\mathrm{x} = \lambda\mathrm{x}$$

Example

$$A = \begin{bmatrix} 1 & -5 \\ -5 & 1 \end{bmatrix}$$

$$det(A - \lambda I) = \begin{vmatrix} 1-\lambda & -5 \\ -5 & 1-\lambda \end{vmatrix} = (1-\lambda)^2 - (-5)^2 = \lambda^2 - 2\lambda - 24$$

$$\lambda^2 - 2\lambda - 24 = 0 \Longleftrightarrow \lambda = 4 \text{ or } \ \lambda = -6$$

Eigenvector associated to eigenvalue $\lambda_1 = 6$

$$\begin{bmatrix} 1 & -5 \\ -5 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 6 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \Longrightarrow \mathrm{v}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Eigenvector associated to eigenvalue $\lambda_2 = -4$

$$\begin{bmatrix} 1 & -5 \\ -5 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = -4 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \Longrightarrow \mathrm{v}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Normalizing:

$$\mathrm{e}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix} \qquad \mathrm{e}_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

Example

$T(x, y) = (y, x)$

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

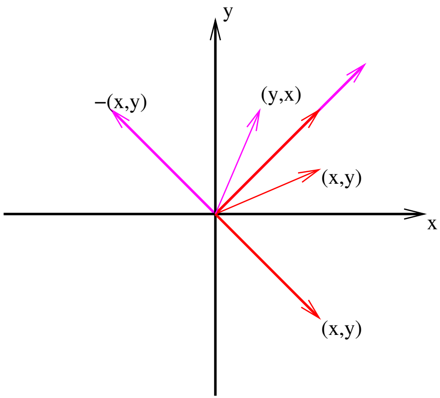
Eigenvalues:

$$(A - \lambda I)\mathsf{x} = 0 \Longleftrightarrow \left| \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right| = 0$$

$$\Longleftrightarrow \begin{vmatrix} -\lambda & 1 \\ 1 & -\lambda \end{vmatrix} = 0 \Longleftrightarrow \lambda^2 - 1 = 0 \Longleftrightarrow \lambda = \pm\sqrt{1}$$

Eigenvectors :

$$\mathsf{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \mathsf{v}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$



In this case $T(x, y) = (y, x)$, the eigenvectors are $(1, 0)^t$ and $(1, -1)^t$

The importance of eigenvectors and eigenvalues

Eigenvectors are vectors that are invariant to T (but to a scalar factor).

If we consider another **basis** (another coordinate system), a **matrix that represents this tranformation** T in this new basis will be **different**.

An important result is that **the basis is built by the eigenvectors of** T. The matrix that represents this T is a **diagonal** matrix built by their respective eigenvalues. This simplifies the algebraic operations.

Principal component

Σ: covariance matrix of x
Because Σ is symmetrical, it has all d real eigenvalues.

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$ be the eigenvalues of Σ and $\mathrm{e}_1, \mathrm{e}_2, \ldots, \mathrm{e}_\mathrm{d}$ the respective normalized eigenvectors.

Now consider the decomposition:

$$\begin{bmatrix} \\ y \\ \\ \end{bmatrix} = \begin{bmatrix} e_1^t \\ e_2^t \\ \vdots \\ e_d^t \end{bmatrix} \begin{bmatrix} \\ x \\ \\ \end{bmatrix}$$

Principal components

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$ be the eigenvalues of Σ and $\mathsf{e}_1, \mathsf{e}_2, \ldots, \mathsf{e}_\mathsf{d}$ the respective normalized eigenvectors.

Principal components

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$ be the eigenvalues of Σ and $\mathrm{e}_1, \mathrm{e}_2, \ldots, \mathrm{e}_\mathrm{d}$ the respective normalized eigenvectors.

One can show that

the i-th principal component is given by:

$$y_i = \mathrm{e}_i^t \mathrm{x}$$

Even more,

$$Var(y_i) = \lambda_i$$

and

$$Cov(y_i, y_j) = 0, \forall j < i$$

Principal components

Why is e_1 the direction of larger dispersion ??
Why is e_2 the second direction of larger dispersion ??

Remember that $Var(y_i) = a_i^t \Sigma a_i$

The explanation is based on the fact that:
Let B is a positive definite matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d > 0$ and their respective normalized eigenvectors, $e_1, e_2, \ldots, e_d$, then,

$$\max_{x \neq 0} \frac{x^t B x}{x^t x} = \lambda_1$$

when $x = e_1$.

$$\max_{x \perp e_1, \ldots, e_k} \frac{x^t B x}{x^t x} = \lambda_{k+1}$$

when $x = e_{k+1}$.

Why

$$Var(y_i) = \lambda_i\,?$$

because

$$\max_{x \neq 0} \frac{x^t \Sigma x}{x^t x} = \lambda_1 = \frac{e_1^t \Sigma e_1}{e_1^t e_1} = e_1^t \Sigma e_1 = Var(y_1)$$

(and $e_1^t e_1 = 1$).

Besides that, $Cov(y_i, y_k) = 0$ because e_i e e_k are orthogonal each other.

Therefore

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{dd} = \sum_{i=1}^{d} Var(x_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_d = \sum_{i=1}^{d} Var(y_i)$$

Therefore

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{dd} = \sum_{i=1}^{d} Var(x_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_d = \sum_{i=1}^{d} Var(y_i)$$

Demo: $\Sigma = M\Lambda M^t$ where Λ is a diagonal matrix where the diagonal is composed by the eigenvalues of Σ, M is the matrix with the respective eigenvectors.

First equality: trivial, because the diagonal of Σ has the variances of x_i.

Second equality: $tr(\Sigma) = tr(M\Lambda M^t) \stackrel{(*)}{=} tr(\Lambda M^t M) \stackrel{(**)}{=} tr(\Lambda)$.

(*) because $tr(AB) = tr(BA)$
(**) because $M^t M = MM^t$ (M is a matrix of normalized eigenvectors)

Interpretation

The previous result show that the total variance of the dataset is equal to the sum of all the eigenvalues.

Therefore, the **percentage of total variance explained by the *k*-th component** is:

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

PCA of normalized variables

Normalization of a Random Variable

- subtract the mean and divide by the standard deviation
- the resulting RV mean 0 and variance 1

PCA of normalized variables

Normalization of a Random Variable

- subtract the mean and divide by the standard deviation
- the resulting RV mean 0 and variance 1

Fact: The covariance matrix of normalized variables is equal to the covariance matrix of the original variables.

PCA of normalized variables

Normalization of a Random Variable

- subtract the mean and divide by the standard deviation
- the resulting RV mean 0 and variance 1

Fact: The covariance matrix of normalized variables is equal to the covariance matrix of the original variables.

Therefore, to compute the normalized RV we can compute the eigenvalues and eigenvectors of the original correlation matrix.

PCA of normalized variables

PCA with normalized variables:
the sum of eigenvalues (variances) is é d

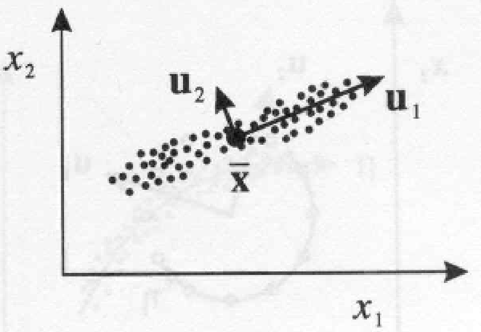
OBS.: the eigenvalues of the correlation matrix are not equal to the eigenvalues of the covariance matrix!

Reduction of dimension using PCA

Idea - choose a new representation in a subspace of lower dimension

$$\mathrm{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_d \end{bmatrix} \Longrightarrow \mathrm{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d'} \end{bmatrix} \qquad d' \ll d$$

In this example, instead of using $x = (x_1, x_2)$, we could use a projection of x on axis u_1 ($y_1 = u_1^t x$)



Dimensionality reduction using PCA

- How many components to choose?
- How large is the error doing that?
- Is the approach acceptable?

Dimensionality reduction using PCA

How many components to choose?

Choose the first d' principal components such that

$$\frac{\sum_{i=1}^{d'} \lambda_i}{\sum_{i=1}^{d} \lambda_i} > T$$

Usually, $T = 0.90$ or $T = 0.95$

Dimensionality reduction using PCA

How large is the error doing that?
M is the eigenvalues of Σ

$$\mathsf{y} = M\mathsf{x}$$

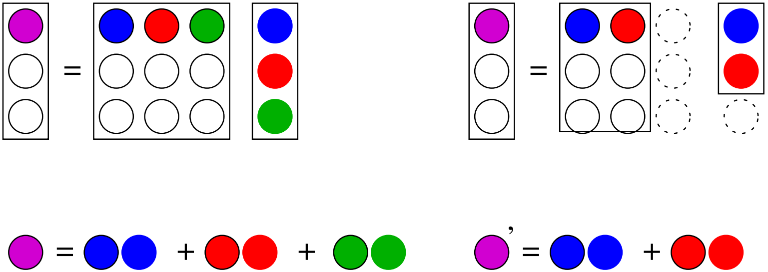
$$\mathsf{x} = M^{-1}\mathsf{y}$$

What if we do not consider all eigenvectors in the reconstruction of x ?

Dimensionality reduction using PCA

How large is the error doing that?

$$x = M^{-1}y$$



(the dimentions of small dispertion are left out)

Dimensionality reduction using PCA

How large is the error doing that?

x' represents x using only d' principal components

Error: $e = ||\mathsf{x} - \mathsf{x}'||$

$$e = \frac{1}{2} \sum_{i=d'+1}^{d} \lambda_i$$

PCA may not be good

PCA is good to simplify the representation (dimensionality reduction) of the dataset, to better preserve the information dispersion.

However, it may not be interesting to discriminate data.