The singular value decomposition

4433LALG3: Linear Algebra

Week 3, Lecture 10, Valente Ramírez

Mathematical & Statistical Methods group — Biometris, Wageningen University & Research

Overview

- Sample principal components
- The singular value decomposition
- Supplementary material

References:

- Nicholson §8.6.1

Section 1

Sample principal components

Population principal components

Last lecture we discussed the **population** principal components.

They were computed directly from the (population) covariance matrix Σ:

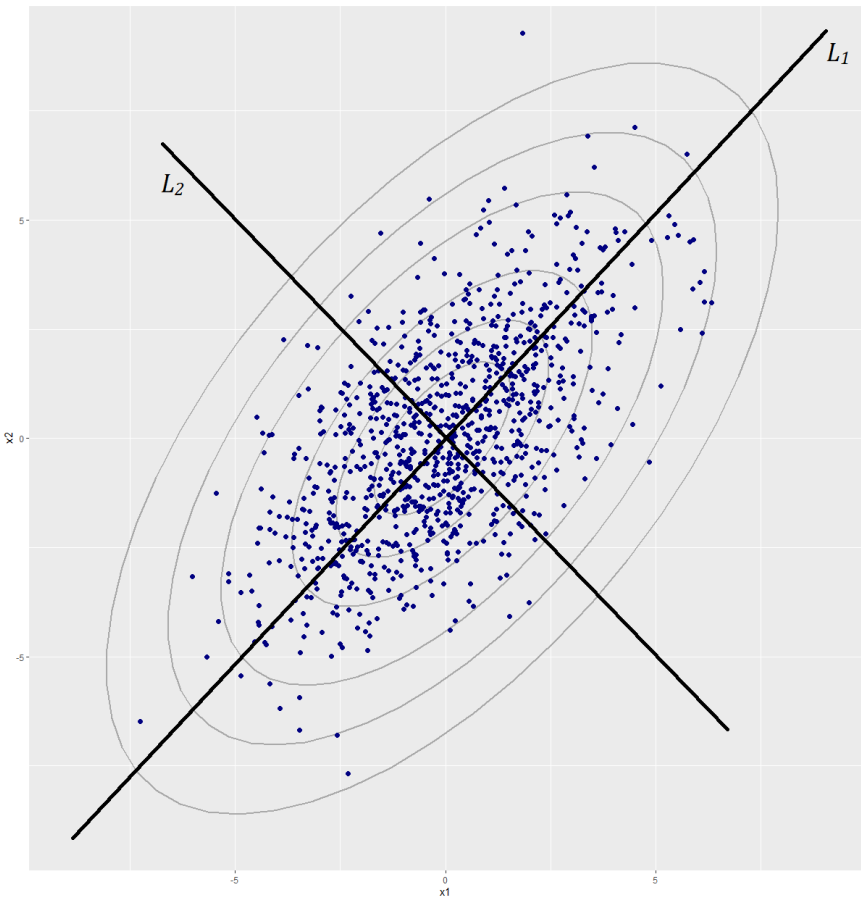
$$W_1 = \mathbf{v}_1^\top \mathbf{X}, \quad W_2 = \mathbf{v}_2^\top \mathbf{X},$$

where $\mathbf{v}_i$ are eigenvectors of Σ (normalized to have length one).

The spectral theorem guarantees that the eigenvectors are *orthogonal*.

But in practice we never know Σ, we only have a sample from $\mathbf{X}$.

Question: How to compute the PCs from the sample?



Sample principal components

The answer is (of course): use the sample covariance matrix S!

Suppose we have an $n \times p$ matrix, $X = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_p \end{bmatrix}$, of independent observations of a random vector $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, with Σ unknown.

To keep things simple, we assume that the data is centered:

$$\overline{\mathbf{x}}_i = 0, \quad \text{for each} \quad i = 1, \dots, p.$$

Then $S = \dfrac{1}{n-1} X^\top X$, is a symmetric $p \times p$ matrix.

By the spectral theorem, there exists an orthogonal $p \times p$ matrix V, such that

$$S = V \widehat{D} V^\top,$$

where $\widehat{D}$ is a diagonal matrix containing the eigenvalues of S:

$$\widehat{D} = \operatorname{diag}\{\hat{\lambda}_1, \dots, \hat{\lambda}_p\}, \quad \text{with} \quad \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_p \geq 0.$$

Sample principal components

Let $\mathbf{v}_k$ denote the k^{th} column of the orthogonal matrix V (thus an eigenvector of S).

The k^{th} **sample** principal component refers to the n–vector:

$$\mathbf{w}_k = X\mathbf{v}_k = v_{1k}\mathbf{x}_1 + v_{2k}\mathbf{x}_2 + \ldots v_{pk}\mathbf{x}_p.$$

The interpretation is the following:

- The eigenvector $\mathbf{v}_k$ provides coefficients to combine the variables $X_1, \ldots, X_p$.
- The new variable $\widehat{W}_k = v_{1k}X_1 + v_{2k}X_2 + \ldots + v_{pk}X_k$ is an *estimate* of the (unknown) k^{th} population principal component.
- The n–vector $\mathbf{w}_k$ is regarded as a sample of size n from the population principal component W_k.

Properties

The first principal component $\mathbf{w}_1$ satisfies:

- $\mathbf{w}_1$ is a linear combination of the columns of X: $\mathbf{w}_1 = X\mathbf{v}$,
- The coefficient vector $\mathbf{v}$ is such that:
  - It maximizes the (sample) variance of $X\mathbf{v}$,
  - subject to the constraint $\mathbf{v}^\top \mathbf{v} = 1$ (e.g. $\|\mathbf{v}\| = 1$) .

Subsequent components satisfy:

- $\mathbf{w}_k$ is a linear combination of the columns of X: $\mathbf{w}_k = X\mathbf{v}$,
- The coefficient vector $\mathbf{v}$ is such that:
  - It maximizes the (sample) variance of $X\mathbf{v}$,
  - subject to the constraint that $\mathbf{w}_k$ is uncorrelated (in the sample) to $\mathbf{w}_1, \ldots, \mathbf{w}_{k-1}$,
  - and subject to $\mathbf{v}^\top \mathbf{v} = 1$ (e.g. $\|\mathbf{v}\| = 1$) .

Summary

The analysis on the previous slides yielded, from the $n \times p$ matrix X:

- A list of values: $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$
  - These numbers are non-negative, because they are eigenvalues of the positive semi-definite matrix S
- A list of p–vectors: $\mathbf{v}_1, \ldots, \mathbf{v}_p$
  - These vectors are orthonormal, because they are eigenvectors of the symmetric matrix S
- A list of n–vectors: $\mathbf{w}_1, \ldots, \mathbf{w}_p$
  - These vectors are obtained by setting: $\mathbf{w}_k = X\mathbf{v}_k$
  - They are orthogonal, because they are represent uncorrelated observations

Note: The vectors $\mathbf{w}_k$ do not form an orthonormal set, because they are not required to have length one.

If we want an orthonormal set, we could normalize them: $\mathbf{u}_k = \dfrac{\mathbf{v}_k}{\|\mathbf{v}_k\|}$.

Application: PCA and dimensionality reduction

Example

The weekly rates of return for five stocks (JP Morgan, Citibank, Wells Fargo, Royal Dutch Shell,and ExxonMobil) listed on the New York Stock Exchange were determined for the period January 2004 through December 2005.

The observations in 103 successive weeks appear to be independently distributed, but the rates of return across stocks are correlated.

Is it possible to capture most of the total sample variance in only two variables?

```
# Data set is stored as 'stock'
> S <- cov(stock)  # Covariance matrix
> round(S, 3)
          JpMorgan Citibank WellsF RDShell ExonMob
JpMorgan     4.333    2.757  1.590   0.641   0.890
Citibank     2.757    4.387  1.800   1.815   1.233
WellsF       1.590    1.800  2.240   0.734   0.605
RDShell      0.641    1.815  0.734   7.225   5.083
ExonMob      0.890    1.233  0.605   5.083   7.657
```

[1] Example from: Johnson, Wichern – Applied Multivariate Statistical Analysis, 6^{th} ed.

Application: PCA and dimensionality reduction

We are working with variables $x_1, \ldots, x_4$ whose sample covariance matrix is

$$S = \begin{bmatrix} s_1^2 & s_{12} & s_{13} & s_{14} \\ s_{21} & s_2^2 & s_{23} & s_{24} \\ s_{31} & s_{32} & s_3^2 & s_{34} \\ s_{41} & s_{42} & s_{43} & s_4^2 \end{bmatrix},$$

where the numerical value of S was given on the previous slide.

The **total sample variance** is: $s_1^2 + s_2^2 + s_3^2 + s_4^2 = \operatorname{tr} S$.

We will compute the principal components w_i and check how much of this variance is explained by the first two PCs. That is:

$$\frac{\operatorname{Var}(w_1) + \operatorname{Var}(w_2)}{\operatorname{tr} S} \times 100\%.$$

Application: PCA and dimensionality reduction

```
# Data set is stored as 'stock'
> S <- cov(stock)  # Covariance matrix
> eigenS <- eigen(S)

# Total variance
> tr(S)
[1] 25.842

# Eigenvalues of S correspond to variance of each PC
> round(eigenS$values, 2)
[1] 13.68  7.01  2.54  1.43  1.19

# The total sum is the same
> sum(eigenS$values)
[1] 25.842

# Percentage of total variance explained by first 2 PCs
> 100*(eigenS$values[1] + eigenS$values[2]) / sum(eigenS$values)
[1] 80.06

# Eigenvectors of S
> round(eigenS$vectors, 2)
     [,1]  [,2]  [,3]  [,4]  [,5]
[1,] 0.22  0.63 -0.33 -0.66 -0.12
[2,] 0.31  0.57  0.25  0.41  0.59
[3,] 0.15  0.34  0.04  0.50 -0.78
[4,] 0.64 -0.25  0.64 -0.31 -0.15
[5,] 0.65 -0.32 -0.65  0.22  0.09
```

Section 2

The singular value decomposition

Motivation

- We begun with a (data) matrix X of size $n \times p$ (thus not squared).
- PCA gave us "a frame of reference" given by the vectors $\mathbf{v}_k$, and "special values" $\hat{\lambda}_k$ that were useful for understanding X.

Warning

The matrix X is not squared in general, so the $\hat{\lambda}_i$ are not eigenvalues of X, and the $\mathbf{v}_i$ are not eigenvectors of X.

Despite the warning, the above were very useful to study X.

It would be nice to have a similar technique for an arbitrary matrix A, not necessarily a data matrix (e.g. one representing a linear transformation).

Strategy

Key idea

Given an arbitrary $m \times n$ matrix A, we can construct two symmetric matrices:

- $A^\top A$, of size $n \times n$ (called the **Gram matrix** of A),
- $AA^\top$, of size $m \times m$.

We could apply the spectral theorem to these, in order to obtain their eigenvalues and (orthogonal) eigenvectors.

Hopefully, this will give us interesting information about A.

Singular values

Theorem

- *The eigenvalues of $A^\top A$ and $AA^\top$ are real and non-negative.*
  - *E.g. They are both positive semi-definite symmetric matrices.*
- *$A^\top A$ and $AA^\top$ have the same set of **positive** eigenvalues.*
- *The above matrix $A^\top A$ has:*
  - *Exactly r positive eigenvalues $\lambda_1, \ldots, \lambda_r$, where r is the rank of A.*
  - *If $r < n$, it also has $n - r$ zero-eigenvalues: $\lambda_{r+1} = \ldots = \lambda_n = 0$.*

Definition

The numbers $\sigma_i = \sqrt{\lambda_i}$ $(i = 1, \ldots, n)$ are called the **singular values** of A.

We always order them in decreasing order:

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0 \quad \text{and} \quad \sigma_i = 0 \ \text{ if } \ i > r.$$

Singular value decomposition

We continue in our attempt to follow the principal component approach on our $m \times n$ matrix A.

We have the singular values:

- $\sigma_1, \ldots, \sigma_r$

The spectral theorem guarantees the existence of an orthogonal matrix V such that $P^\top (A^\top A)\, P$ is diagonal. This gives orthonormal n–vectors:

- $\mathbf{v}_1, \ldots, \mathbf{v}_n$; these form an orthogonal $n \times n$ matrix V.

Finally, let $\mathbf{w}_k = A\mathbf{v}_k$, and $\mathbf{u}_k = \dfrac{\mathbf{w}_k}{\|\mathbf{w}_k\|}$. The set $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ is orthonormal.
If $r < m$, choose, in whichever way you want, vectors $\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m$ such that:

- $\mathbf{u}_1, \ldots, \mathbf{u}_m$ are orthonormal and form an orthogonal $m \times m$ matrix U.

Singular value decomposition

A **singular value decomposition** of A is the process of expressing (e.g. *decomposing*) the $m \times n$ matrix A as:

$$A = U\Sigma_A V^\top,$$

where

- U is an orthogonal $m \times m$ matrix,
- Σ_A is an $m \times n$ matrix containing the singular values σ_i of A:

$$\Sigma_A = \begin{bmatrix} D_A & 0 \\ 0 & 0 \end{bmatrix}_{m \times n}, \quad \text{where} \quad D = \operatorname{diag}\{\sigma_1, \ldots, \sigma_r\}.$$

- V is an orthogonal $n \times n$ matrix.

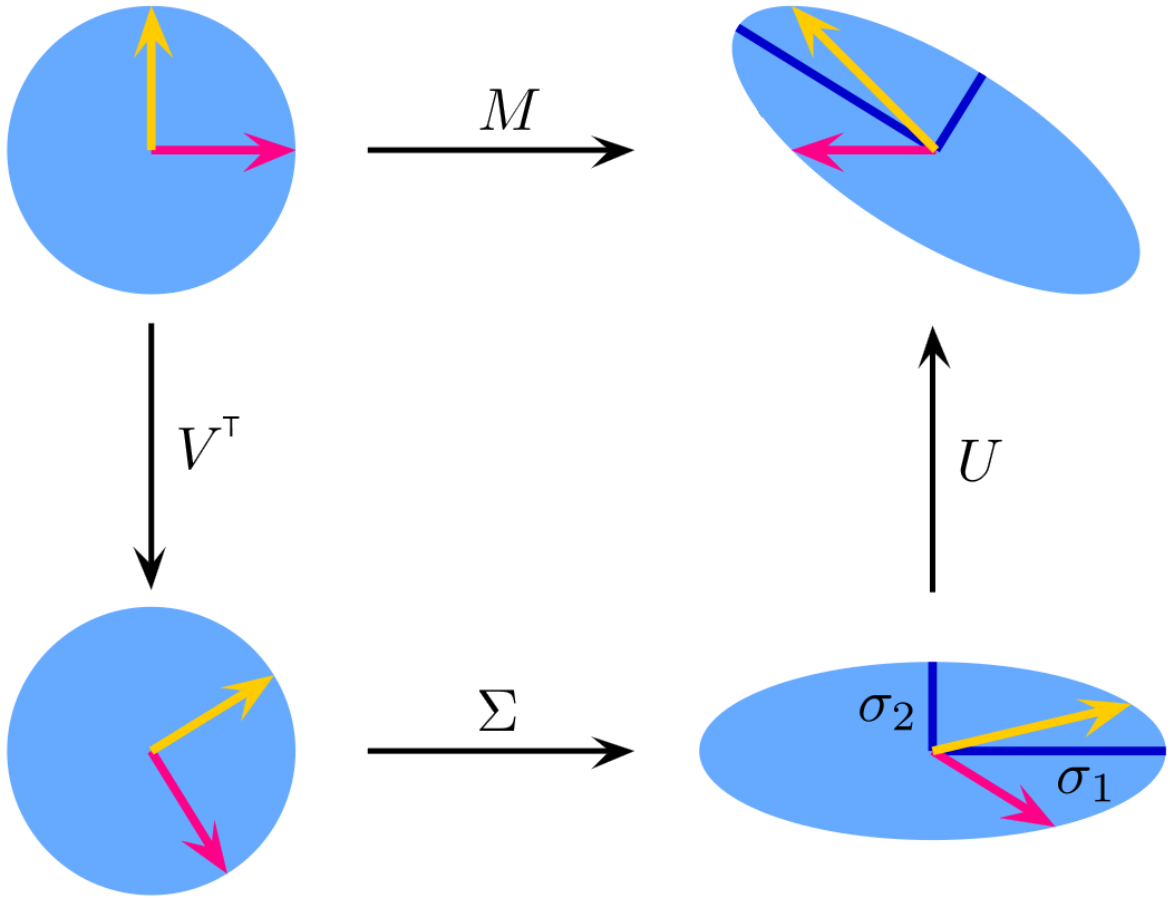
Singular value decomposition

The singular value decomposition, $A = U\Sigma_A V^\top$, is a generalization of the spectral decomposition for symmetric matrices: $S = PDP^\top$.

The differences are:

- The spectral decomposition applies only to symmetric matrices S
  - The singular value decomposition is relevant for any matrix A
- In the spectral decomposition, D is diagonal
  - The matrix Σ_A is not even square, although its elements are arranged along the "diagonal"
- The diagonal elements of D are the eigenvalues of S
  - A does not have eigenvalues, but the non-zero elements of Σ_A are the singular values
- In the spectral decomposition, P and $P^\top$ are inverses of each other
  - U and V can't be inverses because they don't even have the same size!

Geometric interpretation of the SVD



$$M = U \cdot \Sigma \cdot V^\mathsf{T}$$

Modified from: `https://commons.wikimedia.org/wiki/File:Singular-Value-Decomposition.svg`

Applications

- Principal component analysis
- Computing orthonormal bases for column space and null space
- Defining a *"pseudoinverse"* for non-square matrices
- Analysis of *ill-conditioned* problems in numerical linear algebra
- Data compression
- ...

Application: the condition number

Example

We have collected a few observations on three variables: x_1, x_2, y, and wish to fit a multiple linear regression to explain y in terms of the x_i.

We postulate a model of the form

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

and fit parameters using ordinary least-squares.

Question: How sensitive are these estimates to small measuring errors?

Application: the condition number

```
# Original dataset
> X <- dat[,1:2]; X  # design matrix
     x1    x2
1  3.00  3.00
2  5.00  5.00
3  1.00  1.02

> y <- dat[,3]; y  # observations of response variable
[1] 2 2 2

> lm(y ~ x1 + x2 - 1, dat)$coeff  # OLS fit
      x1       x2
  -76.00    76.47
```

Suppose the last observation of x_2 had been 1.01 instead of 1.02.

```
# Modified dataset
> dat[3,2] <- 1.01; dat[,1:2]
     x1    x2
1  3.00  3.00
2  5.00  5.00
3  1.00  1.01

> lm(y ~ x1 + x2 - 1, dat)$coeff  # OLS fit
    x1       x2
-152.5    152.9
```

Application: the condition number

A tiny change in the data resulted in a huge change on the estimates!

The *condition number* of a matrix is a way to quantify this sensitivity.

Definition

Let A be an $m \times n$ matrix, $A \neq 0$. The **condition number** of A is defined as:

$$\kappa(A) = \frac{\sigma_{\text{max}}(A)}{\sigma_{\text{min}}(A)},$$

where

- $\sigma_{\text{max}}(A)$ is the largest (non-zero) singular value of A,
- $\sigma_{\text{min}}(A)$ is the smallest non-zero singular value of A.

Application: the condition number

```
# Compute the singular values of the design matrix
> round(svd(X)$d, 3)
[1] 8.368 0.007

# Compute the condition number
> svd(X)$d[1] / svd(X)$d[2]
[1] 1200.834

# A built-in function for the condition number
> kappa(X, exact=TRUE)
[1] 1200.834
```

The condition number is extremely large!

Which is to be expected, x_1 and x_2 are almost identical!

Conclusion:

This model on this data set is extremely unstable and should be avoided.

Section 3

Supplementary material

SVD in R

```
# SVD example
> A <- matrix(c(
     1,0,1,0,0,
    -1,1,1,0,0,
     0,0,0,2,1,
     0,0,0,1,2), 5,4)

> svd(A)
  $d
  [1] 3.000000 1.732051 1.414214 1.000000

  $u
             [,1]       [,2]       [,3]          [,4]
  [1,]  0.0000000  0.5773503 -0.7071068  2.266233e-17
  [2,]  0.0000000 -0.5773503  0.0000000  4.532467e-17
  [3,]  0.0000000 -0.5773503 -0.7071068 -2.266233e-17
  [4,] -0.7071068  0.0000000  0.0000000 -7.071068e-01
  [5,] -0.7071068  0.0000000  0.0000000  7.071068e-01

  $v
             [,1] [,2] [,3]       [,4]
  [1,]  0.0000000    0   -1  0.0000000
  [2,]  0.0000000   -1    0  0.0000000
  [3,] -0.7071068    0    0 -0.7071068
  [4,] -0.7071068    0    0  0.7071068
```

SVD in R

```
# Copare the output of svd() to an analysis of t(A) %*% A
> Gram <- t(A) %*% A   # The Gram matrix of A

> eigen(Gram)$values
  [1] 9 3 2 1

# The singular values are the square-roots of these eigenvalues
> sqrt(eigen(Gram)$values)
  [1] 3.000000 1.732051 1.414214 1.000000

# The vectors 'v' are +/- the eigenvectors of Gram
> eigen(Gram)$vectors
            [,1] [,2] [,3]       [,4]
  [1,] 0.0000000    0    1  0.0000000
  [2,] 0.0000000    1    0  0.0000000
  [3,] 0.7071068    0    0  0.7071068
  [4,] 0.7071068    0    0 -0.7071068

# For each 'v', Av gives a multiple of the corresponding 'u'
> A %*% eigen(Gram)$vectors
           [,1] [,2] [,3]       [,4]
  [1,] 0.00000   -1    1  0.0000000
  [2,] 0.00000    1    0  0.0000000
  [3,] 0.00000    1    1  0.0000000
  [4,] 2.12132    0    0  0.7071068
  [5,] 2.12132    0    0 -0.7071068
```