Multivariate Verfahren

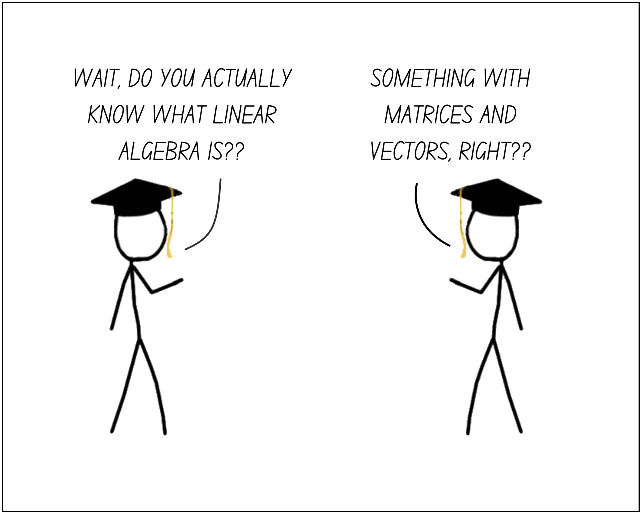
2. Basics of Linear Algebra

Hannah Kümpel

Institut für Statistik, LMU München

Sommersemester 2023

What was Linear Algebra again? And why do we need it? I


WAIT, DO YOU ACTUALLY KNOW WHAT LINEAR ALGEBRA IS??
SOMETHING WITH MATRICES AND VECTORS, RIGHT??

What was Linear Algebra again? And why do we need it? II

- Linear Algebra $\hat{=}$ Study of linear sets of equations & their transformation properties.

- But, "something with matrices and vectors" isn't far off at all! We can think of linear algebra as using mathematical operations on vectors and matrices to create new vectors and matrices.

- Example: We can "create" the vector $(x, y, z)^\top \in \mathbb{R}^3$ by solving

System of Linear Equations **Matrix representation**

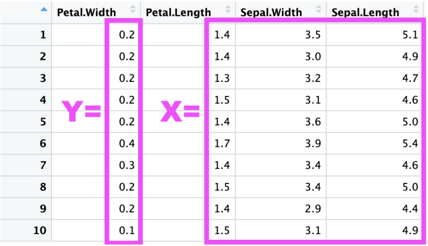
$$\begin{aligned} 2x + 4y + 6z &= 18 \\ 4x + 5y + 6z &= 24 \\ 3x + 1y - 2z &= 4 \end{aligned} \quad \rightarrow \quad \boldsymbol{A} = \begin{pmatrix} 2 & 4 & 6 \\ 4 & 5 & 6 \\ 3 & 1 & 2 \end{pmatrix} \; \boldsymbol{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \; \boldsymbol{b} = \begin{pmatrix} 18 \\ 24 \\ 4 \end{pmatrix}$$

What was Linear Algebra again? And why do we need it? III

- Vectors and matrices *can* have all sorts of mathematical objects as entries, but we (just as most statisticians and data scientists) only concern ourselves with vectors in $\mathbb{R}^p$ and matrices in $\mathbb{R}^{m \times n}$, $p, m, n \in \mathbb{N}$.

- As such, we can also think of Linear Algebra as *the math of data*:

```
View(iris[1:10,4:1])
```



	Petal.Width	Petal.Length	Sepal.Width	Sepal.Length
1	0.2	1.4	3.5	5.1
2	0.2	1.4	3.0	4.9
3	0.2	1.3	3.2	4.7
4	0.2	1.5	3.1	4.6
5	0.2	1.4	3.6	5.0
6	0.4	1.7	3.9	5.4
7	0.3	1.4	3.4	4.6
8	0.2	1.5	3.4	5.0
9	0.2	1.4	2.9	4.4
10	0.1	1.5	3.1	4.9

What was Linear Algebra again? And why do we need it? IV

- Hopefully, the iris-example reminded you of the matrix notation in (linear) regression.
- In fact, solving linear regression using OLS is one of the easiest examples for how Linear Algebra is central to many statistical tasks.
- In today's lecture, we will review some basic methods of Linear Algebra which you will need for this course.
- Throughout, the guiding principle will be: **How can we use linear transformations to make objects easier to deal with?**

Contents

1. Basic calculation rules and definitions
2. Eigenstuff
3. Decompositions
4. Educational example: Ordinary Least Squares

Reminder: Matrix multiplication

- We always multiply *rows* with *columns* **of the same length**, so for two vectors $\mathbf{x} = (x_1, x_2, x_3)^\top, \boldsymbol{y} = (y_1, y_2, y_3)^\top \in \mathbb{R}^3$

$$\boldsymbol{x}^\top \boldsymbol{y} = (x_1, x_2, x_3) \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = x_1 y_1 + x_2 y_2 + x_3 y_3 \,.$$

- And two matrices can only be multiplied **if their inner dimensions agree**, so for $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{B} \in \mathbb{R}^{n \times p}$

$$\boldsymbol{AB} = \boldsymbol{C} \,, \quad \text{where } \boldsymbol{C} \in \mathbb{R}^{m \times p} \text{ with } c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} \,.$$

Some basic definitions I

- **Transpose**: The transpose operator $\boldsymbol{A}^\top$ swaps rows and columns. If $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ then $\boldsymbol{A}^\top \in \mathbb{R}^{n \times m}$ and $(A^\top)_{ij} = A_{ji}$. Also,
    - $(\boldsymbol{A}^\top)^\top = \boldsymbol{A}$
    - $(\boldsymbol{AB})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top$.
- **Symmetric** (*For square matrices only*): A matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is symmetric, if $\boldsymbol{A} = \boldsymbol{A}^\top$.

Some basic definitions II

- **Inverse** (*For square matrices only*): Let $\boldsymbol{I}_n$ denote the $n \times n$ identity matrix. A matrix $\boldsymbol{B} \in \mathbb{R}^{n \times n}$ is invertible, if there exists a matrix $\boldsymbol{B}^{-1} \in \mathbb{R}^{n \times n}$ so that $\boldsymbol{B}\boldsymbol{B}^{-1} = \boldsymbol{B}^{-1}\boldsymbol{B} = \boldsymbol{I}_n$. Also,
  - $\boldsymbol{B}^{-1}$ is unique if it exists.
  - $(\boldsymbol{B}^{-1})^{-1} = \boldsymbol{B}$
  - $(\boldsymbol{B}\boldsymbol{A})^{-1} = \boldsymbol{A}^{-1}\boldsymbol{B}^{-1}$
  - $(\boldsymbol{B}^{-1})^{\top} = (\boldsymbol{B}^{\top})^{-1}$.

Some basic definitions III

- **Linear independence**: A set of vectors $\{\boldsymbol{v}_1, ..., \boldsymbol{v}_n\} \in \mathbb{R}^p$ is linearly independent if, for scalars $c_1, ..., c_n \in \mathbb{R}$,

$$c_1\boldsymbol{v}_1 + ... + c_n\boldsymbol{v}_n = 0 \quad \text{if and only if} \quad c_1 = ... = c_n = 0\,.$$

- **Rank:** For a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$
  - **Row rank:=** the number of linearly independent rows in $\boldsymbol{A}$.
  - **Column rank:=** the number of linearly independent columns in $\boldsymbol{A}$.

Some basic definitions IV

What is the row/column rank of the matrix

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 0 & 4 \\ 2 & 4 & 6 & 8 \end{pmatrix}?$$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 0 & 4 \\ 2 & 4 & 6 & 8 \end{pmatrix} \xrightarrow{R_3 - 2R_1} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 0 & 4 \\ 0 & 0 & 0 & 0 \end{pmatrix} \Rightarrow \text{row rank} = 2.$$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 0 & 4 \\ 2 & 4 & 6 & 8 \end{pmatrix} \xrightarrow{C_3 - 3C_1,\, C_4 - 2C_2} \begin{pmatrix} 1 & 2 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 2 & 4 & 0 & 0 \end{pmatrix} \Rightarrow \text{column rank} = 2.$$

Some basic definitions V

- Cool fact: The row rank always equals the column rank![1]
  $\Rightarrow$ we can just talk about "the rank" of a matrix $\boldsymbol{A}$ ($rank(\boldsymbol{A})$).

- Some properties:
  - $rank(\boldsymbol{A}+\boldsymbol{B}) \leq rank(\boldsymbol{A}) + rank(\boldsymbol{B})$
  - $rank(\boldsymbol{A}) = rank(\boldsymbol{A}^\top) = rank(\boldsymbol{A}\boldsymbol{A}^\top) = rank(\boldsymbol{A}^\top\boldsymbol{A})$
  - $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is invertible if and only if $rank(\boldsymbol{A}) = n$.

Some basic definitions VI

- **Diagonalizability** (*For square matrices only*): A matrix $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ is diagonalizable if $\boldsymbol{M} = \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^{-1}$ for some diagonal $\boldsymbol{D}$ and invertible $\boldsymbol{A}$
- **Definiteness** (*For square matrices only*): A matrix $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ is
  - *positive semi-definite*, if $\forall x \in \mathbb{R}^n \ : \ x^\top \boldsymbol{M} x \geq 0$ and
  - *positive definite*, if $\forall x \in \mathbb{R}^n \ : \ x^\top \boldsymbol{M} x > 0$.
  - *Negative (semi-)definiteness* is defined analogously by replacing $\geq$ and $>$ with $\leq$ and $<$, respectively.
- To diagonalize matrices, determine definiteness as well as do lots of other stuff, we can use eigenvalues and eigenvectors!

[1] See https://ocw.mit.edu/courses/18-701-algebra-i-fall-2010/dfd72d3d4a11988c2335b5e9a79ce48b_MIT18_701F10_rrk_crk.pdf for a short proof.

Contents

1 Basic calculation rules and definitions

2 Eigenstuff

3 Decompositions

4 Educational example: Ordinary Least Squares

Eigenvectors and eigenvalues

- For a square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, consider the mapping

$$\mathcal{M} : \mathbb{R}^n \longrightarrow \mathbb{R}^n, \quad \boldsymbol{x} \mapsto \boldsymbol{A}\boldsymbol{x}\,.$$

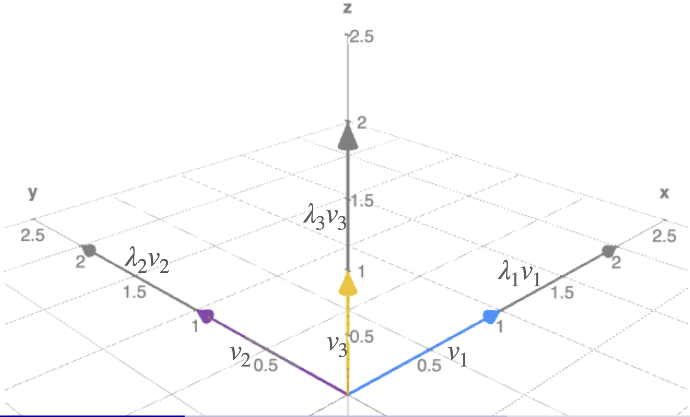
- An **eigenvector** of $\boldsymbol{A}$ is a non-zero vector $\boldsymbol{v} \in \mathbb{R}^n$ so that for some $\lambda \in \mathbb{R}$

$$\mathcal{M}(\boldsymbol{v}) = \boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}\,.$$

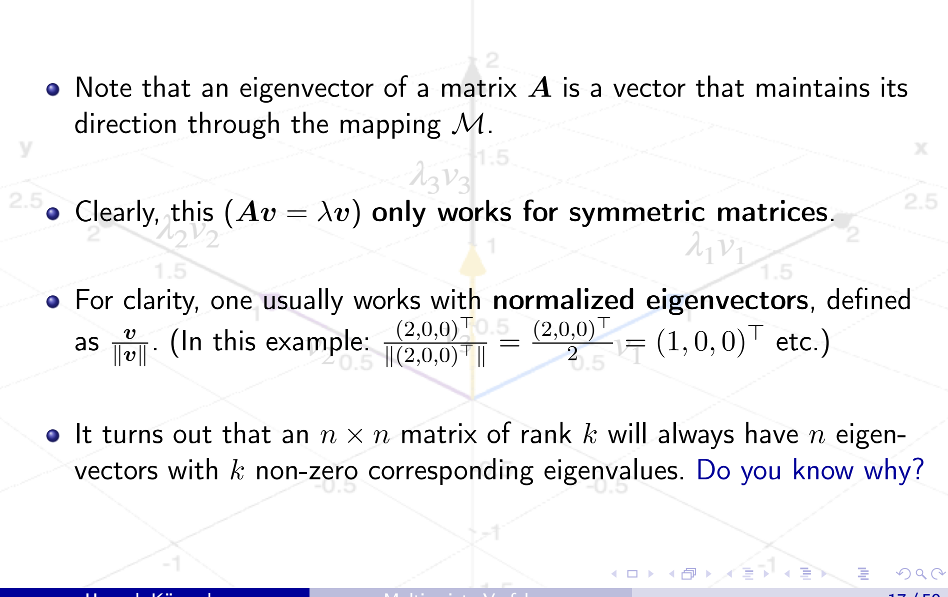
- λ is called **eigenvalue** of $\boldsymbol{A}$ (corresponding to $\boldsymbol{v}$).

Example

- For example, the matrix $\begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$ has eigenvectors $\boldsymbol{v}_1 = (1, 0, 0)$, $\boldsymbol{v}_2 = (0, 1, 0)$, $\boldsymbol{v}_3 = (0, 0, 1)$ and eigenvalues $\lambda_1 = \lambda_2 = \lambda_3 = 2$.



Some observations



- Note that an eigenvector of a matrix $\boldsymbol{A}$ is a vector that maintains its direction through the mapping $\mathcal{M}$.

- Clearly, this ($\boldsymbol{Av} = \lambda \boldsymbol{v}$) **only works for symmetric matrices**.

- For clarity, one usually works with **normalized eigenvectors**, defined as $\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}$. (In this example: $\frac{(2,0,0)^\top}{\|(2,0,0)^\top\|} = \frac{(2,0,0)^\top}{2} = (1,0,0)^\top$ etc.)

- It turns out that an $n \times n$ matrix of rank k will always have n eigenvectors with k non-zero corresponding eigenvalues. Do you know why?

Calculating eigenvalues and eigenvectors - Determinants I

- For situations that aren't as trivial as the previous example, we use **determinants** to calculate eigenvalues and eigenvectors.

- A determinant is a function of a square matrix ($\boldsymbol{A} \in \mathbb{R}^{n \times n}$), denoted by $det(\boldsymbol{A})$ or $|\boldsymbol{A}|$, from which we can get some helpful properties of $\boldsymbol{A}$.

- Determinants can be calculated using the *Leibniz formula*

$$\det(A) = \sum_{\tau \in S_n} \operatorname{sgn}(\tau) \prod_{i=1}^{n} a_{i,\,\tau(i)} = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{\sigma(i),\,i}\,,$$

where sgn is the sign function of elements of the permutation group S_n which returns $+1$ for even and -1 for odd permutations.

Calculating eigenvalues and eigenvectors - Determinants II

For 2×2 and 3×3 matrices, the Leibniz formula gives

- $\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$, and
- $\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = aei + bfg + cdh - ceg - bdi - afh$,

respectively.

How to calculate eigenvalues and eigenvectors

Given a symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{n \times n}$, we can now calculate its eigenstuff by

1. Solving $\det(\boldsymbol{M} - \lambda \boldsymbol{I}_n) \overset{!}{=} 0$ for λ to get all eigenvalues of $\boldsymbol{M}$.
2. For each $i \in \{1, ..., n\}$, determining the eigenvector corresponding to λ_i by solving $(\boldsymbol{M} - \lambda_i \boldsymbol{I}_n)\boldsymbol{x} \overset{!}{=} \boldsymbol{0}$.
   (Where $\boldsymbol{x}$ is a vector of variables which, once solved for, make up the entries of eigenvector $\boldsymbol{v}_i$.)

Example continued

In the previous example, with $\boldsymbol{M} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$, we have

- $det(\boldsymbol{M} - \lambda \boldsymbol{I}_3) \overset{!}{=} 0 \Leftrightarrow (2-\lambda)^3 \overset{!}{=} 0 \quad \Rightarrow \lambda_1 = \lambda_2 = \lambda_3 = 2$ and
- $$\left(\boldsymbol{M} - \lambda_i \boldsymbol{I}_3\right) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \overset{!}{=} \boldsymbol{0} \quad \Leftrightarrow \quad \begin{array}{rcl} 0x_1 + 0x_2 + 0x_3 & = & 0 \\ 0x_1 + 0x_2 + 0x_3 & = & 0 \\ 0x_1 + 0x_2 + 0x_3 & = & 0 \end{array}$$

  $\Rightarrow$ technically, all vectors in $\mathbb{R}^3$ except $(0,0,0)^\top$ are eigenvectors of $\boldsymbol{M}$, but if we want 3 *linearly independent, normalized eigenvectors*, the simplest solution is $\boldsymbol{v}_1 = (1,0,0)$, $\boldsymbol{v}_2 = (0,1,0)$, $\boldsymbol{v}_3 = (0,0,1)$.

Some helpful rules I

For a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ the following holds

- The eigenvalues of $\boldsymbol{A}$ are all real.
- The eigenvectors of $\boldsymbol{A}$ are **orthogonal**.
  This means that, $\forall i, j \in \{1, ..., n\}$ with $i \neq j$: $\boldsymbol{v}_i^\top \boldsymbol{v}_j = 0$.
- Furthermore, the eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{A}^\top$ are identical since $det(\boldsymbol{A} - \lambda \boldsymbol{I}) = det\big((\boldsymbol{A} - \lambda \boldsymbol{I})^\top\big) = det(\boldsymbol{A}^\top - \lambda \boldsymbol{I})$.

Definition: A matrix $\boldsymbol{Q}$ is called *orthogonal*, or *orthonormal*, if it is a real square matrix whose columns and rows are orthonormal vectors (i.e. orthogonal vectors, all of which have length 1). Then $\boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}$.

Some helpful rules II

- We can calculate the inverse of a square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ as

  $$\boldsymbol{A}^{-1} = \frac{1}{det(\boldsymbol{A})} Adj(\boldsymbol{A})$$

  Where $Adj(\boldsymbol{A})$ is the **adjugate matrix** of $\boldsymbol{A}$.

  The general definition of adjugate matrix is *transpose of the cofactor matrix*, but you will at most need to calculate $Adj(\boldsymbol{M})$ for $\boldsymbol{M} \in \mathbb{R}^{2\times 2}$, for which you the following holds

  $$Adj\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right) = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Some helpful rules III

- *For a symmetric matrix* $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, it is very easy to determine definiteness once one has calculated the eigenvalues $\lambda_1, ..., \lambda_n$
  - $\boldsymbol{A}$ is positive/negative semi-definite, iff $\forall i \in \{1, ..., n\}: \quad \lambda_i \geq 0$ / $\lambda_i \leq 0$, respectively.
  - $\boldsymbol{A}$ is positive/negative definite, iff $\forall i \in \{1, ..., n\}: \quad \lambda_i > 0$ / $\lambda_i < 0$, respectively.
  - $\boldsymbol{A}$ is indefinite when there exists both a $\lambda_i > 0$ **and** $\lambda_j < 0$ for $i, j \in \{1, ..., n\}$, $i \neq j$.
- *For a non-symmetric but still square matrix* $\boldsymbol{B} \in \mathbb{R}^{n \times n}$, definiteness may be determined by applying the above rules to the matrix $\frac{1}{2}\left(\boldsymbol{B} + \boldsymbol{B}^\top\right)$.

Contents

1 Basic calculation rules and definitions

2 Eigenstuff

3 Decompositions

4 Educational example: Ordinary Least Squares

Motivation

Matrix decomposition refers to rewriting matrices as products of other matrices. There are many different methods, some of which will be covered in this lecture. Some advantages of matrix decomposition are

- Easier computation of problems using handy matrix properties. (Especially inverses and roots)
- Increasing the informative value of matrices from a mathematical/statistical perspective.
- Avoid correlation related issues by decomposing covariance matrices.

QR decomposition

- You have probably already heard of QR decomposition.

  It is, e.g., used to solve *linear least squares* (which we will get to later) or the *QR algorithm* to determine eigenvalues/-vectors.

- In this case, a matrix $\mathbf{A}$ is decomposed into a product $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ of an orthonormal matrix $\boldsymbol{Q}$ and an upper triangular matrix $\boldsymbol{R}$.

Cholesky decomposition I

- Any symmetric and positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ may be decomposed by writing
$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top ,$$
where $\boldsymbol{L}$ is a *lower triangular matrix* (and, by transition, $\boldsymbol{L}^\top$ an upper triangular matrix).

- Generally, the Cholesky decomposition is slightly less stable than, e.g., QR decomposition, but more efficient for large n!

Cholesky decomposition II

- Specifically, we have

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top = \begin{pmatrix} l_{11} & \dots & 0 \\ \vdots & \ddots & \vdots \\ l_{n1} & \dots & l_{nn} \end{pmatrix} \begin{pmatrix} l_{11} & \dots & l_{1n} \\ \vdots & \ddots & \vdots \\ 0 & \dots & l_{nn} \end{pmatrix}$$

with, $\forall i, k \in \{1, ..., n\}, \quad i \neq k,$

  - $l_{ii} = \sqrt{a_{ii} - \sum_{j=1}^{i-1} l_{ij}^2}$
  - $l_{ki} = \dfrac{a_{ki} - \sum_{j=1}^{i-1} l_{ij} l_{kj}}{l_{ii}}$.

Cholesky decomposition - example

In the previous example, with $\boldsymbol{M} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$, we have

- $l_{11} = \sqrt{2 - \sum_{j=1}^{0} l_{ij}^2} = \sqrt{2-0} = \sqrt{2}, \quad l_{21} = l_{31} = \frac{0}{\sqrt{2}} = 0,$
- $l_{22} = \sqrt{2 - \sum_{j=1}^{1} l_{ij}^2} = \sqrt{2-0} = \sqrt{2}, \quad l_{12} = l_{32} = \frac{0}{\sqrt{2}} = 0,$ and
- $l_{33} = \sqrt{2 - \sum_{j=2}^{2} l_{ij}^2} = \sqrt{2-0} = \sqrt{2}, \quad l_{12} = l_{32} = \frac{0}{\sqrt{2}} = 0.$
- So, clearly, it follows that

$$\boldsymbol{M} = \boldsymbol{L}\boldsymbol{L}^\top .$$

Eigendecomposition (Spektralzerlegung) I

- Note that for any square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, each of the following statements implies the other[2]
  - $\boldsymbol{A}$ is diagonalizable $\Leftrightarrow$
  - $\boldsymbol{A}$ has n linearly independent eigenvectors.
- If one (i.e. both) of these statements holds for a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ we can decompose it into

$$\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{-1},$$

  where $\boldsymbol{Q}$ is the square $n \times n$ matrix whose ith column is the eigenvector $\boldsymbol{v}_i$ of $\boldsymbol{A}$, and $\boldsymbol{\Lambda}$ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues, $\boldsymbol{\Lambda}_{ii} = \lambda_i$.

Eigendecomposition (Spektralzerlegung) II

The decomposition $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$ is actually really intuitive once one considers

- $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ for any eigenvalue λ and eigenvector $\boldsymbol{v}$.
- Therefore, since the columns of $\boldsymbol{Q}$ are the eigenvectors of $\boldsymbol{A}$: $\mathbf{A}\mathbf{Q} = \mathbf{Q}\mathbf{\Lambda}$.
- And it clearly immediately follows that $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$.

[2] See `https://sharmaeklavya2.github.io/theoremdep/nodes/linear-algebra/eigenvectors/diag-linindep.html` for a short proof.

Eigendecomposition (Spektralzerlegung) - continued example

In the previous example, with $\boldsymbol{M} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$, we have

- $\boldsymbol{\Lambda} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$ and
- $\boldsymbol{Q} = \boldsymbol{Q}^{-1} = \boldsymbol{I}_3$.
- So, clearly, it follows that

$$\boldsymbol{M} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1}\,.$$

Singular value decomposition (SVD) I

- Eigendecomposition is extremely helpful in many situation, however, it only applies to diagonizable, and therefore square, matrices.

- For a non-square matrix we can use **singular value decomposition (SVD)**!

- First, note that for any not necessarily symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, the matrix $\boldsymbol{A}^\top \boldsymbol{A} \in \mathbb{R}^{n \times n}$ is symmetric because

$$(\boldsymbol{A}^\top \boldsymbol{A})^\top = \boldsymbol{A}^\top (\boldsymbol{A}^\top)^\top = \boldsymbol{A}^\top \boldsymbol{A}\,.$$

- Now, $\boldsymbol{A}^\top \boldsymbol{A}$ has n real eigenvalues since it is symmetric, and one can show that all of them are positive[3].

Singular value decomposition (SVD) II

- The same of course holds for $\boldsymbol{A}\boldsymbol{A}^\top \in \mathbb{R}^{m \times m}$, except for that it is an $m \times m$ matrix.

- **Importantly**, both $\boldsymbol{A}^\top\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^\top$ will always have *the same* $r = rank(\boldsymbol{A})$ non-zero eigenvalues.

  This follows directly from Sylvester's determinant theorem:
  For two matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}$ it holds that

$$det(\boldsymbol{I}_m + \boldsymbol{A}\boldsymbol{B}^\top) = det(\boldsymbol{I}_n + \boldsymbol{B}^\top\boldsymbol{A})$$

  (See (B.1.16) from Pozrikidis 2014[4]), because setting $\boldsymbol{B} = \boldsymbol{A}$, we can then show that for any non-zero eigenvalue of either $\boldsymbol{A}^\top\boldsymbol{A}$ or $\boldsymbol{A}\boldsymbol{A}^\top$

$$\det(\boldsymbol{A}\boldsymbol{A}^\top - \lambda\boldsymbol{I}_m) = \det(\boldsymbol{A}^\top\boldsymbol{A} - \lambda\boldsymbol{I}_n)\,.$$

Singular value decomposition (SVD) III

- For $r = rank(\boldsymbol{A})$, let $\{\lambda_1, ..., \lambda_r\}$ denote the set of non-zero eigenvalues of $\boldsymbol{A}^\top \boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^\top$, reordered so that

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r > 0\,.$$

The **non-zero sigular values** of $\boldsymbol{A}$ are then defined as

$$\sigma_i := \sqrt{\lambda_i} \quad \forall i = 1, ..., r\,.$$

- Based on this, the **singular value decomposition** of a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is then given by

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top.$$

Singular value decomposition (SVD) IV

- There are two versions of this decomposition regarding the exact form of $\boldsymbol{U}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{V}$, often referred to as **full and compact SVD**, respectively[5].

- We will focus on the compact version, which is also implemented in Base R as the function `svd()`.

[3] See `https://towardsdatascience.com/understanding-singular-value-decomposition-and-its-application-in-data-sci` for proof and further explanation

[4] **Sylvester**

[5] See `http://pfister.ee.duke.edu/courses/ecen601/notes_ch8.pdf` for the formal definitions.

Compact singular value decomposition I

- Here, a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ with $rank(\boldsymbol{A}) = r$ is decomposed into $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, with
  - $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_r \end{pmatrix}$, i.e. an $r \times r$ diagonal matrix with the singular values of $\boldsymbol{A}$ on the diagonal. (Recall that $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0\,$.)
  - $\boldsymbol{U} = \big(\boldsymbol{u}_1, \dots, \boldsymbol{u}_r\big) \in \mathbb{R}^{m \times r}$, where $\boldsymbol{u}_i$ is the eigenvector of $\boldsymbol{A}\boldsymbol{A}^\top$ corresponding to the eigenvalue σ_i^2 and
  - $\boldsymbol{V} = \big(\boldsymbol{v}_1, \dots, \boldsymbol{v}_r\big) \in \mathbb{R}^{n \times r}$, where $\boldsymbol{v}_i$ is the eigenvector of $\boldsymbol{A}^\top\boldsymbol{A}$ corresponding to the eigenvalue σ_i^2.

Compact SVD example I

- Consider the matrix $\boldsymbol{A} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \\ 0 & 0 \end{pmatrix}$.

- To perform SVD, we first calculate

$$\boldsymbol{A}^\top \boldsymbol{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{A}\boldsymbol{A}^\top = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

- Recall that to get the eigenvalues of a matrix we have to solve $\det(\boldsymbol{M} - \lambda \boldsymbol{I}_n) \overset{!}{=} 0$

  $\Rightarrow$ the non-zero eigenvalues of $\boldsymbol{A}^\top \boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^\top$ are $\lambda_1 = \lambda_2 = 1$.

Compact SVD example II

- So, by the definition of singular values σ_i, we get the following for $\boldsymbol{A}$: $\sigma_1 = \sqrt{\lambda_1} = 1$, $\sigma_2 = \sqrt{\lambda_2} = 1$.

$$\Rightarrow \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

- Furthermore, non-zero eigenvectors of $\boldsymbol{A}\boldsymbol{A}^\top$ are $\boldsymbol{u}_1 = (1,0,0)^\top$ and $\boldsymbol{u}_2 = (0,1,0)^\top$, so

$$\boldsymbol{U} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{and}$$

Compact SVD example III

- Non-zero eigenvectors of $\boldsymbol{A}\boldsymbol{A}^\top$ are $\boldsymbol{v}_1 = (1,0)^\top$ and $\boldsymbol{v}_2 = (0,-1)^\top$, so
$$\boldsymbol{V} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

- Check: $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \\ 0 & 0 \end{pmatrix}$ ✓

- For even slightly less simple matrices, compact SVD quickly becomes much more cumbersome to perform by hand. Luckily, R base contains the `svd()` function to perform singular value decomposition - see this Rpubs page for more on SVD and how to perform it in R.

Contents

Reminder: Linear regression setting

- We can write a linear regression with p independent variables (last lecture, we discussed the case $p = 1$) in matrix form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,,$$

with

- $\boldsymbol{y} = (y_1, ..., y_n)^\top$,
- $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)^\top$,
- $\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}$ and
- error terms $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^\top$, usually modelled as $\varepsilon_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$.

OLS in matrix notation I

- If our objective is simply to minimize the least squares (i.e. distance of points to regression line w.r.t. squared loss), our goal is to find the global minimum of $S(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$.

- Recall that last lecture, we considered the approach of maximizing the Likelihood, or, equivalently, the log Likelihood, which is given by:

$$\log\Big(\mathcal{L}\Big(y; \beta = (\beta_0, \beta_1)^\top\Big)\Big) = \log\left(\frac{n}{\sigma\sqrt{2\pi}} e^{\sum_{i=1}^{n} -\frac{1}{2}\left(\frac{y_i - \beta X}{\sigma}\right)^2}\right)$$

$$= n\log(\frac{1}{2}\pi\sigma^2) - \frac{1}{2}\sigma^2 \sum_{i=0}^{n}(y_i - \beta X)^2$$

Since the red term is equal to $S(\boldsymbol{\beta})$, maximizing the Likelihood is equivalent to minimizing the sum of squares in linear regression!

OLS in matrix notation II

- To solve this problem, we need to find a value $\hat{\boldsymbol{\beta}}$ so that the derivative is equal to zero:

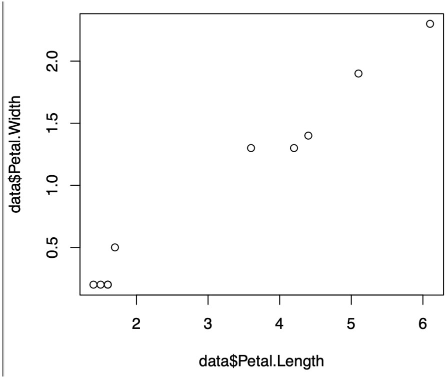
$$\begin{aligned} 0 \overset{!}{=} \frac{dS}{d\boldsymbol{\beta}}(\boldsymbol{\beta}) &= \frac{d}{d\boldsymbol{\beta}}\left(\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}\right) \\ &= -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}\,. \end{aligned}$$

- Given that $\boldsymbol{X}$ has full column rank, and therefore $\boldsymbol{X}^T\boldsymbol{X}$ is invertible, this is solved by

$$\mathbb{R}^{p+1} \ni (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)^\top = \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}\,.$$

Data: Subset of iris

```
set.seed(735)
data<-iris[sample(1:nrow(iris),10),
c("Petal.Width","Petal.Length")]
plot(data$Petal.Length,data$Petal.Width)
```



Example: matrix decomposition in linear regression I

- So we want to use a linear model with `Petal.Width` as dependent and `Petal.Length` as independent variable.
- This is usually achieved via

```
lm(Petal.Width~Petal.Length,data=data)
```

- But we can simply calculate the same coefficients as follows:

```
y<-data$Petal.Width
X<-matrix(c(rep(1,nrow(data)),data$Petal.Length)
,nrow=nrow(data),ncol=2)
solve(t(X)%*%X)%*%t(X)%*%y
```

Example: matrix decomposition in linear regression II

- Notice that, to get the matrix $(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \hat{=} \boldsymbol{A}$, we need to use the `solve()` function to solve the problem

  $$(\boldsymbol{X}^\top \boldsymbol{X})\boldsymbol{A} = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix}$$

  for $\boldsymbol{A}$.

- This is clearly not a problem for simple and small cases of design matrices $\boldsymbol{X}$ but it can quickly get complicated and computationally expensive - and that is one example where matrix decomposition comes into play!

Example: matrix decomposition in linear regression III

Let's look at this problem using QR-decomposition.

- Having decomposed $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$ it follows that, if $\boldsymbol{R}$ is a square matrix,

$$\boldsymbol{X}^\top \boldsymbol{X} = (\boldsymbol{Q}\boldsymbol{R})^\top \boldsymbol{Q}\boldsymbol{R} = \boldsymbol{R}^\top \boldsymbol{Q}^\top \boldsymbol{Q}\boldsymbol{R} \overset{\boldsymbol{Q} \text{ is orthogonal}}{=} \boldsymbol{R}^\top \boldsymbol{R}$$

- Therefore,

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = (\boldsymbol{R}^\top \boldsymbol{R})^{-1} (\boldsymbol{Q}\boldsymbol{R})^\top \boldsymbol{y} \\ &= \boldsymbol{R}^{-1} (\boldsymbol{R}^\top)^{-1} \boldsymbol{R}^\top \boldsymbol{Q}^\top \boldsymbol{y} \\ &= \boldsymbol{R}^{-1} \boldsymbol{Q}^\top \boldsymbol{y}\,,\end{aligned}$$

  which is much easier to compute!!

Example: matrix decomposition in linear regression IV

Let's look at this problem using SVD.

- Having decomposed $\boldsymbol{X} = \boldsymbol{U\Sigma V}^\top$ it directly follows that, if $\boldsymbol{V}$ is a square matrix,

$$\begin{aligned}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = (\boldsymbol{V\Sigma}^T\boldsymbol{U}^T\boldsymbol{U\Sigma V}^T)^{-1}\boldsymbol{V\Sigma}^T\boldsymbol{U}^T \\ =&(\boldsymbol{V\Sigma}^T\boldsymbol{\Sigma V}^T)^{-1}\boldsymbol{V\Sigma}^T\boldsymbol{U}^T = (\boldsymbol{V}^T)^{-1}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma})^{-1}\boldsymbol{V}^{-1}\boldsymbol{V\Sigma}^T\boldsymbol{U}^T \\ =&\boldsymbol{V}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}^T\boldsymbol{U}^T = \boldsymbol{V\Sigma}^{-1}\boldsymbol{U}^T\end{aligned}$$

- Therefore,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} = \boldsymbol{V\Sigma}^{-1}\boldsymbol{U}^T\boldsymbol{y}\,,$$

which is again much easier to compute!!