Multivariate Statistics

Lecture 01

Fudan University

Outline

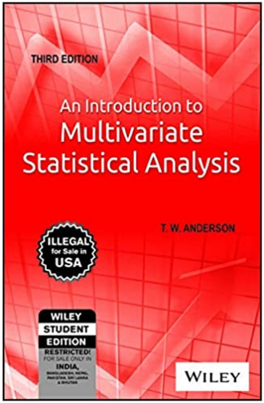
Outline

Course Overview

Homepage: `https://luoluo-sds.github.io/`

Prerequisite course: calculus, linear algebra, probability and statistics

Recommended reading (textbook):



Course Grading Policy

Homework, 40%

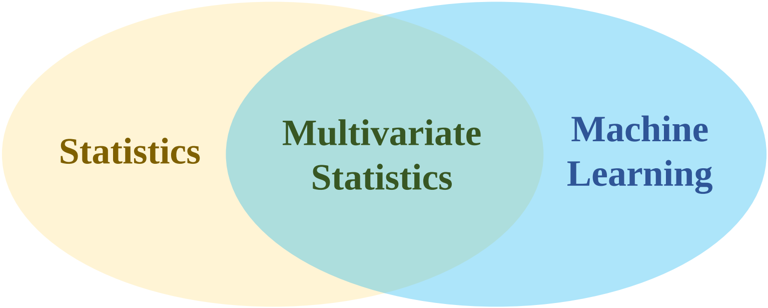
No midterm exam

Final exam, 60%

Multivariate Statistics

Multivariate statistics is a subdivision of statistics encompassing the simultaneous observation and analysis of more than one variable.

Multivariate statistics try to understand the relationships between variables and their relevance to the problem being studied.



Multivariate Statistics

The measurements made on a single individual can be assembled into a column vector.

The set of observations on all individuals in a sample constitutes a sample of vectors, and the vectors set side by side make up the matrix of observations.

The data to be analyzed then are thought of as displayed in a matrix or in several matrices.

We start from the review of linear algebra.

Outline

Notations

We use x_i to denote the entry of the n-dimensional vector $\mathbf{x}$ such that

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n.$$

We use a_{ij} to denote the entry of matrix $\mathbf{A}$ with dimension $m \times n$ such that

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \in \mathbb{R}^{m\times n}.$$

Notations

We can also present the matrix as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1q} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \cdots & \mathbf{A}_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{p1} & \mathbf{A}_{p2} & \cdots & \mathbf{A}_{pq} \end{bmatrix} \in \mathbb{R}^{m \times n}.$$

if the sub-matrices are compatible with the partition.

We define

$$\mathbf{0} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{m \times n}, \quad \mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Matrix Operations: Transpose

The transpose of a matrix results from flipping the rows and columns. Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ such that

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n},$$

then its transpose, written $\mathbf{A}^\top \in \mathbb{R}^{n \times m}$, is an $n \times m$ matrix such that

$$\mathbf{A}^\top = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix} \in \mathbb{R}^{n \times m}.$$

Matrix Operations: Transpose

The following properties of transposes are easily verified

1. $\left(\mathbf{A}^\top\right)^\top = \mathbf{A}$
2. $(c_1\mathbf{A} + c_2\mathbf{B})^\top = c_1\mathbf{A}^\top + c_2\mathbf{A}^\top$

We say a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is symmetric if $\mathbf{A} = \mathbf{A}^\top$. It is common to denote the set of all symmetric matrices of size n as $\mathbb{S}^n$.

We say a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is anti-symmetric if $\mathbf{A} = -\mathbf{A}^\top$.

Matrix Operations: Transpose

Sometimes (not always), we also use $\mathbf{A}'$ the present the transpose of $\mathbf{A}$.

In MATLAB, the notation $\mathbf{A}'$ present the conjugate transpose of $\mathbf{A}$. To avoid ambiguity, we use the superscript H to denote conjugate transpose. Given $\mathbf{A} \in \mathbb{C}^{m \times n}$, we define

$$\mathbf{A}^H = \begin{bmatrix} \bar{a}_{11} & \bar{a}_{21} & \cdots & \bar{a}_{m1} \\ \bar{a}_{12} & \bar{a}_{22} & \cdots & \bar{a}_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{a}_{1n} & \bar{a}_{2n} & \cdots & \bar{a}_{mn} \end{bmatrix} \in \mathbb{C}^{n \times m}.$$

For example,

$$\begin{bmatrix} 2+3\mathrm{i} & 1 & 2-\mathrm{i} \\ 5-6\mathrm{i} & i & 3-2\mathrm{i} \end{bmatrix}^H = \begin{bmatrix} 2-3\mathrm{i} & 5+6\mathrm{i} \\ 1 & -\mathrm{i} \\ 2+\mathrm{i} & 3+2\mathrm{i} \end{bmatrix}$$

Matrix Operations: Addition/Subtraction

If $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{m\times n}$ are two matrices of the same order, then

$$\mathbf{A}+\mathbf{B} = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \cdots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & \cdots & a_{2n}+b_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \cdots & a_{mn}+b_{mn} \end{bmatrix} \in \mathbb{R}^{m\times n}$$

and

$$\mathbf{A}-\mathbf{B} = \begin{bmatrix} a_{11}-b_{11} & a_{12}-b_{12} & \cdots & a_{1n}-b_{1n} \\ a_{21}-b_{21} & a_{22}-b_{22} & \cdots & a_{2n}-b_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}-b_{m1} & a_{m2}-b_{m2} & \cdots & a_{mn}-b_{mn} \end{bmatrix} \in \mathbb{R}^{m\times n}.$$

Matrix Operations: Multiplication

The product of $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times p}$ is the matrix

$$\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{m \times p},$$

where

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1q} \\ c_{21} & c_{22} & \cdots & c_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pq} \end{bmatrix} \in \mathbb{R}^{m \times p}.$$

and $c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$.

Matrix Operations: Inner Product (Dot Product)

Given two vectors $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$, the quantity $\mathbf{x}^\top \mathbf{y} \in \mathbb{R}$ is called the inner product (or dot product) of the vectors, is a real number given by

$$\mathbf{x}^\top \mathbf{y} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i.$$

Matrix Operations: Outer Product

Given two vectors $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$, the matrix $\mathbf{x}\mathbf{y}^\top \in \mathbb{R}^{m\times n}$ is called the outer product of the vectors, that is,

$$\mathbf{x}\mathbf{y}^\top = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & \cdots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_my_1 & x_my_2 & \cdots & x_my_n \end{bmatrix} = \begin{bmatrix} x_1\mathbf{y}^\top \\ x_2\mathbf{y}^\top \\ \vdots \\ x_m\mathbf{y}^\top \end{bmatrix} \in \mathbb{R}^{m\times n}.$$

Example: Let $\mathbf{x} = [x_1, x_2, \ldots, x_m]^\top \in \mathbb{R}^m$ and $\mathbf{1} = [1, 1, \ldots, 1]^\top \in \mathbb{R}^n$, then

$$\begin{bmatrix} | & | & & | \\ \mathbf{x} & \mathbf{x} & \cdots & \mathbf{x} \\ | & | & & | \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} = \mathbf{x}\mathbf{1}^\top \in \mathbb{R}^{m\times n}.$$

Matrix Operations: Linear Combination

Given $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{x} \in \mathbb{R}^n$, the product $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{R}^m$ can be viewed as the linear combination of the columns of $\mathbf{A}$:

$$\mathbf{y} = \mathbf{A}\mathbf{x} = \begin{bmatrix} | & | & & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} | \\ \mathbf{a}_1 \\ | \end{bmatrix} x_1 + \begin{bmatrix} | \\ \mathbf{a}_2 \\ | \end{bmatrix} x_2 \cdots + \begin{bmatrix} | \\ \mathbf{a}_n \\ | \end{bmatrix} x_n.$$

Matrix Operations: Multiplication

Properties of matrix multiplication

1. Matrix multiplication is associative: $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
2. Matrix multiplication is distributive: $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$
3. Matrix multiplication is NOT commutative in general.
4. $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$

Matrix Operations: Trace

The trace of a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, denoted $\mathrm{tr}(\mathbf{A})$, is the sum of diagonal elements in the matrix:

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}.$$

The trace has the following properties

1. For $\mathbf{A} \in \mathbb{R}^{n\times n}$, we have $\mathrm{tr}(\mathbf{A}) = \mathrm{tr}\left(\mathbf{A}^\top\right)$.
2. For $\mathbf{A} \in \mathbb{R}^{n\times n}$, we have $\mathrm{tr}\left(\mathbf{A}^\top\mathbf{A}\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2$.
3. For $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times n}$, $c_1 \in \mathbb{R}$ and $c_2 \in \mathbb{R}$, we have $\mathrm{tr}(c_1\mathbf{A} + c_2\mathbf{B}) = c_1\mathrm{tr}(\mathbf{A}) + c_2\mathrm{tr}(\mathbf{B})$.
4. For $\mathbf{A}$ and $\mathbf{B}$ such that $\mathbf{AB}$ is square, $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$.
5. For $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ such that $\mathbf{ABC}$ is square, we have $\mathrm{tr}(\mathbf{ABC}) = \mathrm{tr}(\mathbf{BCA}) = \mathrm{tr}(\mathbf{CAB})$.

Matrix Operations: Kronecker Product

The Kronecker product, denoted by $\otimes$, is an operation on two matrices of arbitrary size resulting in a block matrix. It is a generalization of the outer product from vectors to matrices.

If $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{B}$ is a $p \times q$ matrix, then the Kronecker product $\mathbf{A} \otimes \mathbf{B}$ is the $mp \times nq$ matrix as follows

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix} \in \mathbb{R}^{mn \times pq}.$$

The following properties of transposes are easily verified

1. $\mathbf{A} \otimes (\mathbf{B} + \mathbf{C}) = \mathbf{A} \otimes \mathbf{B} + \mathbf{A} \otimes \mathbf{C}$
2. $(\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C})$
3. $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD})$
4. $\mathrm{tr}(\mathbf{A} \otimes \mathbf{B}) = \mathrm{tr}(\mathbf{A})\mathrm{tr}(\mathbf{B})$

Inverse

The inverse of a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is denoted by $\mathbf{A}^{-1}$ and is the unique matrix such that

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I} = \mathbf{A}^{-1}\mathbf{A}.$$

Note that not all matrices have inverses. Particular, we say that $\mathbf{A}$ is invertible or non-singular if $\mathbf{A}^{-1}$ exists and non-invertible or singular otherwise.

Inverse

If all the necessary inverse exist, we have

1. $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
2. $(c\mathbf{A})^{-1} = c^{-1}\mathbf{A}^{-1}$
3. $(\mathbf{A}^{-1})^{\top} = (\mathbf{A}^{\top})^{-1}$
4. $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
5. $\mathbf{A}^{-1} = \mathbf{A}^{\top}$ if $\mathbf{A}^{\top}\mathbf{A} = \mathbf{I}$

For $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times p}$, $\mathbf{C} \in \mathbb{R}^{p\times p}$ and $\mathbf{D} \in \mathbb{R}^{p\times n}$, we have

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}.$$

if $\mathbf{A}$ and $\mathbf{A} + \mathbf{BCD}$ are non-singular.

Outline

Vector Norms

A norm of a vector $\mathbf{x} \in \mathbb{R}^n$ written by $\|\mathbf{x}\|$, is informally a measure of the length of the vector. For example, we have the commonly-used Euclidean norm (or ℓ_2 norm),

$$\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^\top \mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Formally, a norm is any function $\mathbb{R}^n \to \mathbb{R}$ that satisfies four properties:

1. For all $\mathbf{x} \in \mathbb{R}^n$, we have $\|\mathbf{x}\| \geq 0$ (non-negativity).
2. $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$ (definiteness).
3. For all $\mathbf{x} \in \mathbb{R}^n$ and $t \in \mathbb{R}$, we have $\|t\mathbf{x}\| = |t| \, \|\mathbf{x}\|$ (homogeneity).
4. For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (triangle inequality).

Vector Norms

There are some examples for $\mathbf{x} \in \mathbb{R}$:

1. The ℓ_1 norm is $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$
2. The ℓ_2 norm is $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$
3. The ℓ_∞ norm is $\|\mathbf{x}\|_2 = \max_i |x_i|$
4. The ℓ_p norm is $\|\mathbf{x}\|_2 = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$ for $p > 1$.

Vector Space

If $\mathcal{W}$ is a subset of $\mathbb{R}^m$ such that for all $\mathbf{x}, \mathbf{y} \in \mathcal{W}$ and $a \in \mathbb{R}$ have $a(\mathbf{x} + \mathbf{y}) \in \mathcal{W}$, then $\mathcal{W}$ is called a vector subspace of $\mathbb{R}^m$. Two simple examples of subspace of $\mathbb{R}^m$ are $\{\mathbf{0}\}$ and $\mathbb{R}^m$ itself.

A set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subseteq \mathbb{R}^m$ is said to be linearly independent if no vector can be represented as a linear combination of the remaining vectors. Conversely, if one vector belonging to the set can be represented as a linear combination of the remaining vectors, then the vectors are said to be linear dependent. That is, if

$$\mathbf{x}_n = \sum_{i=1}^{n-1} \alpha_i \mathbf{x}_i$$

for some scalar values $\alpha_1, \ldots, \alpha_{n-1}$, then we say $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are linearly dependent; otherwise the vectors are linearly independent.

Orthogonality

1. Two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ are orthogonal if $\mathbf{x}^\top \mathbf{y} = 0$.
2. A vector $\mathbf{x} \in \mathbb{R}^n$ is normalized if $\|\mathbf{x}\|_2 = 1$.
3. A square matrix $\mathbf{U} \in \mathbb{R}^{n \times n}$ is orthogonal if all its columns are orthogonal to each other and are normalized (the columns are then referred to as being orthonormal). In other word, we have

$$\mathbf{U}^\top \mathbf{U} = \mathbf{I} = \mathbf{U}\mathbf{U}^\top.$$

   Note that if $\mathbf{U}$ is not square, i.e., $\mathbf{U} \in \mathbb{R}^{m \times n}$, $n < m$, but its columns are still orthonormal, then $\mathbf{U}^\top \mathbf{U} = \mathbf{I}$, but $\mathbf{U}\mathbf{U}^\top \neq \mathbf{I}$, we call that $\mathbf{U}$ is column orthonormal.

Orthogonality

A nice property of orthogonal matrices is that operating on a vector with an orthogonal matrix will not change its Euclidean norm, that is

$$\|\mathbf{U}\mathbf{x}\|_2 = \|\mathbf{x}\|_2$$

for any $\mathbf{x} \in \mathbb{R}^n$ and orthogonal $\mathbf{U} \in \mathbb{R}^{n\times n}$.

Orthogonal matrices can be used to represent a rotation.

A basis $\mathbf{x}_1, \ldots, \mathbf{x}_k$ of a subspace $\mathcal{W}$ of $\mathbb{R}^n$ is called orthonormal basis if all the elements have norm one and are orthogonal to one another.

In particular, if $\mathbf{A} \in \mathbb{R}^n$ is an orthogonal matrix then the columns of $\mathbf{A}$ form an orthogonal basis of $\mathbb{R}^n$.

Rank

The column rank of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the size of the largest subset of columns of $\mathbf{A}$ that constitute a linearly independent set.

In the same way, the row rank is the largest number of rows of A that constitute a linearly independent set.

For any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, the column rank of $\mathbf{A}$ is equal to the row rank of $\mathbf{A}$.

Rank

The following are some basic properties of the rank:

1. $\text{rank}(\mathbf{A}) \leq \min(m, n)$
2. $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^\top)$
3. $\text{rank}(\mathbf{AB}) \leq \min(\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B}))$
4. $\text{rank}(\mathbf{A} + \mathbf{B}) \leq \text{rank}(\mathbf{A}) + \text{rank}(\mathbf{B})$.

If $\text{rank}(\mathbf{A}) = \min(m, n)$, then $\mathbf{A}$ is said to be full rank.

Range and Nullspace

The span of a set of vectors $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ is the set of all vectors that can be expressed as a linear combination of $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$. That is,

$$\text{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_n\} = \left\{ \mathbf{v} : \mathbf{v} = \sum_{i=1}^{n} \beta_i \mathbf{a}_i, \beta_i \in \mathbb{R} \right\}$$

The range (also called the column space) of matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ denote $\mathcal{R}(\mathbf{A})$. In other words,

$$\mathcal{R}(\mathbf{A}) = \{\mathbf{v} : \mathbf{v} = \mathbf{A}\mathbf{x}, \mathbf{x} \in \mathbb{R}^n\} \subseteq \mathbb{R}^m$$

The nullspace of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, denoted $\mathcal{N}(\mathbf{A})$ is the set of all vectors that equal $\mathbf{0}$ when multiplied by $\mathbf{A}$. In other words,

$$\mathcal{N}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = \mathbf{0}\} \subseteq \mathbb{R}^n$$

Range and Nullspace

The subspace $\mathcal{R}(\mathbf{A}^\top)$ is the orthogonal complement of $\mathcal{N}(\mathbf{A})$, that is,

$$\left\{\mathbf{w} : \mathbf{w} = \mathbf{u} + \mathbf{v}, \mathbf{u} \in \mathcal{R}(\mathbf{A}^\top), \mathbf{v} \in \mathcal{N}(\mathbf{A})\right\}$$

and

$$\mathcal{R}(\mathbf{A}^\top) \bigcap \mathcal{N}(\mathbf{A}) = \{\mathbf{0}\}.$$

The first one also can be written as $\mathcal{R}(\mathbf{A}^\top) \bigcup \mathcal{N}(\mathbf{A}) = \mathbb{R}^n$.

QR Factorization

Given a full rank matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, we can construct the column orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{m \times n}$ and upper triangular matrix $\mathbf{R} \in \mathbb{R}^{n \times n}$ such that

$$\mathbf{A} = \mathbf{QR},$$

which also can be written as

$$\begin{bmatrix} | & | & & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \mathbf{q}_1 & \mathbf{q}_2 & \cdots & \mathbf{q}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & r_{nn} \end{bmatrix}$$

QR Factorization

Each $\mathbf{a}_i$ can be presented by a linear combination of $\{\mathbf{q}_1, \cdots, \mathbf{q}_n\}$

$$
\begin{aligned}
\mathbf{a}_1 =& r_{11}\mathbf{q}_1 \\
\mathbf{a}_2 =& r_{12}\mathbf{q}_1 + r_{22}\mathbf{q}_2 \\
&\vdots \\
\mathbf{a}_n =& r_{1n}\mathbf{q}_1 + r_{2n}\mathbf{q}_2 + \cdots + \cdots + r_{nn}\mathbf{q}_n
\end{aligned}
$$

QR Factorization

There is an old idea, known as Gram-Schmidt orthogonalization, which constructs $\mathbf{Q}$ and $\mathbf{R}$ as follows

$$
\begin{aligned}
\mathbf{q}_1 &= \frac{\mathbf{a}_1}{r_{11}} \\
\mathbf{q}_2 &= \frac{\mathbf{a}_2 - r_{12}\mathbf{q}_1}{r_{22}} \\
&\vdots \\
\mathbf{q}_n &= \frac{\mathbf{a}_n - \sum_{i=1}^{n-1} r_{in}\mathbf{q}_i}{r_{nn}},
\end{aligned}
$$

where $r_{ij} = \mathbf{q}_i^\top \mathbf{a}_j$ for any $i \neq j$ and $r_{jj} = \left\| \mathbf{a}_j - \sum_{i=1}^{j-1} r_{ij}\mathbf{q}_i \right\|_2$.

Then we have $\|\mathbf{q}_i\|_2 = 1$ for all $i = 1, \ldots, n$ and $\mathbf{q}_i^\top \mathbf{q}_j = 0$ for all $i \neq j$.

What about $\mathbf{A}$ is not full rank?

Outline

Matrix Norms

Given vector norm $\|\cdot\|$, the corresponding induced matrix norm of $\mathbf{A} \in \mathbb{R}^{m\times n}$ is defined as

$$\|\mathbf{A}\| = \sup_{\mathbf{x}\in\mathbb{R}^n, \mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} = \sup_{\mathbf{x}\in\mathbb{R}^n, \|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|.$$

For example, we define

$$\|\mathbf{A}\|_1 = \sup_{\mathbf{x}\in\mathbb{R}^n, \|\mathbf{x}\|_1=1} \|\mathbf{A}\mathbf{x}\|_1$$

and

$$\|\mathbf{A}\|_\infty = \sup_{\mathbf{x}\in\mathbb{R}^n, \|\mathbf{x}\|_\infty=1} \|\mathbf{A}\mathbf{x}\|_\infty.$$

Matrix Norms

We denote $\mathbf{A} \in \mathbb{R}^{m \times n}$ as

$$\mathbf{A} = \begin{bmatrix} | & | & & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} \mathbf{a}_{(1)}^\top \\ \mathbf{a}_{(2)}^\top \\ \vdots \\ \mathbf{a}_{(m)}^\top \end{bmatrix},$$

then we have

$$\|\mathbf{A}\|_1 = \max_{1 \leq j \leq n} \|\mathbf{a}_j\|_1 \quad \text{and} \quad \|\mathbf{A}\|_\infty = \max_{1 \leq i \leq m} \left\|\mathbf{a}_{(i)}\right\|_1$$

Matrix Norms

General matrix norm norm is any function $\mathbb{R}^{m\times n} \to \mathbb{R}$ that satisfies

1. For all $\mathbf{A} \in \mathbb{R}^{m\times n}$, we have $\|\mathbf{A}\| \geq 0$ (non-negativity).
2. $\|\mathbf{A}\| = 0$ if and only if $\mathbf{A} = \mathbf{0}$ (definiteness).
3. For all $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $t \in \mathbb{R}$, we have $\|t\mathbf{A}\| = |t| \, \|\mathbf{A}\|$ (homogeneity).
4. For all $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$, we have $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$ (triangle inequality).

Matrix Norms

The most important matrix norm which is not induced by a vector norm (why not?) is Frobenius norm (F-norm), which is defined as

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2} = \sqrt{\mathrm{tr}(\mathbf{A}^\top \mathbf{A})}$$

for all $\mathbf{A} \in \mathbb{R}^{m\times n}$. If $\mathbf{Q} \in \mathbb{R}^{m\times m}$ is orthogonal matrix, we have

$$\|\mathbf{QA}\|_F = \|\mathbf{A}\|_F .$$

Homework

For $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$, prove that

$$\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F \|\mathbf{B}\|_F .$$

Determinant

The determinant of a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, is denoted by $\det(\mathbf{A})$ or $|\mathbf{A}|$, which is defined as

$$\det(\mathbf{A}) = \sum_{\tau=(\tau_1,\dots,\tau_n)} \left(\operatorname{sgn}(\tau) \prod_{i=1}^{n} \mathbf{a}_{i,\tau_i} \right)$$

where $\tau = (\tau_1, \dots, \tau_n)$ is permutation of $(1, 2, \dots, n)$. The signature $\operatorname{sgn}(\tau)$ is defined to be $+1$ whenever the reordering given by τ can be achieved by successively interchanging two entries an even number of times, and -1 whenever it can be achieved by an odd number of such interchanges.

Determinant

We can also define determinant recursively

$$\begin{aligned}\det(\mathbf{A}) &= \sum_{i=1}^{n}(-1)^{i+j}a_{ij}\det(\mathbf{A}_{\setminus i,\setminus j}) \quad \text{for any } j \in \{1,\dots,n\} \\ &= \sum_{j=1}^{n}(-1)^{i+j}a_{ij}\det(\mathbf{A}_{\setminus i,\setminus j}) \quad \text{for any } i \in \{1,\dots,n\}\end{aligned}$$

with the initial condition $\det(\mathbf{A}) = a_{11}$ for $\mathbf{A}^{1\times 1}$, where $\mathbf{A}_{\setminus i,\setminus j}$ is the $(n-1)\times(n-1)$ matrix obtained by deleting the i-th row and j-th column from $\mathbf{A}$.

The adjugate of $\mathbf{A}$ is denoted by $\mathrm{adj}(\mathbf{A}) \in \mathbb{R}^{n\times n}$ whose entry at i-th row and j-th column is $(-1)^{i+j}\mathbf{A}_{\setminus j,\setminus i}$. The definition directly implies

$$\mathbf{A}\mathrm{adj}(\mathbf{A}) = \mathrm{adj}(\mathbf{A})\mathbf{A} = \det(\mathbf{A})\mathbf{I}.$$

Determinant

Given square matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ as

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_{(1)}^\top \\ \mathbf{a}_{(2)}^\top \\ \vdots \\ \mathbf{a}_{(m)}^\top \end{bmatrix},$$

the determinant of $\mathbf{A}$ is the "volume" of the set

$$\mathcal{S} = \left\{ \mathbf{v} \in \mathbb{R}^n : \mathbf{v} = \sum_{i=1}^{n} \beta_i \mathbf{a}_{(i)}, \text{where } 0 \leq \beta_j \leq 1, i = 1, \dots, n \right\}.$$

The set $\mathcal{S}$ formed by taking all possible linear combinations of the row vectors, where the coefficients of the linear combination are all between 0 and 1.

Determinant

1. $\det(\mathbf{I}) = 1$
2. If we multiply a single row in $\mathbf{A}$ by a scalar $t \in \mathbb{R}^n$, then the determinant of the new matrix is $t \det(\mathbf{A})$.
3. If we exchange any two rows of the square matrix $\mathbf{A}$, then the determinant of the new matrix is $-\det(\mathbf{A})$.
4. For $\mathbf{A} \in \mathbb{R}^{n \times n}$, we have $\det(\mathbf{A}) = 0$ if and only if $\mathbf{A}$ is singular.

Determinant

1. For $\mathbf{A} \in \mathbb{R}^{n \times n}$ is triangular, then $\det(\mathbf{A}) = \prod_{i=1}^{n} a_{ii}$.
2. For $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{p \times p}$ and $\mathbf{C} \in \mathbb{R}^{n \times p}$, we have
$$\det\left(\begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{0} & \mathbf{B} \end{bmatrix}\right) = \det(\mathbf{A})\det(\mathbf{B})$$
3. For $\mathbf{A} \in \mathbb{R}^{n \times n}$, we have $\det(\mathbf{A}) = \det(\mathbf{A}^\top)$.
4. For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$, we have $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$.
5. For $\mathbf{A} \in \mathbb{R}^{n \times n}$ is orthogonal, we have $\det(\mathbf{A}) = 1$.

Outline

Eigenvalues and Eigenvectors

Given a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, we say that $\lambda \in \mathbb{C}$ is an eigenvalue of $\mathbf{A}$ and $\mathbf{x} \in \mathbb{C}^n$ is the corresponding eigenvector is the corresponding if

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \quad \text{and} \quad \mathbf{x} \neq \mathbf{0}.$$

We define the standardized eigenvector which are normalized to have length 1. Sometimes we also use the word "eigenvector" to refer the standardized eigenvector.

We can define characteristic polynomial as

$$p_{\mathbf{A}}(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = \prod_{i=1}^{n}(\lambda_i - \lambda).$$

We can find the n roots (possibly complex) of $p_{\mathbf{A}}$ to obtain the eigenvalues $\lambda_1, \ldots, \lambda_n$.

Homework

Suppose $\mathbf{A} \in \mathbb{R}^{n \times n}$ has eigenvalues $\lambda_1, \ldots, \lambda_n$. Prove the following statements

1. $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i$,
2. $\det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i$.

Spectral Decomposition Theorem

Any symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ can be written as

$$\mathbf{A} = \mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^\top = \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i^\top$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix elements of its main diagonal are $\lambda_1, \dots, \lambda_n$ and $\mathbf{X}$ is an orthogonal matrix whose columns are corresponding to standardized eigenvectors of $\mathbf{A}$.

Proof Sketch

1. The eigenvalues of $\mathbf{A}$ are real.
2. Two eigenvectors corresponding to distinct eigenvalues of $\mathbf{A}$ are orthogonal.
3. If λ_i is an eigenvector of $\mathbf{A}$ with $m \geq 2$ algebra multiplicity, we can find m orthogonal eigenvectors in its eigenspace.

Outline

Singular Value Decomposition

The singular value decomposition (SVD) of $\mathbf{A} \in \mathbb{R}^{m \times n}$ matrix is

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

where $\mathbf{U} \in \mathbb{R}^{m \times m}$ is orthogonal, $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times n}$ is rectangular diagonal matrix with non-negative real numbers on the diagonal and $\mathbf{V} \in \mathbb{R}^{n \times n}$ is orthogonal.

The diagonal entries of $\boldsymbol{\Sigma}$ are uniquely determined by $\mathbf{A}$ and are known as the singular values of $\mathbf{A}$. The number of non-zero singular values is equal to the rank of $\mathbf{A}$. The columns of $\mathbf{U}$ and the columns of $\mathbf{V}$ are called left-singular vectors and right-singular vectors of $\mathbf{A}$, respectively.

Singular Value Decomposition

We can also write SVD as

$$\mathbf{A} = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$

where $r \leq \min\{m, n\}$ is the rank of $\mathbf{A}$ and σ_i is the diagonal entries of $\mathbf{\Sigma}$.

Homework

The SVD always exists for any $\mathbf{A} \in \mathbb{R}^{m \times n}$.
(Hint: Using spectral decomposition theorem)

The SVD is not unique. It is always possible to choose the decomposition so that the singular values σ_i are in descending order. In this case, $\mathbf{\Sigma}$ (but not always $\mathbf{U}$ and $\mathbf{V}$) is uniquely determined by $\mathbf{A}$.

The term sometimes refers to the compact SVD, a similar decomposition

$$\mathbf{A} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^\top$$

in which $\mathbf{\Sigma}_r$ is square diagonal of size $r \times r$, where $r \leq \min\{m, n\}$ is the rank of $\mathbf{A}$, and has only the non-zero singular values. In this variant, $\mathbf{U}_r$ is an $m \times r$ column orthogonal matrix and $\mathbf{V}_r$ is an $n \times r$ column orthogonal matrix such that $\mathbf{U}_r^\top \mathbf{U}_r = \mathbf{V}_r^\top \mathbf{V}_r = \mathbf{I}$.

Singular Value Decomposition

Based on SVD, we have

1. $\mathcal{R}(\mathbf{A}) = \text{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$
2. $\mathcal{N}(\mathbf{A}) = \text{span}\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$
3. $\mathcal{R}(\mathbf{A}^\top) = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$
4. $\mathcal{N}(\mathbf{A}^\top) = \text{span}\{\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m\}$

Let $\sigma_1 \leq \cdots \leq \sigma_r$ be the non-zero singular values of $\mathbf{A}$. We have $\|\mathbf{A}\|_2 = \sigma_1$ and $\|\mathbf{A}\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}$.

Outline

Quadratic Forms

Given a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and a vector $\mathbf{x} \in \mathbb{R}^n$, the scalar $\mathbf{x}^\top \mathbf{A}\mathbf{x}$ is called a quadratic form and we have

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j.$$

We often implicitly assume that the matrices appearing in a quadratic form are symmetric.

Definiteness

1. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is positive definite (PD) if for all non-zero vectors $\mathbf{x} \in \mathbb{R}^n$ holds that $\mathbf{x}^\top \mathbf{A} \mathbf{x} > 0$. This is usually denoted by $\mathbf{A} \succ \mathbf{0}$.
2. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is positive semi-definite (PSD) if for all vectors $\mathbf{x} \in \mathbb{R}^n$ holds that $\mathbf{x}^\top \mathbf{A} \mathbf{x} \geq 0$. This is usually denoted by $\mathbf{A} \succeq \mathbf{0}$.
3. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is negative definite (ND) if for all non-zero vectors $\mathbf{x} \in \mathbb{R}^n$ holds that $\mathbf{x}^\top \mathbf{A} \mathbf{x} < 0$. This is usually denoted by $\mathbf{A} \prec \mathbf{0}$.
4. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is negative semi-definite (NSD) if for all vectors $\mathbf{x} \in \mathbb{R}^n$ holds that $\mathbf{x}^\top \mathbf{A} \mathbf{x} \leq 0$. This is usually denoted by $\mathbf{A} \preceq \mathbf{0}$.
5. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is indefinite if it is neither positive semi-definite nor negative semi-definite i.e., if there exists $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ such that $\mathbf{x}_1^\top \mathbf{A} \mathbf{x}_1 > 0$ and $\mathbf{x}_2^\top \mathbf{A} \mathbf{x}_2 < 0$.

Definiteness

Let symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ has spectral decomposition

$$\mathbf{A} = \mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^\top$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$.

1. If $\mathbf{A} \succ 0$ then $\lambda_i > 0$ for $i = 1, \ldots, n$.
2. If $\mathbf{A} \succeq 0$ then $\lambda_i \geq 0$ for $i = 1, \ldots, n$.
3. If $\mathbf{A} \prec 0$ then $\lambda_i < 0$ for $i = 1, \ldots, n$.
4. If $\mathbf{A} \preceq 0$ then $\lambda_i \leq 0$ for $i = 1, \ldots, n$.

Schur Complement

Given matrices $\mathbf{A} \in \mathbb{R}^{p\times p}$, $\mathbf{B} \in \mathbb{R}^{p\times q}$, $\mathbf{C} \in \mathbb{R}^{q\times p}$ and $\mathbf{D} \in \mathbb{R}^{q\times q}$ and suppose $\mathbf{D}$ is non-singular. Let

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \in \mathbb{R}^{(p+q)\times(p+q)}.$$

Then the Schur complement of the block $\mathbf{D}$ for $\mathbf{M}$ is

$$\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} \in \mathbb{R}^{p\times p}.$$

Then we can decompose the matrix $\mathbf{M}$ as

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{B}\mathbf{D}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix}$$

and the inverse of $\mathbf{M}$ can be written as

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix} \begin{bmatrix} (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{B}\mathbf{D}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

Schur Complement

The decomposition

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{B}\mathbf{D}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix}$$

means we have $\det(\mathbf{M}) = \det(\mathbf{D})\det(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})$.

Homework

Given the symmetric matrix

$$\mathbf{N} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{D} \end{bmatrix}$$

with non-singular $\mathbf{D}$ and let $\mathbf{S} = \mathbf{D} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B}$, then

1. $\mathbf{N} \succ \mathbf{0} \iff \mathbf{A} \succ \mathbf{0}$ and $\mathbf{S} \succ \mathbf{0}$.
2. If $\mathbf{A} \succ \mathbf{0}$, then $\mathbf{N} \succeq \mathbf{0} \iff \mathbf{S} \succeq \mathbf{0}$.

Cholesky Factorization

The symmetric positive-definite matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ has the decomposition of the form

$$\mathbf{A} = \mathbf{L}\mathbf{L}^\top$$

where $\mathbf{L} \in \mathbb{R}^{n\times n}$ is a lower triangular matrix with real and positive diagonal entries such that

$$\mathbf{L} = \begin{bmatrix} + & 0 & \cdots & 0 \\ \cdot & + & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \cdot & \cdot & \cdots & + \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

Outline

The Gradient

Suppose that $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is a smooth function that takes as input a matrix $\mathbf{X}$ of size $m \times n$ and returns a real value. Then the gradient of f with respect to $\mathbf{X}$ is

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \nabla f(\mathbf{X}) = \begin{bmatrix} \frac{\partial f(\mathbf{X})}{\partial x_{11}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(\mathbf{X})}{\partial x_{m1}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{mn}} \end{bmatrix} \in \mathbb{R}^{m\times n}.$$

Some Basic Results

1. For $\mathbf{X} \in \mathbb{R}^{m \times n}$, we have $\dfrac{\partial(f(\mathbf{X}) + g(\mathbf{X}))}{\partial \mathbf{X}} = \dfrac{\partial f(\mathbf{X})}{\partial \mathbf{X}} + \dfrac{\partial g(\mathbf{X})}{\partial \mathbf{X}}$.
2. For $\mathbf{X} \in \mathbb{R}^{m \times n}$ and $t \in \mathbb{R}$, we have $\dfrac{\partial t f(\mathbf{X})}{\partial \mathbf{X}} = t \dfrac{\partial f(\mathbf{X})}{\partial \mathbf{X}}$.
3. For $\mathbf{A}, \mathbf{X} \in \mathbb{R}^{m \times n}$, we have $\dfrac{\partial \mathrm{tr}(\mathbf{A}^\top \mathbf{X})}{\partial \mathbf{X}} = \mathbf{A}$.
4. For $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{x} \in \mathbb{R}^n$, we have $\dfrac{\partial \mathbf{x}^\top \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$.

   If $\mathbf{A}$ is symmetric, we have $\dfrac{\partial \mathbf{x}^\top \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$.

We can find more results in the matrix cookbook:
https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

The Hessian Matrix

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function that takes as input a matrix $\mathbf{x} \in \mathbb{R}^n$ and returns a real value. Then the Hessian matrix with respect to $\mathbf{x}$, written as $\nabla^2 f(\mathbf{x})$, which is defined as

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_1} & \cdots & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \cdots & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_n} \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Taylor's expansion for multivariable function $f : \mathbb{R}^n \to \mathbb{R}$

$$f(\mathbf{x}) \approx f(\mathbf{a}) + \nabla f(\mathbf{a})^\top (\mathbf{x} - \mathbf{a}) + \frac{1}{2}(\mathbf{x} - \mathbf{a})^\top \nabla^2 f(\mathbf{a})(\mathbf{x} - \mathbf{a})$$

The Hessian Matrix

Suppose $\nabla^2 f(\mathbf{x})$ is continuous in an open neighborhood of $\mathbf{x}^*$ and that $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}^*) \succ \mathbf{0}$. Then $\mathbf{x}^*$ is a strict local minimizer of f.

1. What happens if $\nabla^2 f(\mathbf{x}^*) \succeq \mathbf{0}$?
2. What happens if $\nabla^2 f(\mathbf{x}) \succeq \mathbf{0}$ holds for any $\mathbf{x}$?

Suppose $\mathbf{x}^*$ is a local minimizer of twice differentiable $f(\mathbf{x})$ and $\nabla^2 f(\mathbf{x})$ is continuous in an open neighborhood of $\mathbf{x}^*$, then $\nabla^2 f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}^*) \geq \mathbf{0}$.

Least Squares

Consider the least square problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) = \frac{1}{2}\left\|\mathbf{A}\mathbf{x}-\mathbf{b}\right\|_2^2.$$

where $\mathbf{A}\in\mathbb{R}^{m\times n}$ is full rank, $\mathbf{b}\in\mathbb{R}^m$ and $m\geq n$.

We have

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top(\mathbf{A}^\top\mathbf{A})\mathbf{x} - \mathbf{b}^\top\mathbf{A}\mathbf{x} + \frac{1}{2}\mathbf{b}^\top\mathbf{b}, \quad \nabla f(\mathbf{x}) = \mathbf{A}^\top\mathbf{A}\mathbf{x} - \mathbf{A}^\top\mathbf{b}$$

and

$$\nabla^2 f(\mathbf{x}) = \mathbf{A}^\top\mathbf{A} \succ \mathbf{0}.$$

Least Squares and Projection

The Hessian is positive definite implies the function is convex and we only need to find $\mathbf{x}$ such that $\nabla f(\mathbf{x}) = \mathbf{A}^\top \mathbf{A}\mathbf{x} - \mathbf{A}^\top \mathbf{b} = \mathbf{0}$. Hence, we have

$$\mathbf{A}^\top \mathbf{A}\mathbf{x} = \mathbf{A}^\top \mathbf{b} \Longrightarrow \mathbf{x} = (\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{A}^\top \mathbf{b}$$

For $\mathbf{A} \in \mathbb{R}^{m\times n}$ is full rank and $m > n$, we define the projection of a vector $\mathbf{b} \in \mathbb{R}^m$ onto $\mathcal{R}(\mathbf{A})$ by

$$\mathrm{Proj}(\mathbf{b}, \mathbf{A}) = \operatorname*{arg\,min}_{\mathbf{v}\in\mathcal{R}(\mathbf{A})} \|\mathbf{v} - \mathbf{b}\|_2^2 = \mathbf{A}(\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{A}^\top \mathbf{b}.$$

How to solve the problem when $\mathbf{A}$ is not full rank?

Pseudo Inverse

Let $\mathbf{A} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^\top \in \mathbb{R}^{m\times n}$ be the condense SVD, where r is the rank of $\mathbf{A}$. We define the pseudo inverse of $\mathbf{A}$ as

$$\mathbf{A}^\dagger = \mathbf{V}_r \mathbf{\Sigma}_r^{-1} \mathbf{U}_r^\top \in \mathbb{R}^{n\times m}.$$

In special case, we have

1. If $\mathrm{rank}(\mathbf{A}) = n$, we have $\mathbf{A}^\dagger = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top$.
2. If $\mathrm{rank}(\mathbf{A}) = m$, we have $\mathbf{A}^\dagger = \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top)^{-1}$.
3. If $\mathbf{A}$ is square and non-singular, we have $\mathbf{A}^\dagger = \mathbf{A}^{-1}$.

The solution of the least square problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) = \frac{1}{2}\left\|\mathbf{A}\mathbf{x} - \mathbf{b}\right\|_2^2.$$

is $\hat{\mathbf{x}} = \mathbf{A}^\dagger\mathbf{b} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{y}$, where $\mathbf{y} \in \mathbb{R}^n$.

The Gradient of $\ln\det(\cdot)$

Consider the function $f(\mathbf{A}) = \ln(\det(\mathbf{A}))$ whose domain is $n \times n$ symmetric positive definite matrices. Then we have

$$\nabla f(\mathbf{A}) = \mathbf{A}^{-1}.$$

This also can be viewed as the extension of $(\ln a)' = a^{-1}$ for $a > 0$.

Suppose that $\mathbf{F} : \mathbb{R}^{m\times n} \to \mathbb{R}^{p\times q}$ is a function that takes as input a matrix $\mathbf{X}$ of size $m \times n$ and returns a matrix $\mathbf{F}(\mathbf{X})$ of size $p \times q$ matrix, i.e,

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \in \mathbb{R}^{m\times n}$$

and

$$\mathbf{F}(\mathbf{X}) = \begin{bmatrix} f_{11}(\mathbf{X}) & f_{12}(\mathbf{X}) & \cdots & f_{1n}(\mathbf{X}) \\ f_{21}(\mathbf{X}) & f_{22}(\mathbf{X}) & \cdots & f_{2n}(\mathbf{X}) \\ \vdots & \vdots & \ddots & \vdots \\ f_{m1}(\mathbf{X}) & f_{m2}(\mathbf{X}) & \cdots & f_{mn}(\mathbf{X}) \end{bmatrix} \in \mathbb{R}^{p\times q},$$

where $f_{ij} : \mathbb{R}^{m\times n} \to \mathbb{R}$ for $i = 1, \ldots, p$ and $j = 1, \ldots, q$.

Then the derivative of $\mathbf{F}$ with respect to $\mathbf{X}$ defined as

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \nabla f(\mathbf{X}) = \begin{bmatrix} \frac{\partial f_{11}(\mathbf{X})}{\partial \mathbf{X}} & \cdots & \frac{\partial f_{1n}(\mathbf{X})}{\partial \mathbf{X}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_{m1}(\mathbf{X})}{\partial \mathbf{X}} & \cdots & \frac{\partial f_{mn}(\mathbf{X})}{\partial \mathbf{X}} \end{bmatrix} \in \mathbb{R}^{mp \times nq}$$

where each sub-matrix $\frac{\partial f_{ij}(\mathbf{X})}{\partial \mathbf{X}} \in \mathbb{R}^{p \times q}$ is the gradient of scalar-value function $f_{ij}(\cdot)$.

Based on above notation, we can define the Hessian as

$$\nabla^2 f(\mathbf{x}) = \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}^\top \partial \mathbf{x}}.$$