Matrix Review

Luo Luo

March 18, 2014

Contents

1 Basic Concepts and Notations **2**

2 Matrix Operations and Properties **3**
2.1 Matrix Multiplication . . . 3
2.1.1 Vector-Vector Products . . . 3
2.1.2 Matrix-Vector Products . . . 4
2.1.3 Matrix-Matrix Products . . . 5
2.1.4 Kronecker Products . . . 5
2.1.5 Block Matrix Products . . . 6
2.2 The Transpose . . . 6
2.3 Symmetric Matrices . . . 7
2.4 The Trace . . . 7
2.5 Vector Norms . . . 7
2.6 Vector Spaces . . . 8
2.7 Orthogonality . . . 8
2.8 The Rank . . . 9
2.9 Range and Nullspace . . . 10
2.10 QR Factorization . . . 10
2.11 Matrix Norms . . . 11
2.11.1 Induced Matrix Norms . . . 11
2.11.2 General Matrix Norms . . . 12
2.12 The Inverse . . . 13
2.13 The Determinant . . . 14
2.14 Schur Complement . . . 16

3 Eigenvalues and Eigenvectors **16**
3.1 Definitions . . . 16
3.2 General Results . . . 17
3.3 Spectral Decomposition . . . 18
3.3.1 Spectral Decomposition Theorem . . . 18
3.3.2 Applications . . . 19

4 Singular Value Decomposition **19**
4.1 Definitions and Basic Properties . . . 19
4.2 Low-Rank Approximations . . . 20
4.3 SVD vs. Spectral Decomposition . . . 22

5 Quadratic Forms and Definiteness **22**

6 Matrix Calculus **24**
6.1 Notations . . . 24
6.2 The Gradient . . . 24
6.3 The Hessian . . . 25
6.4 Least Squares . . . 25
6.5 Digression: Generalized Inverse . . . 27
6.6 Gradients of the Determinant . . . 27
6.7 Eigenvalues as Optimization . . . 28
6.8 General Derivatives of Matrix Functions . . . 28

7 Bibliographic Remarks **29**

1 Basic Concepts and Notations

We use the following notation:

- By $\mathbf{A} \in \mathbb{R}^{m\times n}$ we denote a ***matrix*** with m rows and n columns, where the entries of $\mathbf{A}$ are real numbers.
- By $\mathbf{x} \in \mathbb{R}^n$, we denote a ***vector*** with n entries.
- An n-dimensional vector is often thought of as a matrix with n rows and 1 column, known as ***column vector***. A ***row vector*** is a matrix with 1 row and n columns, which we typically write $\mathbf{x}^T$ ($\mathbf{x}^T$ denote the transpose of $\mathbf{x}$, which we will define shortly).
- The ith element of a vector $\mathbf{x}$ is denoted x_i:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

- We use notation a_{ij} to denote the entry of $\mathbf{A}$ in the ith row and jth column:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} = (a_{ij}).$$

- We denote the jth columns of $\mathbf{A}$ by $\mathbf{a}_j$:

$$\mathbf{A} = \begin{bmatrix} | & | & \dots & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_n \\ | & | & \dots & | \end{bmatrix}.$$

- We denote the ith row of $\mathbf{A}$ by $\mathbf{a}_i$:

$$\mathbf{A} = \begin{bmatrix} — & \mathbf{a}_{(1)}^T & — \\ — & \mathbf{a}_{(2)}^T & — \\ & \vdots & \\ — & \mathbf{a}_{(m)}^T & — \end{bmatrix} \text{ or } \mathbf{A} = \begin{bmatrix} — & \mathbf{a}_1^T & — \\ — & \mathbf{a}_2^T & — \\ & \vdots & \\ — & \mathbf{a}_m^T & — \end{bmatrix}.$$

- The meanings of the notations usually should be obvious from its use when they are ambiguous.
- A matrix written in terms of its sub-matrices is called a ***partitioned matrix***:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \dots & \mathbf{A}_{1q} \\ \mathbf{A}_{21} & \mathbf{A}_{12} & \dots & \mathbf{A}_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{p1} & \mathbf{A}_{p2} & \dots & \mathbf{A}_{pq} \end{bmatrix},$$

  where $\mathbf{A}_{ij} \in \mathbb{R}^{r\times s}$, $m = pr$, and $n = qs$. Based on the Matlab colon notation, we have $\mathbf{A}_{ij} = \mathbf{A}\big((i-1)r+1 : ir, (j-1)s+1 : s\big)$.

Table 1: Particular matrices and types of matrix

Name	Definition	Notation
Scalar	$m = n = 1$	$a, b \dots$
Column vector	$n = 1$	$\mathbf{x}, \mathbf{y} \dots$
All-ones vector	$[1, \dots, 1]^T$	$\mathbf{1}$ or $\mathbf{1}_n$
Null	$a_{ij} = 0$	$\mathbf{0}$
Square	$m = n$	$\mathbf{A} \in \mathbb{R}^{n \times n}$
Diagonal	$m = n$, $a_{ij} = 0$ if $i \neq j$	$\text{diag}(a_{11}, \dots, a_{nn})$ or $\text{diag}([a_{11}, \dots, a_{nn}]^T)$
Identity	$\text{diag}(\mathbf{1})$	$\mathbf{I}$ or $\mathbf{I}_n$
Symmetric	$a_{ij} = a_{ji}$	$\mathbf{A} \in \mathbb{S}^n$

Table 2: Basic matrix operations

Operation	Restrictions	Notations and Definitions
Addition / Subtraction	$\mathbf{A}, \mathbf{B}$ of the same order	$\mathbf{A} \pm \mathbf{B} = (a_{ij} \pm b_{ij})$
Scalar multiplication		$c\mathbf{A} = c(a_{ij})$
Inner product	$\mathbf{x}, \mathbf{y}$ of the same order	$\mathbf{x}^T\mathbf{y} = \sum x_i y_i$
Multiplication	#columns of $\mathbf{A}$ equals #rows of $\mathbf{B}$	$\mathbf{AB} = (\mathbf{a}_{(i)}^T \mathbf{b}_j)$
Kronecker Product		$\mathbf{A} \otimes \mathbf{B} = (a_{ij}\mathbf{B})$
Transpose		$\mathbf{A}^T = [\mathbf{a}_{(1)}, \dots, \mathbf{a}_{(m)}]$
Trace	$\mathbf{A}$ is square	$\text{tr}(\mathbf{A}) = \sum a_{ii}$
Determinant	$\mathbf{A}$ is square	$\lvert\mathbf{A}\rvert$ or $\det(\mathbf{A})$
Inverse	$\mathbf{A}$ is sqaure and $\det(\mathbf{A}) \neq 0$	$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$
Pseudo-inverse		$\mathbf{A}^\dagger$

2 Matrix Operations and Properties

2.1 Matrix Multiplication

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

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

where

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

Note that in order for the matrix product to exist, the number of columns in $\mathbf{A}$ must equal the number of rows in $\mathbf{B}$. There are many ways of looking at matrix multiplication. We'll start by examining a few special cases.

2.1.1 Vector-Vector Products

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

$$\mathbf{x}^T\mathbf{y} \in \mathbb{R} = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i.$$

Note that it is always the case that $\mathbf{x}^T\mathbf{y} = \mathbf{y}^T\mathbf{x}$.

Given $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n$ (not for the same size), $\mathbf{x}\mathbf{y}^T$ is called the ***outer product*** of the vectors. It is a matrix whose entries are given by $(\mathbf{x}\mathbf{y}^T)_{ij} = x_i y_j$,

$$\mathbf{x}\mathbf{y}^T \in \mathbb{R}^{m\times n} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \dots & y_n \end{bmatrix} = \begin{bmatrix} x_1y_1 & x_1y_2 & \dots & x_1y_n \\ x_2y_1 & x_2y_2 & \dots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_my_1 & x_my_2 & \dots & x_my_n \end{bmatrix}.$$

Consider the matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ whose columns are equal to some vector $\mathbf{x} \in \mathbb{R}^m$. Using outer products, we can represent $\mathbf{A}$ compactly as,

$$\mathbf{A} = \begin{bmatrix} | & | & & | \\ \mathbf{x} & \mathbf{x} & \dots & \mathbf{x} \\ | & | & & | \end{bmatrix} = \begin{bmatrix} x_1 & x_1 & \dots & x_1 \\ x_2 & x_2 & \dots & x_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_m & x_m & \dots & x_m \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} = \mathbf{x}\mathbf{1}^T.$$

2.1.2 Matrix-Vector Products

Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and a vector $\mathbf{x} \in \mathbb{R}^n$, their product is a vector $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{R}^m$. There are a couple ways of looking at matrix-vector multiplication, and we will look at each of them in turn.

If we write $\mathbf{A}$ by rows, then we can express $\mathbf{A}\mathbf{x}$ as,

$$\mathbf{y} = \mathbf{A}\mathbf{x} = \begin{bmatrix} \text{—} & \mathbf{a}_{(1)}^T & \text{—} \\ \text{—} & \mathbf{a}_{(2)}^T & \text{—} \\ & \vdots & \\ \text{—} & \mathbf{a}_{(m)}^T & \text{—} \end{bmatrix} \mathbf{x} = \begin{bmatrix} \mathbf{a}_{(1)}^T\mathbf{x} \\ \mathbf{a}_{(2)}^T\mathbf{x} \\ \vdots \\ \mathbf{a}_{(m)}^T\mathbf{x} \end{bmatrix}.$$

In other words, the ith entry of $\mathbf{y}$ is equal to the inner product of the ith row of $\mathbf{A}$ and $\mathbf{x}$, $y_i = \mathbf{a}_{(i)}^T\mathbf{x}$.

Alternatively, lets write $\mathbf{A}$ in column form. In this case we see that,

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

In other words, $\mathbf{y}$ is a ***linear combination*** of the columns of $\mathbf{A}$, where the coefficients of the linear combination are given by the entries of $\mathbf{x}$.

So far we have been multiplying on the right by a column vector. We also have similar interpretation of multiplying on the left by a row vector. For $\mathbf{A} \in \mathbb{R}^{m\times n}$ and a vector $\mathbf{x} \in \mathbb{R}^m$, which gives

$$\mathbf{y}^T = \mathbf{x}^T\mathbf{A} = \mathbf{x}^T \begin{bmatrix} | & | & \dots & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_n \\ | & | & \dots & | \end{bmatrix} = \begin{bmatrix} \mathbf{x}^T\mathbf{a}_1 \\ \mathbf{x}^T\mathbf{a}_2 \\ \vdots \\ \mathbf{x}^T\mathbf{a}_n \end{bmatrix},$$

$$\begin{aligned} \mathbf{y}^T &= \mathbf{x}^T\mathbf{A} \\ &= \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} \text{—} & \mathbf{a}_{(1)}^T & \text{—} \\ \text{—} & \mathbf{a}_{(2)}^T & \text{—} \\ & \vdots & \\ \text{—} & \mathbf{a}_{(m)}^T & \text{—} \end{bmatrix} \\ &= x_1 \begin{bmatrix} \text{—} & \mathbf{a}_{(1)}^T & \text{—} \end{bmatrix} + x_2 \begin{bmatrix} \text{—} & \mathbf{a}_{(2)}^T & \text{—} \end{bmatrix} + \cdots + x_n \begin{bmatrix} \text{—} & \mathbf{a}_{(n)}^T & \text{—} \end{bmatrix}. \end{aligned}$$

2.1.3 Matrix-Matrix Products

Armed with this knowledge, we can now look at four different (but, of course, equivalent) ways of viewing the matrix-matrix multiplication $\mathbf{C} = \mathbf{AB}$.

- We can view matrix-matrix multiplication as a set of vector-vector products. The most obvious viewpoint, which follows immediately from the definition, is that the (i, j)th entry of $\mathbf{C}$ is equal to the inner product of the ith row of $\mathbf{A}$ and the jth row of $\mathbf{B}$.

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} \text{—} & \mathbf{a}_{(1)}^T & \text{—} \\ \text{—} & \mathbf{a}_{(2)}^T & \text{—} \\ & \vdots & \\ \text{—} & \mathbf{a}_{(m)}^T & \text{—} \end{bmatrix} \begin{bmatrix} | & | & \cdots & | \\ \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_p \\ | & | & \cdots & | \end{bmatrix} \mathbf{A} = \begin{bmatrix} \mathbf{a}_{(1)}^T\mathbf{b}_1 & \mathbf{a}_{(1)}^T\mathbf{b}_2 & \cdots & \mathbf{a}_{(1)}^T\mathbf{b}_p \\ \mathbf{a}_{(2)}^T\mathbf{b}_1 & \mathbf{a}_{(2)}^T\mathbf{b}_2 & \cdots & \mathbf{a}_{(2)}^T\mathbf{b}_p \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_{(m)}^T\mathbf{b}_1 & \mathbf{a}_{(m)}^T\mathbf{b}_2 & \cdots & \mathbf{a}_{(m)}^T\mathbf{b}_p \end{bmatrix}.$$

- Alternatively, we can represent $\mathbf{A}$ by columns, and $\mathbf{B}$ by rows. This representation leads to a much trickier interpretation of $\mathbf{AB}$ as a sum of outer products.

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} | & | & \cdots & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \\ | & | & \cdots & | \end{bmatrix} \begin{bmatrix} \text{—} & \mathbf{b}_{(1)}^T & \text{—} \\ \text{—} & \mathbf{b}_{(2)}^T & \text{—} \\ & \vdots & \\ \text{—} & \mathbf{b}_{(n)}^T & \text{—} \end{bmatrix} = \sum_{i=1}^{n} \mathbf{a}_i \mathbf{b}_i^T.$$

- We can also view matrix-matrix multiplication as a set of matrix-vector products. Specifically, if we represent $\mathbf{B}$ by columns, we can view the columns of $\mathbf{C}$ as matrix-vector products between $\mathbf{A}$ and the columns of $\mathbf{B}$.

$$\mathbf{C} = \mathbf{AB} = \mathbf{A} \begin{bmatrix} | & | & \cdots & | \\ \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_p \\ | & | & \cdots & | \end{bmatrix} = \begin{bmatrix} | & | & \cdots & | \\ \mathbf{Ab}_1 & \mathbf{Ab}_2 & \cdots & \mathbf{Ab}_p \\ | & | & \cdots & | \end{bmatrix}$$

- We have the analogous viewpoint, where we represent $\mathbf{A}$ by rows, and view the rows of C as the matrix-vector product between the rows of $\mathbf{A}$ and $\mathbf{C}$.

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} \text{—} & \mathbf{a}_{(1)}^T & \text{—} \\ \text{—} & \mathbf{a}_{(2)}^T & \text{—} \\ & \vdots & \\ \text{—} & \mathbf{a}_{(m)}^T & \text{—} \end{bmatrix} \mathbf{B} = \begin{bmatrix} \text{—} & \mathbf{a}_{(1)}^T\mathbf{B} & \text{—} \\ \text{—} & \mathbf{a}_{(2)}^T\mathbf{B} & \text{—} \\ & \vdots & \\ \text{—} & \mathbf{a}_{(m)}^T\mathbf{B} & \text{—} \end{bmatrix}$$

In addition to this, it is useful to know a few basic properties of matrix multiplication at a higher level:

- Matrix multiplication is associative: $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$.
- Matrix multiplication is distributive: $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$.
- Matrix multiplication is, in general, not commutative; that is, it can be the case that $\mathbf{AB} \neq \mathbf{BA}$. (For example, if $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{m \times q}$, the matrix product $\mathbf{BA}$ does not even exist if m and q are not equal!)

2.1.4 Kronecker Products

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 The Kronecker product should not be confused with the usual matrix multiplication, which is an entirely different operation.

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 $mp \times nq$ block matrix:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix},$$

more explicitly:

$$\mathbf{A}\otimes\mathbf{B}=\begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} & \cdots & a_{11}b_{1q} & \cdots & \cdots & a_{1n}b_{11} & a_{1n}b_{12} & \cdots & a_{1n}b_{1q} \\ a_{11}b_{21} & a_{11}b_{22} & \cdots & a_{11}b_{2q} & \cdots & \cdots & a_{1n}b_{21} & a_{1n}b_{22} & \cdots & a_{1n}b_{2q} \\ \vdots & \vdots & \ddots & \vdots & & & \vdots & \vdots & \ddots & \vdots \\ a_{11}b_{p1} & a_{11}b_{p2} & \cdots & a_{11}b_{pq} & \cdots & \cdots & a_{1n}b_{p1} & a_{1n}b_{p2} & \cdots & a_{1n}b_{pq} \\ \vdots & \vdots & & \vdots & \ddots & & \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots & & \ddots & \vdots & \vdots & & \vdots \\ a_{m1}b_{11} & a_{m1}b_{12} & \cdots & a_{m1}b_{1q} & \cdots & \cdots & a_{mn}b_{11} & a_{mn}b_{12} & \cdots & a_{mn}b_{1q} \\ a_{m1}b_{21} & a_{m1}b_{22} & \cdots & a_{m1}b_{2q} & \cdots & \cdots & a_{mn}b_{21} & a_{mn}b_{22} & \cdots & a_{mn}b_{2q} \\ \vdots & \vdots & \ddots & \vdots & & & \vdots & \vdots & \ddots & \vdots \\ a_{m1}b_{p1} & a_{m1}b_{p2} & \cdots & a_{m1}b_{pq} & \cdots & \cdots & a_{mn}b_{p1} & a_{mn}b_{p2} & \cdots & a_{mn}b_{pq} \end{bmatrix}.$$

The following properties of transposes are easily verified:

- $\mathbf{A}\otimes(\mathbf{B}+\mathbf{C})=\mathbf{A}\otimes\mathbf{B}+\mathbf{A}\otimes\mathbf{C}$.
- $(\mathbf{A}\otimes\mathbf{B})\otimes\mathbf{C}=\mathbf{A}\otimes(\mathbf{B}\otimes\mathbf{C})$.
- $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D})=\mathbf{AC}\otimes\mathbf{BD}$.
- $\mathrm{tr}(\mathbf{A}\otimes\mathbf{B})=\mathrm{tr}(\mathbf{A})\mathrm{tr}(\mathbf{B})$.

2.1.5 Block Matrix Products

A block partitioned matrix product can sometimes be used on submatrices forms. The partitioning of the factors is not arbitrary, however, and requires conformable partitions between two matrices such that all submatrix products that will be used are defined. Given a matrix $\mathbf{A}\in\mathbb{R}^{m\times p}$ with q row partitions and s column partitions

$$\mathbf{A}=\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1s} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \cdots & \mathbf{A}_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{q1} & \mathbf{A}_{q2} & \cdots & \mathbf{A}_{qs} \end{bmatrix}$$

and $\mathbf{B}\in\mathbb{R}^{p\times n}$ with s row partitions and r column partitions

$$\mathbf{B}=\begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} & \cdots & \mathbf{B}_{1r} \\ \mathbf{B}_{21} & \mathbf{B}_{22} & \cdots & \mathbf{B}_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}_{s1} & \mathbf{B}_{s2} & \cdots & \mathbf{B}_{sr} \end{bmatrix},$$

that are compatible with the partitions of $\mathbf{A}$, the matrix product $\mathbf{C}=\mathbf{AB}$ can be formed blockwise, yielding as an matrix with q row partitions and r column partitions. The matrices in your matrix $\mathbf{C}$ are calculated by multiplying

$$\mathbf{C}_{ij}=\sum_{k=1}^{s}\mathbf{A}_{ik}\mathbf{B}_{kj}.$$

2.2 The Transpose

The ***transpose*** of a matrix results from flipping the rows and columns. Given a matrix $\mathbf{A}\in\mathbb{R}^{m\times n}$, its transpose, written $\mathbf{A}^T\in\mathbb{R}^{n\times m}$, is the $n\times m$ matrix whose entries are given by $(\mathbf{A}^T)_{ij}=\mathbf{A}_{ji}$. We have in fact already been using the transpose when describing row vectors, since the transpose of a column vector is naturally a row vector. The following properties of transposes are easily verified:

- $(\mathbf{A}^T)^T=\mathbf{A}$
- $(\mathbf{AB})^T=\mathbf{B}^T\mathbf{A}^T$

- $(\mathbf{A}+\mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$

Sometimes, we also use quotation marks $\mathbf{A}'$ to express the transpose of $\mathbf{A}$. But note that the quotation in Matlab is to express the ***hermitian conjugate*** (notated as $\mathbf{A}^*$), not transpose. The transpose operation in Matlab should be written as dot quotation.

2.3 Symmetric Matrices

A ***square matrix*** $\mathbf{A} \in \mathbb{R}^{n\times n}$ is ***symmetric*** if $\mathbf{A} = \mathbf{A}^T$. It is ***anti-symmetric*** if $\mathbf{A} = -\mathbf{A}^T$. It is easy to show that for any matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, the matrix $\mathbf{A} - \mathbf{A}^T$ is anti-symmetric. From this it follows that any square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ can be represented as a sum of a symmetric matrix and an anti-symmetric matrix, since

$$\mathbf{A} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^T) + \frac{1}{2}(\mathbf{A} - \mathbf{A}^T)$$

and the first matrix on the right is symmetric, while the second is anti-symmetric. It turns out that symmetric matrices occur a great deal in practice, and they have many nice properties which we will look at shortly. It is common to denote the set of all symmetric matrices of size n as $\mathbb{S}^n$, so that $\mathbf{A} \in \mathbb{S}^n$ means that $\mathbf{A}$ is a symmetric $n \times n$ matrix. Another property is that for $\mathbf{A} \in \mathbb{S}^n, \mathbf{x} \in \mathbb{R}^n, y \in \mathbb{R}^n$, we have $\mathbf{x}^T\mathbf{A}\mathbf{y} = \mathbf{y}^T\mathbf{A}\mathbf{x}$, which can be verified easily.

2.4 The Trace

The ***trace*** of a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, denoted $\mathrm{tr}(\mathbf{A})$ (or just $\mathrm{tr}\mathbf{A}$ if the parentheses are obviously implied), 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:

- For $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathrm{tr}\mathbf{A} = \mathrm{tr}\mathbf{A}^T$.
- For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$, $\mathrm{tr}(\mathbf{A} \pm \mathbf{B}) = \mathrm{tr}\mathbf{A} \pm \mathrm{tr}\mathbf{B}$.
- For $\mathbf{A} \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}$, $\mathrm{tr}(c\mathbf{A}) = c\mathrm{tr}\mathbf{A}$.
- For $\mathbf{A}, \mathbf{B}$ such that $\mathbf{AB}$ is square, $\mathrm{tr}\mathbf{AB} = \mathrm{tr}\mathbf{BA}$.
- For $\mathbf{A}, \mathbf{B}, \mathbf{C}$ such that $\mathbf{ABC}$ is square, $\mathrm{tr}\mathbf{ABC} = \mathrm{tr}\mathbf{BCA} = \mathrm{tr}\mathbf{CAB}$, and so on for the product of more matrices.
- For $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathrm{tr}(\mathbf{A}^T\mathbf{A}) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2$.

The fourth equality, where the main work occurs, uses the commutativity of scalar multiplication in order to reverse the order of the terms in each product, and the commutativity and associativity of scalar addition in order to rearrange the order of the summation.

2.5 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),

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

More formally, a norm is any function $\mathbb{R}^n \to \mathbb{R}$ that satisfies 4 properties:

1. For all $\mathbf{x} \in \mathbb{R}^n$, $||\mathbf{x}|| \geq 0$ (non-negativity).
2. $||\mathbf{x}|| = 0$ if and only if $\mathbf{x} = \mathbf{0}$ (definiteness).

3. For all $\mathbf{x} \in \mathbb{R}^n, t \in \mathbb{R}, ||t\mathbf{x}|| = |t|||\mathbf{x}||$ (homogeneity).

4. For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n, ||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$ (triangle inequality).

Other examples of norms are the ℓ_1 norm,

$$||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i|,$$

and the ℓ_∞ norm,

$$||\mathbf{x}||_\infty = \max_i |x_i|.$$

In fact, all three norms presented so far are examples of the family of ℓ_p norms, which are parameterized by a real number $p \geq 1$, and defined as

$$||\mathbf{x}||_p = \Big(\sum_{i=1}^{n} |x_i|^p \Big)^{1/p}.$$

2.6 Vector Spaces

The set of vectors in $\mathbb{R}^m$ satisfies the following properties for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$ and all $a, b \in \mathbb{R}$.

- $a(\mathbf{x} + \mathbf{y}) = a\mathbf{x} + a\mathbf{y}$
- $(a + b)\mathbf{x} = a\mathbf{x} + b\mathbf{x}$
- $(ab)\mathbf{x} = a(b\mathbf{y})$
- $1\mathbf{x} = \mathbf{x}$

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 $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^m$ 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 that the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are linearly dependent; otherwise the vectors are linearly independent.

Let $\mathcal{W}$ be a subspace of $\mathbb{R}^n$. Then a ***basis*** of $\mathcal{W}$ is a maximal linear independent set of vectors. The following properties hold for a basis of $\mathcal{W}$:

- Every basis of $\mathcal{W}$ contains the same number of elements. This number is called the ***dimension*** of $\mathcal{W}$ and denoted $\dim(\mathcal{W})$. In particular $\dim(\mathbb{R}^n) = n$.
- If $\mathbf{x}_1, \ldots, \mathbf{x}_k$ is a basis for $\mathcal{W}$ then every element $\mathbf{x} \in \mathcal{W}$ can be express as a linear combination of $\mathbf{x}_1, \ldots, \mathbf{x}_k$.

2.7 Orthogonality

Two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ are ***orthogonal*** if $\mathbf{x}^T\mathbf{y} = 0$. A vector $\mathbf{x} \in \mathbb{R}^n$ is ***normalized*** if $||\mathbf{x}||_2 = 1$. A square matrix $\mathbf{U} \in \mathbb{R}^{n \times n}$ is orthogonal (note the different meanings when talking about vectors versus matrices) if all its columns are orthogonal to each other and are normalized (the columns are then referred to as being orthonormal). It follows immediately from the definition of orthogonality and normality that

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

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 $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, but $\mathbf{U}\mathbf{U}^T \neq \mathbf{I}$, we call that $\mathbf{U}$ is ***column orthonormal***.

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

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

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

Orthogonal matrices can be used to represent a rotation, since the cosine of the angle $\theta \in [0, \pi]$ between lines from $\mathbf{0}$ to $\mathbf{x}$ and $\mathbf{0}$ to $\mathbf{y}$ will not be changed by orthogonal transformation. That is, for a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is orthogonal:

$$\cos\theta' = \frac{(\mathbf{Ax})^T\mathbf{Ay}}{||\mathbf{Ax}||_2||\mathbf{Ay}||_2} = \frac{\mathbf{x}^T(\mathbf{A}^T\mathbf{A})\mathbf{y}}{||\mathbf{x}||_2||\mathbf{y}||_2} = \cos\theta.$$

In two dimension space, the orthogonal matrix can be represented as

$$\begin{bmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{bmatrix}$$

and represents a rotation of the coordinate axes counterclockwise through an angle θ.

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 1 and are orthogonal to one another. In particular, if $\mathbf{A} \in \mathbb{R}^{n\times n}$ is an orthogonal matrix then the columns of $\mathbf{A}$ form an orthogonal basis of $\mathbb{R}^n$.

2.8 The Rank

The ***column rank*** of a matrix $A \in \mathbb{R}^{m\times n}$ is the size of the largest subset of columns of $\mathbf{A}$ that constitute a linearly independent set. With some abuse of terminology, this is often referred to simply as the number of linearly independent columns of $\mathbf{A}$. In the same way, the ***row rank*** is the largest number of rows of $\mathbf{A}$ that constitute a linearly independent set.

For any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, it turns out that the column rank of $\mathbf{A}$ is equal to the row rank of $\mathbf{A}$. Then we present a short proof, which uses only basic properties of linear combinations of vectors. Let the column rank of $\mathbf{A}$ be r and let $\mathbf{c}_1, ..., \mathbf{c}_r$ be any basis for the column space of $\mathbf{A}$. Place these as the columns of an matrix

$$\mathbf{C} = \begin{bmatrix} | & | & \cdots & | \\ \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_r \\ | & | & \cdots & | \end{bmatrix} \in \mathbb{R}^{m\times r}.$$

Every column of $\mathbf{A}$ can be expressed as a linear combination of the r columns in $\mathbf{C}$. This means that there is a matrix $\mathbf{Q} = (q_{ij}) \in \mathbb{R}^{r\times n}$ such that $\mathbf{A} = \mathbf{CQ}$. $\mathbf{Q}$ is the matrix whose i-th column is formed from the coefficients giving the ith column of $\mathbf{A}$ as a linear combination of the r columns of $\mathbf{C}$.

$$\begin{bmatrix} | \\ \mathbf{a}_j \\ | \end{bmatrix} = \sum_{i=1}^{r} q_{ij}\mathbf{c}_i = \begin{bmatrix} | & | & \cdots & | \\ \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_r \\ | & | & \cdots & | \end{bmatrix} \begin{bmatrix} | \\ \mathbf{q}_j \\ | \end{bmatrix}$$

$$\implies \mathbf{A} = \begin{bmatrix} | & | & \cdots & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \\ | & | & \cdots & | \end{bmatrix} = \begin{bmatrix} | & | & \cdots & | \\ \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_r \\ | & | & \cdots & | \end{bmatrix} \begin{bmatrix} | & | & & | \\ \mathbf{q}_1 & \mathbf{q}_2 & \cdots & \mathbf{q}_n \\ | & | & & | \end{bmatrix} = \mathbf{CQ}.$$

Now, each row of $\mathbf{A}$ is given by a linear combination of the r rows of $\mathbf{Q}$,

$$\mathbf{A} = \mathbf{CQ} \implies \begin{bmatrix} — & \mathbf{a}_{(1)}^T & — \\ — & \mathbf{a}_{(2)}^T & — \\ & \vdots & \\ — & \mathbf{a}_{(m)}^T & — \end{bmatrix} = \begin{bmatrix} — & \mathbf{c}_{(1)}^T & — \\ — & \mathbf{c}_{(2)}^T & — \\ & \vdots & \\ — & \mathbf{c}_{(m)}^T & — \end{bmatrix} \begin{bmatrix} — & \mathbf{q}_{(1)}^T & — \\ — & \mathbf{q}_{(2)}^T & — \\ & \vdots & \\ — & \mathbf{q}_{(r)}^T & — \end{bmatrix}$$

$$\implies \begin{bmatrix} — & \mathbf{a}_{(i)}^T & — \end{bmatrix} = \begin{bmatrix} — & \mathbf{c}_{(i)}^T & — \end{bmatrix} \mathbf{Q}.$$

Therefore, the row rank of $\mathbf{A}$ cannot exceed r. This proves that the row rank of $\mathbf{A}$ is less than or equal to the column rank of $\mathbf{A}$. This result can be applied to any matrix, so apply the result to the transpose of $\mathbf{A}$. Since the row rank of the transpose of $\mathbf{A}$ is the column rank of $\mathbf{A}$ and the column rank of the transpose of $\mathbf{A}$ is the row rank of $\mathbf{A}$, this establishes the reverse inequality and we obtain the equality of the row rank and the column rank of $\mathbf{A}$, so both quantities are referred to collectively as the ***rank*** of $\mathbf{A}$, denoted as $\text{rank}(\mathbf{A})$.

The following are some basic properties of the rank:

- For $A \in \mathbb{R}^{m\times n}, \text{rank}(\mathbf{A}) \leq \min(m, n)$. If $\text{rank}(\mathbf{A}) = \min(m, n)$, then $\mathbf{A}$ is said to be full rank.
- For $\mathbf{A} \in \mathbb{R}^{m\times n}, \text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^T)$.
- For $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{B} \in \mathbb{R}^{n\times p}, \text{rank}(\mathbf{AB}) \leq \min(\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B}))$.
- For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}, \text{rank}(\mathbf{A} + \mathbf{B}) \leq \text{rank}(\mathbf{A}) + \text{rank}(\mathbf{B})$.

2.9 Range and Nullspace

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

$$\text{span}(\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}) = \left\{ v : v = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i, \alpha_i \in \mathbb{R} \right\}.$$

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,

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

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 $\mathbf{A}$,

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

Note that vectors in $\mathcal{R}(\mathbf{A})$ are of size m, while vectors in the $\mathcal{N}(\mathbf{A})$ are of size n, so vectors in $\mathcal{R}(\mathbf{A}^T)$ and $\mathcal{N}(\mathbf{A})$ are both in $\mathbb{R}^n$. In fact, we can say much more that

$$\{w : w = u + v, u \in \mathcal{R}(\mathbf{A}^T), v \in \mathcal{N}(\mathbf{A})\} = \mathbb{R}^n \text{ and } \mathcal{R}(\mathbf{A}^T) \cap \mathcal{N}(\mathbf{A}) = \{\mathbf{0}\}.$$

In other words, $\mathcal{R}(\mathbf{A}^T)$ and $\mathcal{N}(\mathbf{A})$ together span the entire space of $\mathbb{R}^n$. Furthermore, the Table 3 summarizes some properties of four fundamental subspace of $\mathbf{A}$ including the range and the nullspace, where $\mathbf{A}$ has singular value decomposition[1]:

$$\mathbf{A} = \mathbf{U\Sigma V}^T.$$

Table 3: Properties of four fundamental subspaces of rank r matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$

Name of Subspace	Notation	Containing Space	Dimension	Basis
range, column space or image	$\mathcal{R}(\mathbf{A})$	$\mathbb{R}^m$	r (rank)	The first r columns of $\mathbf{U}$
nullspace or kernel	$\mathcal{N}(\mathbf{A})$	$\mathbb{R}^n$	$n - r$ (nullity)	The last $(n - r)$ columns of $\mathbf{V}$
row space or coimage	$\mathcal{R}(\mathbf{A}^T)$	$\mathbb{R}^n$	r (rank)	The first r columns of $\mathbf{V}$
left nullspace or cokernel	$\mathcal{N}(\mathbf{A}^T)$	$\mathbb{R}^m$	$m - r$ (corank)	The last $(m - r)$ columns of $\mathbf{U}$

2.10 QR Factorization

Given a full rank matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ $(m \geq n)$, we can construct the vectors $\mathbf{q}_1, \ldots, \mathbf{q}_n$ and entries r_{ij} such that

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

[1]Singular value decomposition will be introduced in section 4.

where the diagonal entries r_{kk} are nonzero, then $\mathbf{a}_1, \ldots, \mathbf{a}_k$ can be expressed as linear combination of $\mathbf{q}_1, \ldots, \mathbf{q}_k$ and the invertibility of the upper-left $k \times k$ block of the triangular matrix implies that, conversely, $\mathbf{q}_1, \ldots, \mathbf{q}_k$ can be expressed as linear combination of $\mathbf{a}_1, \ldots, \mathbf{a}_k$. Written out, these equations take the form

$$\begin{aligned}
\mathbf{a}_1 &= r_{11}\mathbf{q}_1 \\
\mathbf{a}_2 &= r_{12}\mathbf{q}_1 + r_{22}\mathbf{q}_2 \\
\mathbf{a}_3 &= r_{13}\mathbf{q}_1 + r_{23}\mathbf{q}_2 + r_{33}\mathbf{q}_3 \\
&\vdots \\
\mathbf{a}_n &= r_{1n}\mathbf{q}_1 + r_{2n}\mathbf{q}_2 + \cdots + r_{nn}\mathbf{q}_n
\end{aligned}$$

As matrix formula, we have

$$\mathbf{A} = \hat{\mathbf{Q}}\hat{\mathbf{R}},$$

where $\hat{\mathbf{Q}}$ is $m \times n$ with orthonormal columns and $\hat{\mathbf{R}}$ is $n \times n$ and upper-triangular. Such a factorization is called a ***reduced QR factorization*** of $\mathbf{A}$.

A ***full QR factorization*** of $\mathbf{A} \in \mathbb{R}^{m\times n}$ $(m \geq n)$ goes futher, appending an additional $m - n$ orthogonal columns to $\hat{\mathbf{Q}}$ so that it becomes an $m \times m$ orthogonal matrix $\mathbf{Q}$. In the process, rows of zeros are appended to $\hat{\mathbf{R}}$ so that it becomes an $m \times n$ matrix $\mathbf{R}$. Then we have

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

There is an old idea, known as ***Gram-Schmidt orthogonalization*** to construct the vectors $\mathbf{q}_1 \ldots, \mathbf{q}_n$ and entires r_{ij} for reduced QR factorization. Let's rewrite the relationship between $\mathbf{q}_1 \ldots, \mathbf{q}_n$ and $\mathbf{a}_1 \ldots, \mathbf{a}_n$, in the form

$$\begin{aligned}
\mathbf{q}_1 &= \frac{\mathbf{a}_1}{r_{11}} \\
\mathbf{q}_2 &= \frac{\mathbf{a}_2 - r_{12}\mathbf{q}_1}{r_{22}} \\
\mathbf{q}_3 &= \frac{\mathbf{a}_3 - r_{13}\mathbf{q}_1 - r_{23}\mathbf{q}_2}{r_{33}} \\
&\vdots \\
\mathbf{q}_n &= \frac{\mathbf{a}_n - \sum_{i=1}^{n-1} r_{in}\mathbf{q}_i}{r_{nn}}.
\end{aligned}$$

Let $r_{ij} = \mathbf{q}_i^T\mathbf{a}_j$ for $i \neq j$, and $|r_{jj}| = ||\mathbf{a}_j - \sum_{i=1}^{j-1} r_{ij}\mathbf{q}_i||_2$, then we have $\mathbf{q}_i^T\mathbf{q}_j = 0$ (can be proved by induction). But how to obtain full QR factorization?

2.11 Matrix Norms

2.11.1 Induced Matrix Norms

An $m \times n$ matrix can be viewed as a vector in an mn-dimensional space: each of the mn entries of the matrix is an independent coordinate. Any mn-dimensional norm can therefore be used for measuring the 'size' of such a matrix. In dealing with a space of matrices, these are the ***induced matrix norms***, defined in terms of the behavior of a matrix as an operator between its normed domain and range space.

Given vector norms $||\cdot||_{(m)}$, $||\cdot||_{(n)}$ and on the domain and the range of $\mathbf{A} \in \mathbb{R}^{m\times n}$, respectively, the induced matrix norm $||\mathbf{A}||_{(m,n)}$ is the smallest number C for which the following inequality holds for all $\mathbf{x} \in \mathbb{R}^n$:

$$||\mathbf{Ax}||_{(m)} \leq C||\mathbf{x}||_{(n)}.$$

The induced matrix norm can be defined equivalently as following:

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

Note that the (m) and (n) in subscript are only represented the dimensionality of vectors, rather than ℓ_m or ℓ_n norm. The notations $||\cdot||_{(m)}$ and $||\cdot||_{(n)}$ can be any the same type of vector norm.

For example, if $\mathbf{A} \in \mathbb{R}^{m\times n}$, then its 1-norm $||\mathbf{A}||_1$ is equal to the 'maximum column sum' of $\mathbf{A}$. We explain and derive this result as follows. Let's write $\mathbf{A}$ in terms of its columns

$$\mathbf{A} = \begin{bmatrix} | & | & \dots & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_n \\ | & | & \dots & | \end{bmatrix}.$$

Any vector $\mathbf{x}$ in the set $\{\mathbf{x} \in \mathbb{R}^n : \sum_{i=1}^n |x_i| = 1\}$ satisfies

$$||\mathbf{A}\mathbf{x}||_1 = \Big\| \sum_{i=1}^n x_i \mathbf{a}_i \Big\|_1 \leq \sum_{i=1}^n |x_i| ||\mathbf{a}_i|| \leq \sum_{i=1}^n |x_i| \Big(\max_{1\leq j\leq n} ||\mathbf{a}_j||_1 \Big) = \max_{1\leq j\leq n} ||\mathbf{a}_j||_1.$$

By choosing $\mathbf{x}$ be the vector whose jth entry is 1 and others are zero, where j maximizes $||\mathbf{a}_j||_1$, we attain this bound, and thus the matrix norm is

$$||\mathbf{A}||_1 = \max_{1\leq j\leq n} ||\mathbf{a}_j||_1.$$

By much the same argument, it can be shown that the ∞-norm of an $m \times n$ matrix is equal to the 'maximum row sum'

$$||\mathbf{A}||_\infty = \max_{1\leq i\leq m} ||\mathbf{a}_{(i)}||_1.$$

Computing matrix p-norms with $p \neq 1, \infty$ is more difficult.

Then we provide a geometric example of some induced matrix norms. Let's consider the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 0 & 2 \end{bmatrix},$$

which maps $\mathbb{R}^2$ to $\mathbb{R}^2$ by multiplication.

Figure 1 depicts the action of $\mathbf{A}$ on the unit balls of $\mathbb{R}^2$ defined by the 1-, 2- and ∞-norms. Regardless of the norm, $\mathbf{A}$ maps $e_1 = [1, 0]^T$ to the first column of $\mathbf{A}$ itself, and $e_2 = [0, 1]^T$ to the second column of $\mathbf{A}$.

In the 1-norm, the vector whose norm is 1 that is amplified most by $\mathbf{A}$ is $[0, 1]^T$ (or its negative), and the amplification factor is 4. In the ∞-norm, the vector whose norm is 1 that is amplified most by $\mathbf{A}$ is $[1, 1]^T$ (or its negative), and the amplification factor is 3. In the 2-norm, the vector whose norm is 1 that is amplified most by $\mathbf{A}$ is $[0, 1]^T$ is the vector indicated by the dash line in the figure (or its negative), and the amplification factor is approximately 2.9208 (the exact value is $\sqrt{(9 + \sqrt{65})/2}$). We will consider how to get 2-norm again in section 3.3.2.

2.11.2 General Matrix Norms

Matrix norms do not have to be induced by vector norms. In general a matrix norm must merely satisfy the vector norm conditions applied in the mn-dimensional vector space of matrices:

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

The most important matrix norm which is not induced by a vector norm is ***Frobenius norm*** (or F-norm), for $\mathbf{A} \in \mathbb{R}^{m\times n}$, defined by

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

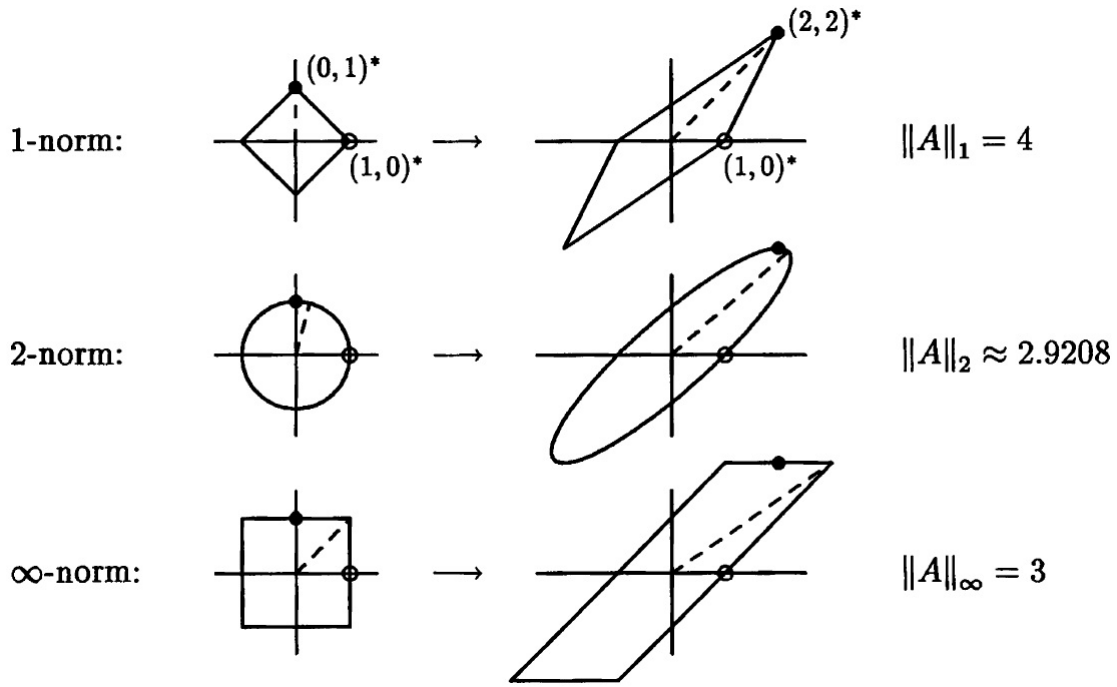

Figure 1: On the left, the unit balls of $\mathbb{R}^2$ with respect to $||\cdot||_1$, $||\cdot||_2$ and $||\cdot||_\infty$. On the right, their images under multiplying the matrix $\mathbf{A}$ on left. Dashed lines mark the vectors that are amplified most by $\mathbf{A}$ in each norm.

The Frobenius norm can be used to bound products of matrices. Let $\mathbf{C} = \mathbf{AB} = (c_{ij}) \in \mathbb{R}^{m\times n}$, by Cauchy-Schwarz inequality we have $c_{ij} \le ||\mathbf{a}_{(i)}||_2||\mathbf{b}_j||_2$. Then we obtain

$$\begin{aligned} ||\mathbf{AB}||_F^2 &= \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}^2 \\ &\le \sum_{i=1}^{m}\sum_{j=1}^{n} (||\mathbf{a}_{(i)}||_2||\mathbf{b}_j||_2)^2 \\ &= \Big(\sum_{i=1}^{m}(||\mathbf{a}_{(i)}||_2^2\Big)\Big(\sum_{j=1}^{n}||\mathbf{b}_j||_2^2\Big) \\ &= ||\mathbf{A}||_F^2||\mathbf{B}||_F^2. \end{aligned}$$

One of the many special properties of Frobenius norm is that, it is invariant under multiplication by orthogonal matrix. The same property holds for matrix 2-norm, that is for any $\mathbf{A} \in \mathbb{R}^{m\times n}$ and orthogonal $\mathbf{Q} \in \mathbb{R}^{m\times m}$, we have

$$||\mathbf{QA}||_F = ||\mathbf{A}||_F, \qquad ||\mathbf{QA}||_2 = ||\mathbf{A}||_2.$$

To the end of this section, consider why Frobenius norm can not be induced by a vector norm?

2.12 The Inverse

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

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

Note that not all matrices have inverses. Non-square matrices, for example, do not have inverses by definition. However, for some square matrices A, it may still be the case that $\mathbf{A}^{-1}$ may not exist. In particular, we say that $\mathbf{A}$ is ***invertible*** or ***non-singular*** if $\mathbf{A}$ exists and ***non-invertible*** or ***singular otherwise***.

In order for a square matrix A to have an inverse $\mathbf{A}^{-1}$, then $\mathbf{A}$ must be full rank. We will soon see that there are many alternative sufficient and necessary conditions, in addition to full rank, for invertibility.

The following are properties of the inverse; all assume that $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ are non-singular and $c \in \mathbb{R}$:

- $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
- $(c\mathbf{A})^{-1} = c^{-1}\mathbf{A}^{-1}$
- $(\mathbf{A}^{-1})^T = (\mathbf{A}^T)^{-1}$
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
- $\mathbf{A}^{-1} = \mathbf{A}^T$ if $\mathbf{A} \in \mathbb{R}^{n\times n}$ is orthogonal matrix.

Additionally, if all the necessary inverse exist, then 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 $\mathbf{D} \in \mathbb{R}^{p\times n}$, the Woodbury matrix identity says that

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

This conclusion is very useful to compute inverse of matrix more efficiently when $\mathbf{A} = \mathbf{I}$ and $n \gg p$.

For non-singular matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and vectors $\mathbf{x}, \mathbf{b}, \in \mathbb{R}^n$, when writing the product $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ it is important not to let the inverse matrix notation obscure what is really going on! Rather than thinking of $\mathbf{x}$ as the result of applying $\mathbf{A}^{-1}$ to $\mathbf{b}$, we should understand it as the unique vector that satisfies the equation $\mathbf{Ax} = \mathbf{b}$. This means that $\mathbf{x}$ is the vector of coefficients of the unique linear expansion of $\mathbf{b}$ in the basis of columns of $\mathbf{A}$. Multiplication by $\mathbf{A}^{-1}$ is a change of basis operation as Figure 2 displayed.

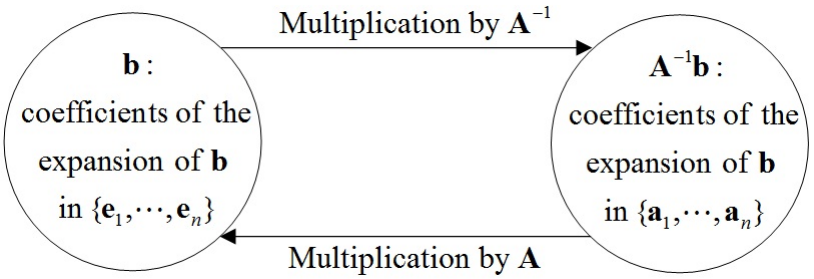

Figure 2: The vector $\mathbf{e}_i$ represent the vector in $\mathbb{R}^n$ whose ith entry is 1 and the others are 0.

2.13 The Determinant

The ***determinant*** of a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, is a function $\det : \mathbb{R}^{n\times n} \to \mathbb{R}$ and is denoted $\det(A)$ or $|A|$. (like the trace operator, we usually omit parentheses). Algebraically, one could write down an explicit formula for the determinant of A, but this unfortunately gives little intuition about its meaning. Instead, well start out by providing a geometric interpretation of the determinant and then visit some of its specific algebraic properties afterwards. Given a square matrix

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

consider the set of points $\mathcal{S} \subset R^n$ formed by taking all possible linear combinations of the row vectors, where the coefficients of the linear combination are all between 0 and 1; that is, the set $\mathcal{S}$ is the restriction of $\text{span}(\{\mathbf{a}_1, ..., \mathbf{a}_n\})$ to only those linear combinations whose coefficients satisfy $0 \le \alpha_i \le 1, i = 1, \ldots, n$. Formally, we have

$$\mathcal{S} = \{v \in \mathbb{R}^n : v = \sum_{i=1}^{n} \alpha_i \mathbf{a}_i, \text{ where } 0 \le \alpha_i \le 1, i = 1, \ldots, n\}.$$

The absolute value of the determinant of $\mathbf{A}$, it turns out, is a measure of the 'volume' [2] of the set $\mathcal{S}$.

For example, consider the 2×2 matrix,

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix}.$$

[2] Admittedly, we have not actually defined what we mean by 'volume' here, but hopefully the intuition should be clear enough. When $n = 2$, our notion of 'volume' corresponds to the area of $\mathcal{S}$ in the Cartesian plane. When $n = 3$, 'volume' corresponds with our usual notion of volume for a three-dimensional object.

Here, the rows of the matrix are

$$\mathbf{a}_1 = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \quad \mathbf{a}_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}.$$

The set $\mathcal{S}$ corresponding to these rows is shown in Figure 3. For two-dimensional matrices, *setS* generally has the shape of a parallelogram. In our example, the value of the determinant is $\det(\mathbf{A}) = -7$ (as can be computed using the formulas shown later in this section), so the area of the parallelogram is 7.

In three dimensions, the set $\mathcal{S}$ corresponds to an object known as a parallelepiped which is a three-

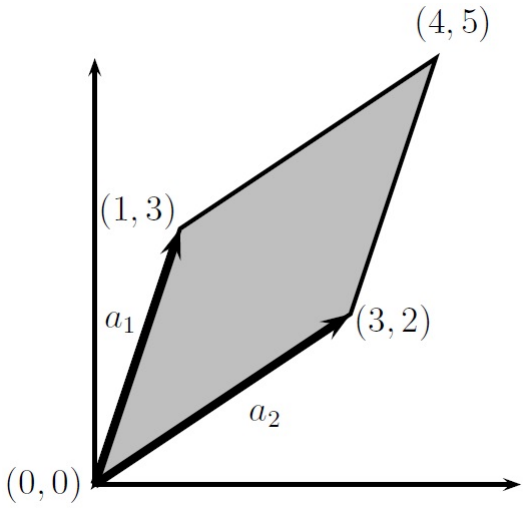

Figure 3: Illustration of the determinant for the matrix $\mathbf{A}$. Here, $\mathbf{a}_1$ and $\mathbf{a}_2$ are vectors corresponding to the rows of $\mathbf{A}$, and the set $\mathcal{S}$ corresponds to the shaded region (i.e., the parallelogram). The absolute value of the determinant, $|\det(A)|$, is the area of the parallelogram.

dimensional box with skewed sides, such that every face has the shape of a parallelogram and the trivial example is a cubic. The absolute value of the determinant of the matrix whose rows define $\mathcal{S}$ give the three-dimensional volume of the parallelepiped. In even higher dimensions, the set $\mathcal{S}$ is an object known as an n-dimensional parallelotope.

Algebraically, the determinant satisfies the following three properties (from which all other properties follow, including the general formula):

- The determinant of the identity is 1, $\det(\mathbf{I}) = 1$. (Geometrically, the volume of a unit hypercube is 1).
- Given a matrix $A \in \mathbb{R}^{n \times n}$, if we multiply a single row in $\mathbf{A}$ by a scalar $t \in \mathbb{R}$, then the determinant of the new matrix is $t \det(\mathbf{A})$. (Geometrically, multiplying one of the sides of the set *setS* by a factor t causes the volume to increase by a factor t.)
- If we exchange any two rows $\mathbf{a}_{(i)}^T$ and $\mathbf{a}_{(j)}^T$ of A, then the determinant of the new matrix is $\det(\mathbf{A})$.
- For $\mathbf{A} \in \mathbb{R}^{n \times n}$ is triangular, $\det(\mathbf{A}) = \prod_{i=1}^{n} a_{ii}$
- For square sub-matrices $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{p \times p}$ and $\mathbf{C} \in \mathbb{R}^{n \times p}$,

$$\det\left(\begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{0} & \mathbf{B} \end{bmatrix}\right) = \det(\mathbf{A}) \det(\mathbf{B}).$$

- For $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\det(\mathbf{A}) = \det(\mathbf{A}^T)$.
- For $\mathbf{A} \in \mathbb{R}^{n \times n}$ is orthogonal, $\det(\mathbf{A}) = \pm 1$.
- For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$, $\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$.
- For $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\det(\mathbf{A}) = 0$ if and only if $\mathbf{A}$ is singular. (If A is singular then it does not have full rank, and hence its columns are linearly dependent. In this case, the set $\mathcal{S}$ corresponds to a 'flat sheet' within the n-dimensional space and hence has zero volume.)

Before giving the general definition for the determinant, we define, for $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{A}_{\setminus i,\setminus j} \in \mathbb{R}^{(n-1)\times(n-1)}$ to be the matrix that results from deleting the ith row and jth column from $\mathbf{A}$. The general (recursive) formula for the determinant is

$$\begin{aligned}\det(\mathbf{A}) &= \sum_{i=1}^{n}(-1)^{i+j}a_{ij}\det(\mathbf{A}_{\setminus i,\setminus j}) \quad (\text{for any } j \in \{1,2,\ldots,n\}) \\ &= \sum_{j=1}^{n}(-1)^{i+j}a_{ij}\det(\mathbf{A}_{\setminus i,\setminus j}) \quad (\text{for any } i \in \{1,2,\ldots,n\})\end{aligned}$$

with the initial case that $\det(\mathbf{A}) = a_{11}$ for $\mathbf{A} \in \mathbb{R}^{1\times 1}$. If we were to expand this formula completely for $\mathbf{A} \in \mathbb{R}^{n\times n}$, there would be a total of n (n factorial) different terms. For this reason, we hardly ever explicitly write the complete equation of the determinant for matrices bigger than 3×3. The classical adjoint (often just called the adjoint) of a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is denoted $\text{adj}(\mathbf{A})$, and defined as

$$\text{adj}(\mathbf{A}) \in \mathbb{R}^{n\times n}, \quad \big(\text{adj}(\mathbf{A})\big)_{ij} = (-1)^{i+j}\det(\mathbf{A}_{\setminus j,\setminus i})$$

(note the switch in the indices $\mathbf{A}_{\setminus j,\setminus i}$). It can be shown that for any nonsingular $A \in \mathbb{R}^{n\times n}$,

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}\text{adj}(\mathbf{A}).$$

2.14 Schur Complement

Suppose $\mathbf{A} \in \mathbb{R}^{p\times p}$, $\mathbf{B} \in \mathbb{R}^{p\times q}$, $\mathbf{C} \in \mathbb{R}^{q\times p}$, $\mathbf{D} \in \mathbb{R}^{q\times q}$, and $\mathbf{D}$ is invertible. Let

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

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

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

This is analogous to an ***LDU decomposition***. That is, we have shown that

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

and the he inverse of $\mathbf{M}$ may be expressed involving $\mathbf{D}^{-1}$ and the inverse of Schur complement (if it exists) such as

$$\begin{aligned}\begin{bmatrix}\mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D}\end{bmatrix}^{-1} &= \begin{bmatrix}\mathbf{I}_p & \mathbf{0} \\ -\mathbf{D}^{-1}\mathbf{C} & \mathbf{I}_q\end{bmatrix}\begin{bmatrix}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{I}_p & -\mathbf{B}\mathbf{D}^{-1} \\ \mathbf{0} & I_q\end{bmatrix} \\ &= \begin{bmatrix}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})\mathbf{B}\mathbf{D}^{-1}\end{bmatrix}.\end{aligned}$$

Moreover, the determinant of $\mathbf{M}$ is also clearly seen to be given by

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

3 Eigenvalues and Eigenvectors

3.1 Definitions

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 ***eigenvector*** if

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

Intuitively, this definition means that multiplying $\mathbf{A}$ by the vector $\mathbf{x}$ results in a new vector that points in the same direction as $\mathbf{x}$, but scaled by a factor λ. Also note that for any eigenvector $\mathbf{x} \in \mathbb{C}^n$, and

scalar $t \in \mathbb{C}, \mathbf{A}(c\mathbf{x}) = c\mathbf{A}\mathbf{x} = c\lambda\mathbf{x} = \lambda(c\mathbf{x})$, so $c\mathbf{x}$ is also an eigenvector. And if $\mathbf{x}$ and $\mathbf{y}$ are eigenvectors for λ_j and $\alpha \in \mathbb{R}$, then $\mathbf{x}+\mathbf{y}$ and $\alpha\mathbf{x}$ are also eigenvectors for λ_i. Thus, the set of all eigenvectors for λ_j forms a subspace which is called ***eigenspace*** of $\mathbf{A}$ for λ_i.

For this reason when we talk about 'the' eigenvector associated with λ, we define the standardized eigenvector which are normalized to have length 1 (this still creates some ambiguity, since $\mathbf{x}$ and $-\mathbf{x}$ will both be eigenvectors, but we will have to live with this). Sometimes we also use the term 'eigenvector' to refer the ***standardized eigenvector***.

We can rewrite the equation above to state that $(\lambda, \mathbf{x})$ is an eigenvalue-eigenvector pair of $\mathbf{A}$ if,

$$(\lambda\mathbf{I} - \mathbf{A})\mathbf{x} = \mathbf{0}, \quad \mathbf{x} \neq \mathbf{0}.$$

$(\lambda\mathbf{I} - \mathbf{A})\mathbf{x} = \mathbf{0}$ has non-zero solution to $\mathbf{x}$ if and only if $(\lambda\mathbf{I} - \mathbf{A})\mathbf{x}$ has a non-zero dimension nullspace. Considering that Table 3 implies $\text{rank}\Big(\mathcal{N}(\lambda\mathbf{I} - \mathbf{A})\Big) = n - \text{rank}(\lambda\mathbf{I} - \mathbf{A})$, which means such $\mathbf{x}$ exists only the case if $(\lambda\mathbf{I} - \mathbf{A})$ is singular, that is

$$\det(\lambda\mathbf{I} - \mathbf{A}) = 0.$$

We can now use the previous definition of the determinant to expand this expression into a (very large) ***characteristic polynomial*** in λ, defined as

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

where λ will have maximum degree n. We can find the n roots (possibly complex) of $p_{\mathbf{A}}(\lambda)$ to find the n eigenvalues $\lambda_1, \ldots, \lambda_n$. To find the eigenvector corresponding to the eigenvalue λ_i, we simply solve the linear equation $(\lambda\mathbf{I} - \mathbf{A})\mathbf{x} = 0$. It should be noted that this is not the method which is actually used in practice to numerically compute the eigenvalues and eigenvectors (remember that the complete expansion of the determinant has $n!$ terms), although it maybe works in the linear algebra exam you have taken.

3.2 General Results

The following are properties of eigenvalues and eigenvectors (in all cases assume $\mathbf{A} \in \mathbb{R}^{n\times n}$ has eigenvalues $\lambda_i, \ldots, \lambda_n$ and associated eigenvectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$):

- The trace of a $\mathbf{A}$ is equal to the sum of its eigenvalues,

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{n}\lambda_i.$$

- The determinant of $\mathbf{A}$ is equal to the product of its eigenvalues,

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

- Let $\mathbf{C} \in \mathbb{R}^{n\times n}$ be a non-singular matrix, then

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \det(\mathbf{C})\det(\mathbf{A} - \lambda\mathbf{I}\mathbf{C}^{-1}\mathbf{C})\det(\mathbf{C}^{-1}) = \det(\mathbf{C}\mathbf{A}\mathbf{C}^{-1} - \lambda\mathbf{I}).$$

  Thus $\mathbf{A}$ and $\mathbf{C}\mathbf{A}\mathbf{C}^{-1}$ have the same eigenvalues and we call they are similar. Further, if $\mathbf{x}_i$ is an eigenvector of $\mathbf{A}$ for λ_i then $\mathbf{C}\mathbf{x}_i$ is an eigenvector of $\mathbf{C}\mathbf{A}\mathbf{C}^{-1}$ for λ_i.

- Let λ_1 denote any particular eigenvalue of $\mathbf{A}$, with eigenspace $\mathcal{W}$ of dimension r (called ***geometric multiplicity***). If k denotes the multiplicity of λ_1 in as a root of $p_A(\lambda)$ (called ***algebra multiplicity***), then $1 \leq r \leq k$

To prove the last conclusion, let $\mathbf{e}_1, \ldots, \mathbf{e}_r$ be an orthogonal basis of $\mathcal{W}$ and extend it to an orthogonal basis of $\mathbb{R}^n$, $\mathbf{e}_1, \ldots, \mathbf{e}_r, \mathbf{f}_1, \ldots, \mathbf{f}_{n-r}$. Write $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_r]$ and $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_{n-r}]$. Then $[\mathbf{E}, \mathbf{F}]$ is an orthogonal matrix so that

$$\begin{aligned}
\mathbf{I}_n &= [\mathbf{E},\mathbf{F}][\mathbf{E},\mathbf{F}]^T = \mathbf{E}\mathbf{E}^T + \mathbf{F}\mathbf{F}^T, \\
\det([\mathbf{E},\mathbf{F}][\mathbf{E},\mathbf{F}]^T) &= \det([\mathbf{E},\mathbf{F}])\det([\mathbf{E},\mathbf{F}]^T) = 1, \\
\mathbf{E}^T\mathbf{A}\mathbf{E} &= \lambda_1\mathbf{E}^T\mathbf{E} = \lambda_1\mathbf{I}_r, \\
\mathbf{F}^T\mathbf{F} &= \mathbf{I}_{n-r}, \\
\mathbf{F}^T\mathbf{A}\mathbf{E} &= \lambda_1\mathbf{F}^T\mathbf{E} = \mathbf{0}.
\end{aligned}$$

Thus

$$\begin{aligned} p_{\mathbf{A}}(\lambda) &= \det(\mathbf{A}-\lambda\mathbf{I}_n) = \det([\mathbf{E},\mathbf{F}]^T)\det(\mathbf{A}-\lambda\mathbf{I}_n)\det([\mathbf{E},\mathbf{F}]) \\ &= \det([\mathbf{E},\mathbf{F}]^T(\mathbf{A}\mathbf{E}\mathbf{E}^T+\mathbf{A}\mathbf{F}\mathbf{F}^T-\lambda\mathbf{E}\mathbf{E}^T-\lambda\mathbf{F}\mathbf{F}^T)[\mathbf{E},\mathbf{F}]) \\ &= \det\left(\begin{bmatrix}(\lambda_1-\lambda)\mathbf{I}_r & \mathbf{E}^T\mathbf{A}\mathbf{F} \\ \mathbf{0} & \mathbf{F}^T\mathbf{A}\mathbf{F}-\lambda\mathbf{I}_{n-r}\end{bmatrix}\right) \\ &= (\lambda_1-\lambda)^r p_{\mathbf{A}_1}(\lambda), \end{aligned}$$

then the multiplicity of λ_1 as a root of $p_{\mathbf{A}}(\lambda)$ is at least r. In section 3.3.1, we will show that if $\mathbf{A}$ is symmetric then $r=k$. However, if $\mathbf{A}$ is not symmetric, it is possible that $r<k$, For example,

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

has eigenvalue 0 with multiplicity 2 but the corresponding eigenspace which is generated by $[1,0]^T$ only has dimension 1.

3.3 Spectral Decomposition

3.3.1 Spectral Decomposition Theorem

Two remarkable properties come about when we look at the eigenvalues and eigenvectors of symmetric matrix $\mathbf{A}\in\mathbb{S}^n$. First, it can be shown that all the eigenvalues of $\mathbf{A}$ are real. Let an eigenvalue-eigenvector pair $(\lambda,\mathbf{x})$ of $\mathbf{A}$, where $\mathbf{x}=\mathbf{y}+i\mathbf{z}$ and $\mathbf{y},\mathbf{z}\in\mathbb{R}^n$. Then we have

$$\begin{aligned} & \mathbf{A}\mathbf{x}=\lambda\mathbf{x} \\ \Longrightarrow\quad & \mathbf{A}(\mathbf{y}+i\mathbf{z})=\lambda(\mathbf{y}+i\mathbf{z}) \\ \Longrightarrow\quad & (\mathbf{y}-i\mathbf{z})^T\mathbf{A}(\mathbf{y}+i\mathbf{z})=\lambda(\mathbf{y}-i\mathbf{z})^T(\mathbf{y}+i\mathbf{z}) \\ \Longrightarrow\quad & \mathbf{y}^T\mathbf{A}\mathbf{y}-i\mathbf{z}^T\mathbf{A}\mathbf{y}+i\mathbf{y}^T\mathbf{A}\mathbf{z}+\mathbf{z}^T\mathbf{A}\mathbf{z}=\lambda(\mathbf{y}^T\mathbf{y}+\mathbf{z}^T\mathbf{z}) \\ \Longrightarrow\quad & \lambda=\frac{\mathbf{y}^T\mathbf{A}\mathbf{y}+\mathbf{z}^T\mathbf{A}\mathbf{z}}{\mathbf{y}^T\mathbf{y}+\mathbf{z}^T\mathbf{z}}\in\mathbb{R}. \end{aligned}$$

Secondly, two eigenvectors corresponding to distinct eigenvalues of $\mathbf{A}$ are orthogonal. In other words if $\lambda_i\neq\lambda_j$ are eigenvalues of $\mathbf{A}$ with eigenvector $\mathbf{x}_i$ and $\mathbf{x}_j$ respectively, then

$$\lambda_j\mathbf{x}_i^T\mathbf{x}_j=\mathbf{x}_i^T\mathbf{A}\mathbf{x}_j=\mathbf{x}_j^T\mathbf{A}\mathbf{x}_i=\lambda_i\mathbf{x}_j^T\mathbf{x}_i,$$

so that $\mathbf{x}_i^T\mathbf{x}_j=0$, which means the eigenspaces of distinct eigenvalues are orthogonal.

Furthermore, if λ_i is an eigenvector of $\mathbf{A}$ with $m\geq 2$ algebra multiplicity, we can find m orthogonal eigenvectors in its eigenspace. Then we prove this result. Let $\mathbf{x}_i$ be one of eigenvector with λ_i and extend $\mathbf{x}_i$ to an orthogonal bias $\mathbf{x}_i,\mathbf{y}_2,\ldots,\mathbf{y}_n$. and denote $\mathbf{B}=[\mathbf{x}_i,\mathbf{Y}]$, where $\mathbf{Y}=[\mathbf{y}_2,\ldots,\mathbf{y}_n]$. Then $\mathbf{B}$ is orthogonal and $\mathbf{Y}$ is column orthonormal, and we have

$$\mathbf{B}^T\mathbf{A}\mathbf{B}=\begin{bmatrix}\lambda_i & \mathbf{0}\\ \mathbf{0} & \mathbf{Y}^T\mathbf{A}\mathbf{Y}\end{bmatrix}$$

and

$$\begin{aligned} \det(\lambda\mathbf{I}_n-\mathbf{A}) &= \det(\mathbf{B}^T)\det(\lambda\mathbf{I}_n-\mathbf{A})\det(\mathbf{B}) \\ &= \det(\lambda\mathbf{I}_n-\mathbf{B}^T\mathbf{A}\mathbf{B}) \\ &= (\lambda-\lambda_i)\det(\lambda\mathbf{I}_{n-1}-\mathbf{Y}^T\mathbf{A}\mathbf{Y}) \end{aligned}$$

Since $m\geq 2$, $\det(\lambda_i\mathbf{I}_{n-1}-\mathbf{Y}^T\mathbf{A}\mathbf{Y})$ must be zero, which means λ_i is an eigenvalue of $\mathbf{Y}^T\mathbf{A}\mathbf{Y}$. Suppose $\mathbf{x}_{i2}$ is the eigenvector of $\mathbf{Y}^T\mathbf{A}\mathbf{Y}$ with λ_i, then $\mathbf{Y}^T\mathbf{A}\mathbf{Y}\mathbf{x}_{i2}=\lambda_2\mathbf{x}_{i2}$. Hence, we have

$$\mathbf{B}^T\mathbf{A}\mathbf{B}\begin{bmatrix}0\\ \mathbf{x}_{i2}\end{bmatrix}=\begin{bmatrix}\lambda_i & \mathbf{0}\\ \mathbf{0} & \mathbf{Y}^T\mathbf{A}\mathbf{Y}\end{bmatrix}\begin{bmatrix}0\\ \mathbf{x}_{i2}\end{bmatrix}=\lambda_i\begin{bmatrix}0\\ \mathbf{x}_{i2}\end{bmatrix}$$

$$\Longrightarrow\quad \mathbf{A}\mathbf{B}\begin{bmatrix}0\\ \mathbf{x}_{i2}\end{bmatrix}=\lambda_i\mathbf{B}\begin{bmatrix}0\\ \mathbf{x}_{i2}\end{bmatrix}$$

which implies

$$\mathbf{x}_j = \mathbf{B}\begin{bmatrix} 0 \\ \mathbf{x}_{i2} \end{bmatrix}$$

another eigenvectors of $\mathbf{A}$ and $\mathbf{x}_i^T\mathbf{x}_j = 0$. Repeat this type of procedure by induction with $\mathbf{B}_2 = [\mathbf{x}_i, \mathbf{x}_j, \mathbf{y}_3, \ldots, \mathbf{y}_n]$, we can find m m orthogonal eigenvectors in this eigenspace.

Based on previous results, we can derive ***spectral decomposition theorem*** that any symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ can be written as

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

where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix of eigenvalues of $\mathbf{A}$, and $\mathbf{X}$ is an orthogonal matrix whose columns are corresponding standardized eigenvectors.

3.3.2 Applications

An application where eigenvalues and eigenvectors come up frequently is in maximizing some function of a matrix. In particular, for a matrix $\mathbf{A} \in \mathbb{S}^n$, consider the following maximization problem,

$$\max_{\mathbf{x}\in\mathbb{R}^n} \mathbf{x}^T\mathbf{A}\mathbf{x} \qquad \text{subject to } ||\mathbf{x}||_2^2 = 1$$

i.e., we want to find the vector (of ℓ_2-norm 1) which maximizes the objective function (which is a quadratic form). Assuming the eigenvalues of $\mathbf{A}$ are ordered as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, the optimal $\mathbf{x}$ for this optimization problem is $\mathbf{x}_1$, the eigenvector corresponding to λ_1. In this case the maximal value of the quadratic form is λ_1, also called the spectral radius of $\mathbf{A}$ and denoted as $\rho(\mathbf{A})$, which satisfied $\rho(\mathbf{A}) \leq ||\mathbf{A}||$ for any induced matrix norm.

Additionally, the solution of above maximization problem indicate us how to compute the 2-norm of a matrix. Recall that for any $\mathbf{B} \in \mathbb{R}^{m\times n}$, we can define its 2-norm as

$$||\mathbf{B}||_2 = \sup_{\mathbf{x}\in\mathbb{R}^n, ||\mathbf{x}||_2=1} ||\mathbf{B}\mathbf{x}||_2.$$

and consider when $\mathbf{A} = \mathbf{B}^T\mathbf{B} \in \mathbb{S}^n$. The solution of this optimization problem can be proved by appealing to the eigenvector-eigenvalue form of $\mathbf{A}$ and the properties of orthogonal matrices. However, in section 6.7 we will see a way of showing it directly using matrix calculus.

4 Singular Value Decomposition

4.1 Definitions and Basic Properties

For general matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, we can use spectral decomposition theorem to derive that if $\text{rank}(\mathbf{A}) = r$, then $\mathbf{A}$ can be written as ***reduced singular value decomposition*** (SVD)

$$\mathbf{A} = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^T.$$

When $m \geq n$, $\widehat{\mathbf{U}} \in \mathbb{R}^{m\times n}$ is column orthonoromal and $\widehat{\mathbf{V}} \in \mathbb{R}^{n\times n}$ are orthonoromal matrices and $\widehat{\mathbf{\Sigma}} = \text{diag}(\sigma_1(\mathbf{A}), \ldots, \sigma_n(\mathbf{A}))$ is a diagonal matrix with positive elements where $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}), \ldots, \geq \sigma_n(\mathbf{A}) \geq 0$, called singular values. The result of SVD for general matrix $\mathbf{A}$ can be derived by using spectral decomposition on $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ (consider their eigenvalues are non-negative, we can get singular values by the square roots of the eigenvalues of $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$). Furthermore, we have $\sigma_1(\mathbf{A}) = ||\mathbf{A}||_2$.

By adjoining additional orthonomal columns, $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ can be extended to a orthogonal matrix if they are not. Accordingly, let $\mathbf{\Sigma}$ be the $m \times n$ matrix consisting of $\widehat{\mathbf{\Sigma}}$ in upper $n \times n$ block together with zeros for other entries. We now have a new factorization, the ***full SVD*** of $\mathbf{A}$.

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

where $\mathbf{\Sigma}_{ii} = \sigma_i(\mathbf{A}) = 0$ for $i > r$.

For following properties, we assume that $\mathbf{A} \in \mathbb{R}^{m\times n}$. Let p be the minimum of m and n, let $r < p$ denote the number of nonzero singular values of $\mathbf{A}$

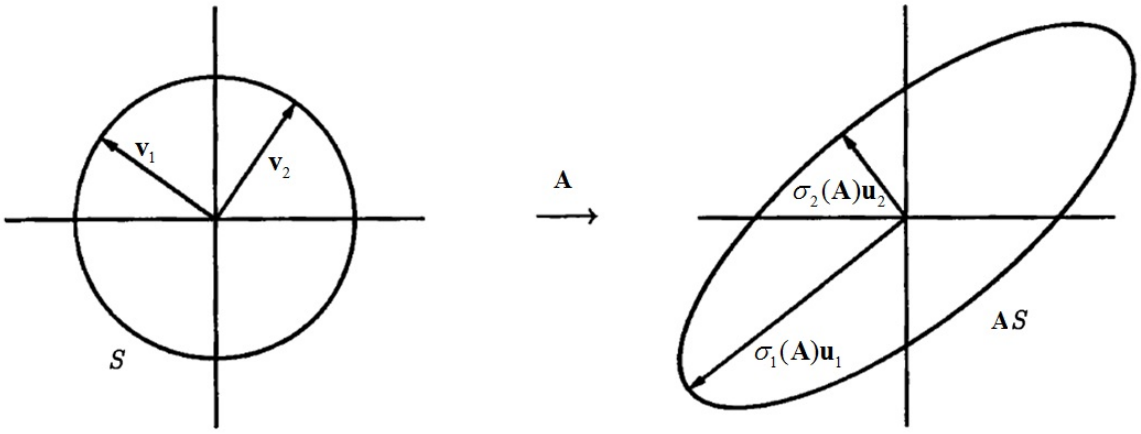

Figure 4: SVD of $\mathbf{A} \in \mathbb{R}^{2\times 2}$ matrix.

- $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{\Sigma}) = r$.
- $\mathcal{R}(\mathbf{A}) = \text{span}(\{\mathbf{u}_1, \ldots, \mathbf{u}_r\})$ and $\mathcal{N}(\mathbf{A}) = \text{span}(\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\})$ (in Table 3).
- $||\mathbf{A}||_2 = \sigma_1(\mathbf{A})$ and $||\mathbf{A}||_F = \sqrt{\sigma_1(\mathbf{A})^2 + \sigma_2(\mathbf{A})^2 + \cdots + \sigma_r(\mathbf{A})^2}$.

The SVD is motivated by the geometric fact that the image of the unit sphere under $m\times n$ is a hyperellipse. Hyperellipse is just the m-dimensional generalization of an ellipse. We may define a hyperellipse in $\mathbb{R}^m$ by some factors $\sigma_1, \ldots, \sigma_m$ (possible 0) in some orthogonal directions $\mathbf{u}_1, \ldots, \mathbf{u}_m \in \mathbb{R}^m$ and let $||\mathbf{u}_i||_2 = 1$. Let $\mathcal{S}$ be the unit sphere in $\mathbb{R}^n$, and take any $\mathbf{A} \in \mathbb{R}^{m\times n}$ with $m \geq n$ and has the SVD $\mathbf{A} = \mathbf{U\Sigma V}^T$. The Figure 4 shows the transformation of $\mathcal{S}$ in two dimension by 2×2 matrix $\mathbf{A}$.

4.2 Low-Rank Approximations

SVD can also be written as a sum of rank-one matrices, for $\mathbf{A} \in \mathbb{R}^{m\times n}$ has rank r and SVD $\mathbf{A} = \mathbf{U\Sigma V}^T$,

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i(\mathbf{A})\mathbf{u}_i\mathbf{v}_i^T.$$

This form has a deeper property: the kth partial sum capture as much of the energy of $\mathbf{A}$ as possible. This statement holds with 'energy' defined by either the 2-norm or the Frobenius norm. We can make it precise by formulating a problem of best approximation of a matrix $\mathbf{A}$ of lower rank. That is for any k with $0 \leq k \leq r$, define

$$\mathbf{A}_k = \sum_{i=1}^{k} \sigma_i(\mathbf{A})\mathbf{u}_i\mathbf{v}_i^T;$$

if $k = p = \min\{m, n\}$, define $\sigma_{k+1}(\mathbf{A}) = 0$. Then

$$\begin{aligned} ||\mathbf{A} - \mathbf{A}_k||_2 &= \inf_{\mathbf{B}\in\mathbb{R}^{m\times n}, \text{rank}(\mathbf{B})\leq k} ||\mathbf{A} - \mathbf{B}||_2 = \sigma_{k+1}(\mathbf{A}), \\ ||\mathbf{A} - \mathbf{A}_k||_F &= \inf_{\mathbf{B}\in\mathbb{R}^{m\times n}, \text{rank}(\mathbf{B})\leq k} ||\mathbf{A} - \mathbf{B}||_F = \sqrt{\sigma_{k+1}^2 +, \ldots, +\sigma_r^2}. \end{aligned}$$

In detial, for Frobenius norm, if it can be shown that for arbitrary $\mathbf{x}_i \in \mathbb{R}^m, \mathbf{y}_i \in \mathbb{R}^n$

$$\left\|\mathbf{A} - \sum_{i=1}^{k} \mathbf{x}_i\mathbf{y}_i^T\right\|_F^2 \geq ||\mathbf{A}_k||_F^2 = ||\mathbf{A}||_F^2 - \sum_{i=1}^{k} \sigma_i^2(\mathbf{A})$$

then $\mathbf{A}_k$ will be the desired approximation.

Without loss of generality we may assume that the vector $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are orthonormal. For if they are not, we can use Gram-Schmidt orthogonalization (not included in this review) to express them as linear

combinations of orthonormal vectors, substitute these expressions in $\sum_{i=1}^k \mathbf{x}_i\mathbf{y}_i^T$, and collect therms in the new vectors. Now

$$\begin{aligned}\left\|\mathbf{A}-\sum_{i=1}^k \mathbf{x}_i\mathbf{y}_i^T\right\|_F^2 &= \operatorname{tr}\left(\left(\mathbf{A}-\sum_{i=1}^k \mathbf{x}_i\mathbf{y}_i^T\right)^T\left(\mathbf{A}-\sum_{i=1}^k \mathbf{x}_i\mathbf{y}_i^T\right)\right)\\ &= \operatorname{tr}\left(\mathbf{A}\mathbf{A}^T+\sum_{i=1}^k(\mathbf{y}_i-\mathbf{A}^T\mathbf{x}_i)(\mathbf{y}_i-\mathbf{A}^T\mathbf{x}_i)^T-\sum_{i=1}^k \mathbf{A}^T\mathbf{x}_i\mathbf{x}_i\mathbf{A}\right).\end{aligned}$$

Since $\operatorname{tr}((\mathbf{y}_i-\mathbf{A}^T\mathbf{x}_i)(\mathbf{y}_i-\mathbf{A}^T\mathbf{x}_i)^T)\geq 0$ and $\operatorname{tr}(\mathbf{A}\mathbf{x}_i\mathbf{x}_i^T\mathbf{A}^T)=||\mathbf{A}\mathbf{x}_i||_F^2$, the result will be established if it can be shown that

$$\sum_{i=1}^k ||\mathbf{A}\mathbf{x}_i||_F^2 \leq \sum_{i=1}^k \sigma_i^2(\mathbf{A}).$$

Let $\mathbf{V}=[\mathbf{V}_1,\mathbf{V}_2]$, where $\mathbf{V}_1$ has k columns, and let $\mathbf{\Sigma}=\operatorname{diag}(\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)$ be conformal partition of $\mathbf{\Sigma}$. Then

$$\begin{aligned}||\mathbf{A}\mathbf{x}_i||_F^2 &= \sigma_k^2(\mathbf{A})+\big(||\mathbf{\Sigma}_1\mathbf{V}_1^T\mathbf{x}_i||_2^2-\sigma_k^2(\mathbf{A})||\mathbf{V}_1^T\mathbf{x}_i||_2^2\big)\\ &\quad -\big(\sigma_k^2(\mathbf{A})||\mathbf{V}_2^T\mathbf{x}_i||_2^2-||\mathbf{\Sigma}_2\mathbf{V}_2^T\mathbf{x}_i||_2^2\big)\\ &\quad -\sigma_k^2(\mathbf{A})\big(1-||\mathbf{V}^T\mathbf{x}_i||_2^2\big).\end{aligned}$$

Now the last two terms in above equation are clearly nonnegative. Hence

$$\begin{aligned}\sum_{i=1}^k ||\mathbf{A}\mathbf{x}_i||_2^2 &\leq k\sigma_k^2(\mathbf{A})+\sum_{i=1}^k\big(||\mathbf{\Sigma}_1\mathbf{V}_1^T\mathbf{x}_i||_2^2-\sigma_k^2(\mathbf{A})||\mathbf{V}_1^T\mathbf{x}_i||_2^2\big)\\ &= k\sigma_k^2(\mathbf{A})+\sum_{i=1}^k\sum_{j=1}^k\big(\sigma_j^2(\mathbf{A})-\sigma_k^2(\mathbf{A})\big)||\mathbf{v}_j^T\mathbf{x}_i||_2^2\\ &= \sum_{j=1}^k\left(\sigma_k^2(\mathbf{A})+\big(\sigma_j^2(\mathbf{A})-\sigma_k^2(\mathbf{A})\big)\sum_{i=1}^k||\mathbf{v}_j^T\mathbf{x}_i||_2^2\right)\\ &\leq \sum_{j=1}^k\left(\sigma_k^2(\mathbf{A})+\big(\sigma_j^2(\mathbf{A})-\sigma_k^2(\mathbf{A})\big)\right)\\ &= \sum_{j=1}^k\sigma_j^2(\mathbf{A}),\end{aligned}$$

which establishes the result.

For 2-norm, we can prove by contradiction. Suppose there is some $\mathbf{B}$ with $\operatorname{rank}(\mathbf{B})\leq k$ such that $||\mathbf{A}-\mathbf{B}||_2<||\mathbf{A}-\mathbf{A}_k||_2=\sigma_{k+1}(\mathbf{A})$. Then $\operatorname{rank}(\mathcal{N}(\mathbf{B}))=n-\operatorname{rank}(\mathbf{B})\geq n-k$, hence there is an $(n-k)$-dimensional subspace $\mathcal{W}\subset\mathbb{R}^n$ such that $\mathbf{w}\in\mathcal{W}\Rightarrow\mathbf{B}\mathbf{w}=\mathbf{0}$. Accordingly, for any $\mathbf{w}\in\mathbf{W}$, we have $\mathbf{A}\mathbf{w}=(\mathbf{A}-\mathbf{B})\mathbf{w}$ and

$$||\mathbf{A}\mathbf{w}||_2=||(\mathbf{A}-\mathbf{B})\mathbf{w}||_2\leq||\mathbf{A}-\mathbf{B}||_2||\mathbf{w}||_2<\sigma_{k+1}(\mathbf{A})||\mathbf{w}||_2.$$

Thus $\mathcal{W}$ is and an $(n-k)$-dimensional subspace such that $\mathbf{w}\in\mathcal{W}\Rightarrow||\mathbf{A}\mathbf{w}||_2<\sigma_{k+1}(\mathbf{A})||\mathbf{w}||_2$. But there is a $(k+1)$-dimensional subspace where for any $\mathbf{w}$ of it has $||\mathbf{A}\mathbf{w}||_2\geq\sigma_{k+1}(\mathbf{A})||\mathbf{w}||_2$, namely the space spanned by the first $k+1$ columns of $\mathbf{V}$, since for any $\mathbf{w}=\sum_{i=1}^{k+1}c_i\mathbf{v}_i$, we have $||\mathbf{w}||_2^2=\sum_{i=1}^{k+1}c_i^2$ and

$$||\mathbf{A}\mathbf{w}||_2=\left\|\sum_{j=1}^{k+1}c_i\mathbf{A}\mathbf{v}_i\right\|_2=\left\|\sum_{j=1}^{k+1}c_i\sigma_i(\mathbf{A})\mathbf{u}_i\right\|_2\geq\sigma_{k+1}(\mathbf{A})\left\|\sum_{j=1}^{k+1}c_i\mathbf{u}_i\right\|_2=\sigma_{k+1}(\mathbf{A})||\mathbf{w}||_2.$$

Since the sum of the dimensions of these spaces exceeds n, there must be a nonzero vector lying in both, and this is a contradiction.

4.3 SVD vs. Spectral Decomposition

The SVD makes it possible for us to say that every matrix is diagonal if only one uses the proper bases for the domain and range spaces.

Here is how the change of bases works. For $\mathbf{A} \in \mathbb{R}^{m\times n}$ has SVD $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, any $\mathbf{b} \in \mathbb{R}^m$ can be expanded in the basis of left singular vectors of $\mathbf{A}$ (columns of $\mathbf{U}$), and any $\mathbf{x} \in \mathbb{R}^n$ can be expanded in the basis of right singular vectors of $\mathbf{A}$ (columns of $\mathbf{V}$). The coordinate vectors for these expansions are

$$\mathbf{b}' = \mathbf{U}^T\mathbf{b}, \qquad \mathbf{x}' = \mathbf{V}^T\mathbf{x},$$

which means the relation $\mathbf{b} = \mathbf{A}\mathbf{x}$ can be expressed in terms of $\mathbf{b}'$ and $\mathbf{x}'$:

$$\mathbf{b} = \mathbf{A}\mathbf{x} \iff \mathbf{U}^T\mathbf{b} = \mathbf{U}^T\mathbf{A}\mathbf{x} = \mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{x} \iff \mathbf{b}' = \mathbf{\Sigma}\mathbf{x}'.$$

Whenever $\mathbf{b} = \mathbf{A}\mathbf{x}$, we have $\mathbf{b}' = \mathbf{\Sigma}\mathbf{x}'$. Thus $\mathbf{A}$ reduces to the diagonal matrix $\mathbf{\Sigma}$ when the range is expressed in the basis of column $\mathbf{U}$ and the domain is expressed in the basis of column of $\mathbf{V}$.

If $\mathbf{A} \in \mathbb{S}^n$ has spectral decomposition

$$\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^T,$$

then it implies that if we define, for $\mathbf{b}, \mathbf{x} \in \mathbb{R}^n$ satisfying $\mathbf{b} = \mathbf{A}\mathbf{x}$,

$$\mathbf{b}' = \mathbf{X}^T\mathbf{b}, \qquad \mathbf{x}' = \mathbf{X}^T\mathbf{x}$$

then the newly expended vectors $\mathbf{b}'$ and $\mathbf{x}'$ satisfy $\mathbf{b}' = \mathbf{\Lambda}\mathbf{x}'$.

There are fundamental differences between the SVD and the eigenvalue decomposition.

- The SVD uses two different bases (the sets of left and right singular vectors), whereas the eigenvalue decompositions uses just one (the eigenvectors).
- Not all matrices (even sqaure ones) have a spectral decomposition, but all matrices (even rectangular ones) have a singular value decomposition.
- In applications, eigenvalues tend to be relevant to problems involving the behavior of iterated forms of $\mathbf{A}$, such as matrix powers $\mathbf{A}^k$, whereas singular value vectors tend to be relevant to problems involving the behavior of $\mathbf{A}$ itself or its inverse.

5 Quadratic Forms and Definiteness

Given a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and a vector $\mathbf{x} \in \mathbb{R}^n$, the scalar value $\mathbf{x}^T\mathbf{A}\mathbf{x}$ is called a ***quadratic form***. Written explicitly, we see that

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i=1}^n x_i(\mathbf{A}x)_i = \sum_{i=1}^n x_i\left(\sum_{j=1}^n A_{ij}x_j\right) = \sum_{i=1}^n\sum_{j=1}^n A_{ij}x_ix_j.$$

Note that

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = (\mathbf{x}^T\mathbf{A}\mathbf{x})^T = \mathbf{x}^T\mathbf{A}^T\mathbf{x} = \mathbf{x}^T(\frac{1}{2}\mathbf{A} + \frac{1}{2}\mathbf{A}^T)\mathbf{x},$$

where the first equality follows from the fact that the transpose of a scalar is equal to itself, and the second equality follows from the fact that we are averaging two quantities which are themselves equal. From this, we can conclude that only the symmetric part of $\mathbf{A}$ contributes to the quadratic form. For this reason, we often implicitly assume that the matrices appearing in a quadratic form are symmetric.

We give the following definitions:

- A symmetric matrix $\mathbf{A} \in \mathbb{S}^n$ is ***positive definite (PD)*** if for all non-zero vectors $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}^T\mathbf{A}\mathbf{x} > \mathbf{0}$. This is usually denoted $\mathbf{A} \succ \mathbf{0}$ (or just $\mathbf{A} > \mathbf{0}$), and often times the set of all positive definite matrices is denoted $\mathbb{S}^n_{++}$.
- A symmetric matrix $\mathbf{A} \in \mathbb{S}^n$ is ***positive semidefinite (PSD)*** if for all vectors $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq \mathbf{0}$. This is written $A \succeq \mathbf{0}$ (or just $\mathbf{A} \geq \mathbf{0}$), and the set of all positive semidefinite matrices is often denoted $\mathbb{S}^n_+$.

- Likewise, a symmetric matrix $\mathbf{A} \in \mathbb{S}^n$ is ***negative definite (ND)***, denoted $\mathbf{A} \prec \mathbf{0}$ (or just $\mathbf{A} < \mathbf{0}$) if for all non-zero $\mathbf{x} \in \mathbb{R}^n, \mathbf{x}^T\mathbf{A}\mathbf{x} < \mathbf{0}$.
- Similarly, a symmetric matrix $\mathbf{A} \in \mathbb{S}^n$ is ***negative semidefinite (NSD)***, denoted $\mathbf{A} \preceq \mathbf{0}$ (or just $\mathbf{A} \leq \mathbf{0}$) if for all $\mathbf{x} \in \mathbb{R}^n, \mathbf{x}^T\mathbf{A}\mathbf{x} \leq \mathbf{0}$.
- Finally, a symmetric matrix $\mathbf{A} \in \mathbb{S}^n$ is ***indefinite***, if it is neither positive semidefinite nor negative semidefinite i.e., if there exists $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ such that $\mathbf{x}_1^T\mathbf{A}\mathbf{x}_1 > 0$ and $\mathbf{x}_2^T\mathbf{A}\mathbf{x}_2 < 0$.

It should be obvious that if $\mathbf{A}$ is positive definite, then $-\mathbf{A}$ is negative definite and vice versa. Likewise, if $\mathbf{A}$ is positive semidefinite then $-\mathbf{A}$ is negative semidefinite and vice versa. If $\mathbf{A}$ is indefinite, then so is $-\mathbf{A}$.

There are some important properties of definiteness as following.

- All positive definite and negative definite matrices are always full rank, and hence, invertible. To see why this is the case, suppose that some matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is not full rank. Then, suppose that the jth column of $\mathbf{A}$ is expressible as a linear combination of other $n-1$ columns:

  $$\mathbf{a}_j = \sum_{i\neq j} x_i \mathbf{a}_i,$$

  for some $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n \in \mathbb{R}$. Setting $x_j = -1$, we have

  $$\mathbf{A}\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{a}_i = 0.$$

  This implies $\mathbf{x}^T\mathbf{A}\mathbf{x} = 0$ for some non-zero vector $\mathbf{x}$, so $\mathbf{A}$ must be neither positive definite nor negative definite. Therefore, if $\mathbf{A}$ is either positive definite or negative definite, it must be full rank.
- Let $\mathbf{A} \in \mathbb{S}^n$ has spectral decomposition $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^T$, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_n)$. For $i = 1, \ldots, n$, if $\mathbf{A} \succeq \mathbf{0}$, then $\lambda_i \geq 0$ and if $\mathbf{A} \succ \mathbf{0}$, then $\lambda_i > 0$. Consider if $\mathbf{A} \succ \mathbf{0}$, we have for all $\mathbf{z} \neq \mathbf{0}$,

  $$\mathbf{z}^T\mathbf{A}\mathbf{z} = \mathbf{z}^T\mathbf{X}\mathbf{\Lambda}\mathbf{X}^T\mathbf{z} = \mathbf{y}^T\mathbf{\Lambda}\mathbf{y} = \sum_{i=1}^{n} \lambda_i y_i^2,$$

  where $\mathbf{y} = \mathbf{X}^T\mathbf{z}$. Note that for any $\mathbf{y}$, we can get the corresponding $\mathbf{z}$, by $\mathbf{z} = \mathbf{X}\mathbf{y}$. Thus, by choosing $y_1 = 1, y_2 = \cdots = y_n = 0$, we reduced that $\lambda_1 > 0$. Similarly $\lambda_i > 0$ for all $1 \leq i \leq n$. If $\mathbf{A} \succeq \mathbf{0}$, the above inequalities are weak.
- If $\mathbf{A} \succeq \mathbf{0}$, then $\mathbf{A}$ is non-singular and $\det(\mathbf{A}) > 0$.

Finally, there is one type of positive definite matrix that comes up frequently, and so deserves some special mention. Given any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ (not necessarily symmetric or even square), the matrix $\mathbf{G} = \mathbf{A}^T\mathbf{A}$ (sometimes called a Gram matrix) is always positive semidefinite. Further, if $m \geq n$ (and we assume for convenience that $\mathbf{A}$ is full rank), then $\mathbf{G} = \mathbf{A}^T\mathbf{A}$ is positive definite.

Let $\mathbf{X}$ be a symmetric matrix given by

$$\mathbf{X} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix}$$

Let $\mathbf{S}$ be the Schur complement of $\mathbf{A}$ in $\mathbf{X}$, that is:

$$\mathbf{S} = \mathbf{C} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}.$$

Then we have following results:

- $\mathbf{X} \succ \mathbf{0} \iff \mathbf{A} \succ \mathbf{0}, \mathbf{S} \succ \mathbf{0}$.
- If $\mathbf{A} \succ \mathbf{0}$ then $\mathbf{X} \succeq \mathbf{0} \iff \mathbf{S} \succeq \mathbf{0}$.

To prove it, you can consider the following minimization problem (can be solved easily by matrix calculus)

$$\min_{\mathbf{u}} \mathbf{u}^T\mathbf{A}\mathbf{u} + 2\mathbf{v}^T\mathbf{B}^T\mathbf{u} + \mathbf{v}^T\mathbf{C}\mathbf{v}.$$

6 Matrix Calculus

6.1 Notations

While the topics in the previous sections are typically covered in a standard course on linear algebra, one topic that does not seem to be covered very often (and which we will use extensively) is the extension of calculus to the matrix setting. Despite the fact that all the actual calculus we use is relatively trivial, the notation can often make things look much more difficult than they are.

In this section we present some basic definitions of matrix calculus and provide a few examples. There are two competing notational conventions (***denominator notations*** and ***numerator-layout notation***) which split the field of matrix calculus into two separate groups. The two groups can be distinguished by whether they write the derivative of a scalar with respect to a vector as a column vector (in denominator notations) or a row vector (in numerator-layout notations).

The discussion in this review assumes the denominator notations. The choice of denominator layout does not imply that this is the correct or superior choice. There are advantages and disadvantages to the various layout types. Serious mistakes can result when combining results from different authors without carefully verifying that compatible notations are used. Therefore great care should be taken to ensure notational consistency.

Matrix calculus refers to a number of different notations that use matrices and vectors to collect the derivative of each component of the dependent variable with respect to each component of the independent variable. In general, the independent variable can be a scalar, a vector, or a matrix while the dependent variable can be any of these as well.

6.2 The Gradient

Suppose that $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is a function that takes as input a matrix $\mathbf{A}$ of size $m \times n$ and returns a real value. Then the ***gradient*** of f with respect to $\mathbf{A} \in \mathbb{R}^{m\times n}$ is the matrix of partial derivatives, defined as

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

i.e., an $m \times n$ matrix with

$$\Big(\nabla_{\mathbf{A}} f(\mathbf{A})\Big)_{ij} = \frac{\partial f(\mathbf{A})}{\partial a_{ij}}.$$

So if, in particular, the independent variable is just a vector $\mathbf{x} \in \mathbb{R}^n$,

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \nabla_{\mathbf{x}} f(\mathbf{x}) \in \mathbb{R}^n = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}.$$

It is very important to remember that the gradient (denoted as ∇) of a function is only defined if the function is real-valued, that is, if it returns a scalar value, and the gradient $\nabla_{\mathbf{A}} f(\mathbf{A})$ is always the same as the size of $\mathbf{A}$. We can not, for example, take the gradient of $\mathbf{Ax}, \mathbf{x} \in \mathbb{R}^n$ with respect to $\mathbf{x}$, since this quantity is vector-valued.

It follows directly from the equivalent properties of partial derivatives with that:

- For $\mathbf{x} \in \mathbb{R}^n$, $\dfrac{\partial\Big(f(\mathbf{x}) + g(\mathbf{x})\Big)}{\partial \mathbf{x}} = \dfrac{\partial f(\mathbf{x})}{\partial \mathbf{x}} + \dfrac{\partial g(\mathbf{x})}{\partial \mathbf{x}}$.

- For $\mathbf{x} \in \mathbb{R}^n$, $t \in \mathbb{R}$, $\dfrac{\partial t f(\mathbf{x})}{\partial \mathbf{x}} = t\dfrac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$.
- For $\mathbf{x}, \mathbf{a} \in \mathbb{R}^n$, $\dfrac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$.
- For $\mathbf{x} \in \mathbb{R}^n$, $\dfrac{\partial \mathbf{x}^T\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}$.
- For $\mathbf{x} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{n\times n}$, $\dfrac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$. If $\mathbf{A} \in \mathbb{S}^n$, then $\dfrac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$.
- For $\mathbf{x} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{n\times n}$, $\dfrac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$. If $\mathbf{A} \in \mathbb{S}^n$, then $\dfrac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$.

6.3 The Hessian

Suppose that $f : \mathbb{R}^n \to R^n$ is a function that takes a vector in $\mathbb{R}^n$ and returns a real number. Then the ***Hessian matrix*** with respect to $\mathbf{x}$, written $\nabla^2_{\mathbf{x}} f(\mathbf{x})$ or $\mathbf{H}_f(\mathbf{x})$ (simply as $\mathbf{H}(x)$ when f is clear from context, even $\mathbf{H}$ when $\mathbf{x}$ is also clear), is the $n \times n$ matrix of partial derivatives.

$$\nabla^2_{\mathbf{x}} f(\mathbf{x}) = \mathbf{H}_f(\mathbf{x}) \in \mathbb{R}^{m\times n} = \begin{bmatrix} \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{bmatrix}.$$

If the second derivatives of f are all continuous in a neighborhood of $\mathbf{x}$, then the Hessian is a symmetric matrix, i.e,

$$\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} = \frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_i}.$$

Similar to the gradient, the Hessian is defined only when $f(x)$ is real-valued. It is natural to think of the Hessian as the analogue of the second derivative (and the symbols we use also suggest this relation). But it is **not** the case that the Hessian is the second order gradient or derivative of the independent variable, i.e,

$$\nabla^2_{\mathbf{x}} f(\mathbf{x}) \neq \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}^2}.$$

6.4 Least Squares

Consider the quadratic function $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$ for $\mathbf{A} \in \mathbb{S}^n$. Remember that

$$f(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j.$$

To take the partial derivative, well consider the terms including x_k and x_k^k factors separately:

$$\begin{aligned} \frac{\partial f(\mathbf{x})}{\partial x_k} &= \frac{\partial}{\partial x_k}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j \\ &= \frac{\partial}{\partial x_k}\left[\sum_{i\neq k}\sum_{j\neq k} a_{ij}x_i x_j + \sum_{i\neq k} a_{ik}x_i x_k + \sum_{j\neq k} a_{kj}x_k x_j + a_{kk}x_k^2\right] \\ &= \sum_{i\neq k} a_{ik}x_i + \sum_{j\neq k} a_{kj}x_j + 2a_{kk}x_k \\ &= \sum_{i=1}^{n} a_{ik}x_i + \sum_{j=1}^{n} a_{kj}x_j = 2\sum_{i=1}^{n} a_{ik}x_i. \end{aligned}$$

Lets look at the Hessian of the quadratic function $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$ (it should be obvious that the Hessian of a linear function. In this case,

$$\frac{\partial^2 f(\mathbf{x})}{\partial x_k \partial x_l} = \frac{\partial}{\partial x_k}\left[\frac{\partial f(\mathbf{x})}{\partial x_l}\right] = \frac{\partial}{\partial x_k}\left[\sum_i^n a_{li}x_i\right] = 2a_{lk} = 2a_{kl}.$$

There it should be clear that $\nabla^2_{\mathbf{x}} f(\mathbf{x}) = 2\mathbf{A}$.

Then, let's apply the equations we obtained to derive the least squares equations. Suppose we are given matrices $\mathbf{A} \in \mathbb{R}^{m\times n}$ with full rank and a vector $\mathbf{b} \in \mathbb{R}^m$ such that $\mathbf{b} \in \mathcal{R}(\mathbf{A})$, where $m > n$. In this situation we will not be able to find a vector $\mathbf{x} \in \mathbb{R}^n$, such that $\mathbf{A}\mathbf{x} = \mathbf{b}$, so instead we want to find a vector $\mathbf{x}$ such that $\mathbf{A}\mathbf{x}$ is as close as possible to $\mathbf{b}$, as measured by the square of the Euclidean norm. Let $l(\mathbf{x}) = ||\mathbf{A}\mathbf{x} - \mathbf{b}||_2^2$. Using the fact that $||\mathbf{x}||_2^2 = \mathbf{x}^T\mathbf{x}$, we have

$$l(\mathbf{x}) = ||\mathbf{A}\mathbf{x} - \mathbf{b}||_2^2 = (\mathbf{A}\mathbf{x} - \mathbf{b})^T(\mathbf{A}\mathbf{x} - \mathbf{b}) = \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} - 2\mathbf{b}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{b}.$$

Taking the gradient with respect to $\mathbf{x}$ we have, and using the properties we has derived, we have

$$\begin{aligned}
\frac{\partial l(\mathbf{x})}{\partial \mathbf{x}} &= \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T\mathbf{A}^T\mathbf{x} - 2\mathbf{b}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{b}) \\
&= \frac{\partial}{\partial \mathbf{x}}\mathbf{x}^T\mathbf{A}^T\mathbf{x} - \frac{\partial}{\partial \mathbf{x}}2\mathbf{b}^T\mathbf{A}\mathbf{x} + \frac{\partial}{\partial \mathbf{x}}\mathbf{b}^T\mathbf{b} \\
&= 2\mathbf{A}^T\mathbf{A}\mathbf{x} - 2\mathbf{A}^T\mathbf{b}.
\end{aligned}$$

Since we have $\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} = \sum_{k=1}^m \Big(\sum_{l=1}^m a_{kl}x_l\Big)^2$, the fact

$$\begin{aligned}
&\frac{\partial^2}{\partial x_i \partial x_j}\sum_{k=1}^m \Big(\sum_{l=1}^m a_{kl}x_l\Big)^2 \\
= &\frac{\partial}{\partial x_i}\left(\frac{\partial}{\partial x_j}\sum_{k=1}^m \Big(\sum_{l=1}^m a_{kl}x_l\Big)^2\right) \\
= &\frac{\partial}{\partial x_i}\sum_{k=1}^m \Big(2a_{kj}\sum_{l=1}^n a_{kl}x_j\Big) \\
= &\ 2\left(\sum_{k=1}^m a_{kj}\right)\left(\sum_{l=1}^n a_{ki}\right)
\end{aligned}$$

means

$$\nabla^2_{\mathbf{x}}\Big(\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x}\Big) = 2\mathbf{A}^T\mathbf{A},$$

and it is clear that $\nabla^2_{\mathbf{x}}(\mathbf{b}^T\mathbf{A}\mathbf{x}) = \mathbf{0}$ and $\nabla^2_{\mathbf{x}}(\mathbf{b}^T\mathbf{b}) = \mathbf{0}$, Hence we have $\nabla^2_{\mathbf{x}} l(\mathbf{x}) = 2\mathbf{A}^T\mathbf{A} \succ \mathbf{0}$. Then the positive definite Hessian leads we can get the optimal $\widehat{\mathbf{x}}$ by setting the gradient $l(\mathbf{x})$ with respect to $\mathbf{x}$ be zero, i.e,

$$\begin{aligned}
&\frac{\partial l(\mathbf{x})}{\partial \mathbf{x}}\bigg|_{\mathbf{x}=\widehat{\mathbf{x}}} = \mathbf{0} \\
\Longrightarrow\ &2\mathbf{A}^T\mathbf{A}\widehat{\mathbf{x}} - 2\mathbf{A}^T\mathbf{b} = \mathbf{0} \\
\Longrightarrow\ &\widehat{\mathbf{x}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}.
\end{aligned}$$

Note that in the computation of $\nabla^2_{\mathbf{x}} l(\mathbf{x})$ above, we **haven't** taken the derivative of $\nabla_{\mathbf{x}} l(\mathbf{x})$ with respect with $\mathbf{x}$.

For $\mathbf{A} \in \mathbb{R}^{m\times n}$ is full rank and $m > n$, we define the ***projection*** of a vector $\mathbf{y} \in \mathbb{R}^m$ onto the range of $\mathbf{A}$ is given by

$$\text{Proj}(\mathbf{y}, \mathbf{A}) = \underset{\mathbf{v}\in\mathcal{R}(\mathbf{A})}{\text{argmin}}\, ||\mathbf{v} - \mathbf{y}||_2 = \underset{\mathbf{v}\in\mathcal{R}(\mathbf{A})}{\text{argmin}}\, ||\mathbf{v} - \mathbf{y}||_2^2 = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}.$$

6.5 Digression: Generalized Inverse

Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ has reduced singular value decomposition $\mathbf{A} = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^T$, with $\text{rank}(\mathbf{A}) = r$. We define the ***pseudo-inverse*** or ***Moore-Pseudo inverse*** of $\mathbf{A}$ as

$$\mathbf{A}^\dagger = \widehat{\mathbf{V}}\widehat{\mathbf{\Sigma}}'\widehat{\mathbf{U}}^T,$$

where $\widehat{\mathbf{\Sigma}}' = \text{diag}(\sigma_1^{-1}, \ldots, \sigma_r^{-1}, 0, \ldots, 0)$ is a $m \times m$ diagonal matrix. Alternative expressions are

$$\mathbf{A}^\dagger = \lim_{\epsilon\to 0^+} (\mathbf{A}^T\mathbf{A} + \epsilon\mathbf{I})^{-1}\mathbf{A}^T = \lim_{\epsilon\to 0^+} \mathbf{A}^T(\mathbf{A}\mathbf{A}^T + \epsilon\mathbf{I})^{-1}.$$

In special cases, we have

- If $\text{rank}(\mathbf{A}) = n$, then $\mathbf{A}^\dagger = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$.
- If $\text{rank}(\mathbf{A}) = m$, then $\mathbf{A}^\dagger = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}$.
- If $\mathbf{A}$ is square and non-singular, then $\mathbf{A}^\dagger = \mathbf{A}^{-1}$.

By the result of last subsection, $\mathbf{A}^\dagger\mathbf{b}$ is the solution of the least-squares problem.

More generally, we can extend the pseudo-inverse to the ***generalized inverse***. There four conditions are used to define different generalized inverses, let $\mathbf{X} \in \mathbb{R}^{n\times m}$

1. $\mathbf{AXA} = \mathbf{A}$
2. $\mathbf{XAX} = \mathbf{X}$
3. $(\mathbf{AX})^T = \mathbf{AX}$
4. $(\mathbf{XA})^T = \mathbf{XA}$

If $\mathbf{X}$ satisfies condition 1, it is a 1-generalized inverse of $\mathbf{A}$, if it satisfies conditions 1 and 2 then it is a generalized reflexive inverse of $\mathbf{A}$ (1-2 generalized inverse), and if it satisfies all 4 conditions, then it is a pseudo-inverse of $\mathbf{A}$.

6.6 Gradients of the Determinant

Recall from our discussion of determinants that

$$\det(\mathbf{A}) = \sum_{i=1}^{n}(-1)^{i+j}a_{ij}\det(\mathbf{A}_{\setminus i,\setminus j}) \quad (\text{for any } j \in \{1, 2, \ldots, n\})$$

so

$$\begin{aligned}
\frac{\partial}{\partial a_{kl}}\det(\mathbf{A}) &= \frac{\partial}{\partial a_{kl}}\sum_{i=1}^{n}(-1)^{i+j}a_{ij}\det(\mathbf{A}_{\setminus i,\setminus j}) \\
&= \frac{\partial}{\partial a_{kl}}\sum_{l=1}^{n}(-1)^{i+l}a_{il}\det(\mathbf{A}_{\setminus i,\setminus l}) \\
&= (-1)^{k+l}a_{kl}\det(\mathbf{A}_{\setminus k,\setminus l}) = \big(\text{adj}(\mathbf{A})\big)_{lk}.
\end{aligned}$$

From this it immediately follows from the properties of the adjoint that

$$\nabla_{\mathbf{A}}\det(\mathbf{A}) = \big(\text{adj}(\mathbf{A})\big)^T.$$

Now lets consider the function $f : \mathbb{S}^n_{++} \to \mathbb{R}, f(\mathbf{A}) = \ln\big(\det(\mathbf{A})\big)$. Note that we have to restrict the domain of f to be the positive definite matrices, since this ensures that $\det(\mathbf{A}) > 0$, so that the log of $\det(\mathbf{A})$ is a real number. In this case we can use the chain rule from single-variable calculus (nothing fancy) to see that

$$\frac{\partial \ln\big(\det(\mathbf{A})\big)}{\partial a_{ij}} = \frac{\partial \ln\big(\det(\mathbf{A})\big)}{\partial \det(\mathbf{A})}\frac{\partial \det(\mathbf{A})}{\partial a_{ij}} = \frac{1}{\det(\mathbf{A})}\frac{\partial \det(\mathbf{A})}{\partial a_{ij}}$$

From this it should be obvious that

$$\nabla_{\mathbf{A}} \ln\big(\det(\mathbf{A})\big) = \frac{1}{\det(\mathbf{A})} \nabla_{\mathbf{A}} \det(\mathbf{A}) = \mathbf{A}^{-1}.$$

where we can drop the transpose in the last expression because A is symmetric. Note the similarity to the single-valued case, i.e, for x is a positive real number,

$$\frac{\partial \ln x}{\partial x} = \frac{1}{x}.$$

6.7 Eigenvalues as Optimization

Let's use matrix calculus to solve an optimization problem in a way that leads directly to eigenvalue/eigenvector analysis. Consider the following, equality constrained optimization problem (by solving this problem, we will know how to get the 2-norm of a matrix eventually):

$$\max_{\mathbf{x} \in \mathbb{R}^n} \mathbf{x}^T \mathbf{A} \mathbf{x} \qquad \text{subject to } ||\mathbf{x}||_2^2 = 1$$

for a symmetric matrix $\mathbf{A} \in \mathbb{S}^n$. A standard way of solving optimization problems with equality constraints is by forming the ***Lagrangian***, an objective function that includes the equality constraints The Lagrangian in this case can be given by

$$\mathcal{L}(\mathbf{x}, \lambda) = \mathbf{x}^T \mathbf{A} \mathbf{x} - \lambda \mathbf{x}^T \mathbf{x}$$

where λ is called the Lagrange [3] multiplier associated with the equality constraint. It can be established that for $\widehat{\mathbf{x}}$ to be a optimal point to the problem, the gradient of the Lagrangian has to be zero at $\widehat{\mathbf{x}}$ (this is not the only condition, but it is required). That is,

$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda) = \nabla_{\mathbf{x}} (\mathbf{x}^T \mathbf{A} \mathbf{x} - \lambda \mathbf{x}^T \mathbf{x}) = 2\mathbf{A}^T \mathbf{x} - 2\lambda \mathbf{x} = \mathbf{0}.$$

Notice that this is just the linear equation $\mathbf{A}\mathbf{x} = \lambda \mathbf{x}$. This shows that the only points which can possibly maximize $\mathbf{x}^T \mathbf{A} \mathbf{x}$ assuming $\mathbf{x}^T \mathbf{x} = 1$ are the standard eigenvectors of $\mathbf{A}$. Hence $\widehat{\mathbf{x}}$ is the standard eigenvector of $\mathbf{A}$ corresponding to the largest eigenvalue of $\mathbf{A}$.

6.8 General Derivatives of Matrix Functions

To the end of this section, let's see how to define the derivative of a function whose dependent value is also a matrix (or vector). In this case, we use the term ***derivative***, rather than gradient. Suppose that $\mathbf{F} : \mathbb{R}^{m \times n} \to \mathbb{R}^{p \times q}$ is a function that takes as input a matrix $\mathbf{A}$ of size $m \times n$ and returns a matrix $\mathbf{F}(\mathbf{X})$ of size $p \times q$ matrix, i.e,

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

and

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

Then the derivative of $\mathbf{F}$ with respect to $\mathbf{A}$ defined as the following partition form

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

[3] We will cover the Lagrangian in greater detail later in our class.

where each sub-matrix $\frac{\partial f_{ij}(\mathbf{X})}{\partial \mathbf{X}}$ with size $p \times q$ by definition of gradient of scalar-value function.

Let's consider the Hessian matrix again. Remember that for $f : \mathbb{R}^n \to \mathbb{R}$ is a function that takes a vector in $\mathbb{R}^n$ and returns a real number, the gradient of f is

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}.$$

Then we can also define Hessian as following

$$\begin{aligned} \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}^T \partial \mathbf{x}} = \frac{\partial}{\partial \mathbf{x}^T}\Big(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\Big) &= \frac{\partial}{\partial \mathbf{x}^T} \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial \mathbf{x}^T}\Big(\frac{\partial f(\mathbf{x})}{\partial x_1}\Big) \\ \frac{\partial}{\partial \mathbf{x}^T}\Big(\frac{\partial f(\mathbf{x})}{\partial x_2}\Big) \\ \vdots \\ \frac{\partial}{\partial \mathbf{x}^T}\Big(\frac{\partial f(\mathbf{x})}{\partial x_n}\Big) \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{bmatrix} \\ &= \nabla_{\mathbf{x}}^2 f(\mathbf{x}). \end{aligned}$$

Jan R. Magnus and H. Neudecker said in this type of definition, the determinant of $\partial \mathbf{F}(\mathbf{X})/\partial \mathbf{X}$ has no interpretation and a useful chain rule does not exist, and they also claimed this is a bad one. In practise, it is not frequently to deal with the derivative of matrix-by-matrix function. Hence we will not discuss this case any more.

7 Bibliographic Remarks

All of things in this review are standard and its details can be found in many books. The following reading list may be useful for your further learning.

1. Horn, R. A., & Johnson, C. R. (2012). Matrix analysis. Cambridge university press.
2. Kolter, Z., & Chuong Do. (2008). Linear Algebra Review and Reference. Available online: http.
3. Lütkepohl, H. (1996). Handbook of Matrices.
4. Magnus, J. R., & Neudecker, H. (1985). Matrix differential calculus with applications to simple, Hadamard, and Kronecker products. Journal of Mathematical Psychology, 29(4), 474-492.
5. Matrix Calculus. In Wikipedia. Retrieved March 3, 2014, from http://en.wikipedia.org/wiki/Matrix_calculus
6. Petersen, K. B., & Pedersen, M. S. (2008). The matrix cookbook. Technical University of Denmark, 7-15.
7. Stewart, G. W. (1993). On the early history of the singular value decomposition. SIAM review, 35(4), 551-566.
8. Trefethen, L. N., & Bau III, D. (1997). Numerical linear algebra (Vol. 50). Siam.