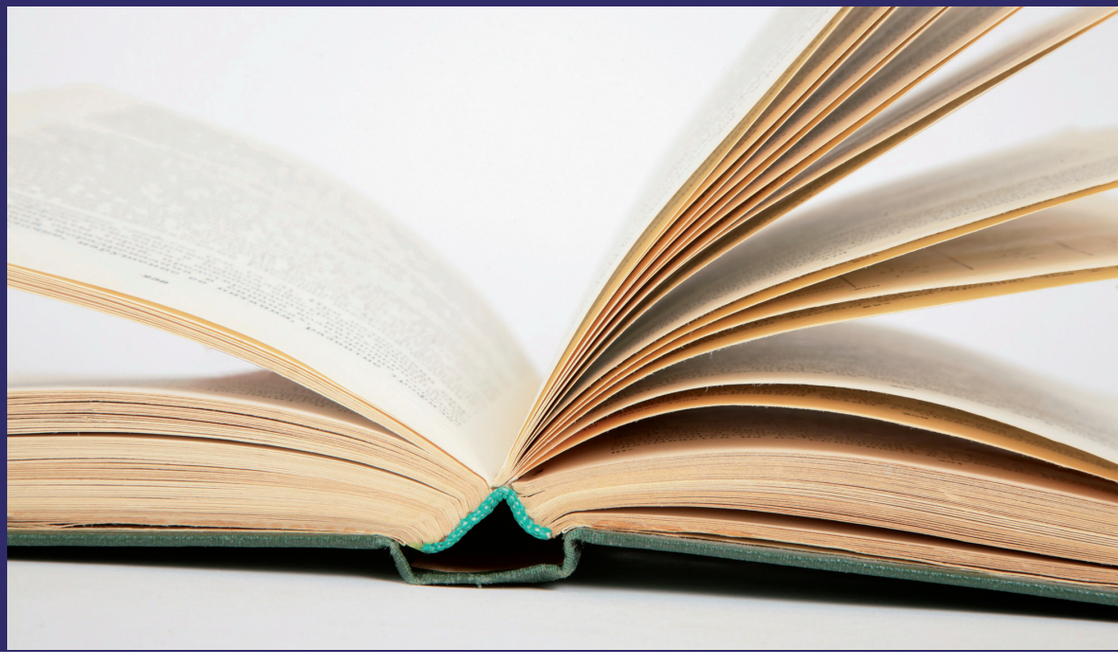
Statistics, multivariate data analysis and convex optimization are applied widely in many scientific domains and most analytical techniques are developed based on matrix analysis and matrix calculus because matrix is abstract representation of multivariate data. Although it is slightly confused for us to comprehend their concepts and theories, matrix analysis and calculus give us exciting results which enhance data analysis techniques to be more plentiful and accurate. So the report is survey of matrix analysis and calculus, which includes five main sections such as basic concepts, matrix analysis, matrix derivative, composite derivative, and applications of matrix. Matrix derivative and composite derivative are subjects of matrix calculus.



Loc Nguyen is a director at Sunflower Soft Company, Vietnam. He is interested in computer science, statistics and mathematics. He serves as reviewer and editor in a wide range of international journals. Now he is a volunteer of Statistics Without Borders of American Statistics Association.

978-3-659-69400-4

Loc Nguyen

Matrix Analysis and Calculus

LAP LAMBERT Academic Publishing

Loc Nguyen

Matrix Analysis and Calculus

Loc Nguyen

Matrix Analysis and Calculus

LAP LAMBERT Academic Publishing

Impressum / Imprint
Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.
Alle in diesem Buch genannten Marken und Produktnamen unterliegen warenzeichen-, marken- oder patentrechtlichem Schutz bzw. sind Warenzeichen oder eingetragene Warenzeichen der jeweiligen Inhaber. Die Wiedergabe von Marken, Produktnamen, Gebrauchsnamen, Handelsnamen, Warenbezeichnungen u.s.w. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutzgesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Bibliographic information published by the Deutsche Nationalbibliothek: The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.
Any brand names and product names mentioned in this book are subject to trademark, brand or patent protection and are trademarks or registered trademarks of their respective holders. The use of brand names, product names, common names, trade names, product descriptions etc. even without a particular marking in this work is in no way to be construed to mean that such names may be regarded as unrestricted in respect of trademark and brand protection legislation and could thus be used by anyone.

Coverbild / Cover image: www.ingimage.com

Verlag / Publisher:
LAP LAMBERT Academic Publishing
ist ein Imprint der / is a trademark of
OmniScriptum GmbH & Co. KG
Heinrich-Böcking-Str. 6-8, 66121 Saarbrücken, Deutschland / Germany
Email: info@lap-publishing.com

Herstellung: siehe letzte Seite /
Printed at: see last page
ISBN: 978-3-659-69400-4

Copyright © 2015 OmniScriptum GmbH & Co. KG
Alle Rechte vorbehalten. / All rights reserved. Saarbrücken 2015

Matrix Analysis and Calculus

Loc Nguyen

Vietnam Institute of Mathematics

Abstract

Statistics, multivariate data analysis and convex optimization are applied widely in many scientific domains and most analytical techniques are developed based on matrix analysis and matrix calculus because matrix is abstract representation of multivariate data. Although it is slightly confused for us to comprehend their concepts and theories, matrix analysis and calculus give us exciting results which enhance data analysis techniques to be more plentiful and accurate. So the report is survey of matrix analysis and calculus, which includes five main sections such as basic concepts, matrix analysis, matrix derivative, composite derivative, and applications of matrix. Matrix derivative and composite derivative are subjects of matrix calculus.

Contents

1. Basic concepts

We begin matrix analysis and matrix calculus with some basic concepts of matrix. Before discussing main subjects, there are some conventions. Firstly, if there is no additional explanation, normal letters denote scalar variables and numbers; vectors are denoted by bold letters and matrices are denoted as bold and uppercase letters. For example, letters x, $\boldsymbol{x}$ and $\boldsymbol{X}$ indicate a scalar variable x, a vector $\boldsymbol{x}$ and a matrix $\boldsymbol{X}$. Note that term "scalar" refers real or complex number in opposite to vector and matrix. All constants are denoted by lowercase letters, for example, letters *c, a* and *A* express a scalar constant, a vector constant and a matrix constant, respectively and so their difference is based on study context. If there is the mix of constant vectors and constant scalars in the same expression, Greek lowercase letters such as α, β and γ often denote constant vectors and Latin lowercase letters such as a, b and c often denote scalar constants. Analysis spaces or algebra structures are denoted by letter-like symbols such as complex number field $\mathbb{C}$**,** real number field $\mathbb{R}$, rational number field $\mathbb{Q}$, integer ring $\mathbb{Z}$, set of natural number $\mathbb{N}$, vector space $\mathbb{V}$, n-dimension vector space over real number field $\mathbb{R}^n$. The default vector space is $\mathbb{R}^n$.

Vector and matrix properties

Vector $\boldsymbol{x}$ is a range of n values denoted $\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ or $(x_1, x_2, \dots, x_n)^T$ called n-dimension vector where T denotes transpose operation which changes row to column and otherwise. Please pay attention to transpose operation because it is used over the whole of report. There are two kinds of vector such as aforementioned column vector $\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ and row vector $(x_1, x_2, \dots, x_n)$. By default, $\boldsymbol{x}$ denotes column vector and $\boldsymbol{x}^T$ denotes row vector. Each x_i is partial value (partial component, partial element, or partial scalar) of vector $\boldsymbol{x}$. If $\boldsymbol{x}$ is random variable, x_i is also random variable. Simple vector operations include transposition, addition, subtraction, scalar multiplication and scalar division.

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}^T = (x_1, x_2, \dots, x_n) \text{ and } (x_1, x_2, \dots, x_n)^T = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

$$\boldsymbol{x} \pm \boldsymbol{y} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \pm \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 \pm y_1 \\ x_2 \pm y_2 \\ \vdots \\ x_n \pm y_n \end{pmatrix}$$

$$c\boldsymbol{x} = c\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} cx_1 \\ cx_2 \\ \vdots \\ cx_n \end{pmatrix} \text{ and } \frac{\boldsymbol{x}}{c} = \frac{1}{c}\boldsymbol{x} = \begin{pmatrix} x_1/c \\ x_2/c \\ \vdots \\ x_n/c \end{pmatrix}$$

The *dot product* of two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ denoted $\boldsymbol{x} \cdot \boldsymbol{y}$ or $\boldsymbol{x}\boldsymbol{y}^T$ is a scalar value which is sum of multiplications of their components. The dot product is also called inner product or scalar product.

$$\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{x}^T \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i$$

The length also called norm, module, or magnitude of vector $\boldsymbol{x}$ is:

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

Both notations |.| and ||.|| are used to indicate length of vector but the notation |.| is used more frequently. We also use the notation |.| to indicate absolute value of scalar and determinant of matrix. Normalized vector is defined as vector whose length is 1. Arbitrary vector is normalized by dividing itself by its length.

$$normalize(\boldsymbol{x}) = \frac{\boldsymbol{x}}{|\boldsymbol{x}|} = \begin{pmatrix} x_1/|\boldsymbol{x}| \\ x_2/|\boldsymbol{x}| \\ \vdots \\ x_n/|\boldsymbol{x}| \end{pmatrix}$$

The cosine of angle between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is the ratio of its product to multiplication of their lengths.

$$cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^T \boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|}$$

Cosine ranges in interval [–1, 1]. The *cross product* or outer product of two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ denoted $\boldsymbol{x} \times \boldsymbol{y}$ is the vector that is perpendicular to both $\boldsymbol{x}$ and $\boldsymbol{y}$ according to right-hand rule. The length of cross product vector is:

$$|\boldsymbol{x} \times \boldsymbol{y}| = |\boldsymbol{x}||\boldsymbol{y}| sin(\boldsymbol{x}, \boldsymbol{y})$$

Where $sin(\boldsymbol{x}, \boldsymbol{y})$ is the sine of angle between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, which ranges in $[0, \pi]$ according to right-hand rule, hence $sin(\boldsymbol{x}, \boldsymbol{y})$ is always greater than or equal to 0 and is calculated by formula $sin(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{1 - cos^2(\boldsymbol{x}, \boldsymbol{y})} = \sqrt{1 - \left(\frac{\boldsymbol{x}^T \boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|}\right)^2}$. In 3-dimensional vector space, the cross product can be computed via determinant of second order square matrix. Given vectors $\boldsymbol{x} = (x_1, x_2, x_3)$ and $\boldsymbol{y} = (y_1, y_2, y_3)$, their cross product $\boldsymbol{z} = \boldsymbol{x} \times \boldsymbol{y}$ is determined as follows:

$$\boldsymbol{z} = \boldsymbol{x} \times \boldsymbol{y} = \left(\begin{vmatrix} x_2 & x_3 \\ y_2 & y_3 \end{vmatrix}, \begin{vmatrix} x_3 & x_1 \\ y_3 & y_1 \end{vmatrix}, \begin{vmatrix} x_1 & x_2 \\ y_1 & y_2 \end{vmatrix}\right)$$

Where $\begin{vmatrix} x_i & x_j \\ y_i & y_j \end{vmatrix} = (x_i y_j - x_j y_i)$ is determinant of square matrix $\begin{pmatrix} x_i & x_j \\ y_i & y_j \end{pmatrix}$, which is discussed later. Note that the semantic of cross product $\boldsymbol{x} \times \boldsymbol{y} = \boldsymbol{z}$ is the same to the semantic of normal multiplication like 2 * 3 = 6.

According to (Baker K. , 2013), given an integer $k > 1$, the k^{th} power of vector $\boldsymbol{x}$ is defined as below:

$$\begin{cases} \boldsymbol{x}^k = \prod_{i=1}^{k/2} \boldsymbol{x}^T \boldsymbol{x} \text{ if } k \text{ is even} \\ \boldsymbol{x}^k = \boldsymbol{x} \times \prod_{i=1}^{(k-1)/2} \boldsymbol{x}^T \boldsymbol{x} \text{ if } k \text{ is odd} \end{cases}$$

Where Π denotes cross product or scalar product. So we infer that $\boldsymbol{x}^k$ is scalar if k is even and $\boldsymbol{x}^k$ is vector if k is odd. We have summarization table for $k = 2, 3, 4, 5$ as follows (Baker M. J., n.d.):

Power	Value	Type
$\boldsymbol{x}^2$	$\boldsymbol{x} \cdot \boldsymbol{x}$	Scalar
$\boldsymbol{x}^3$	$(\boldsymbol{x} \cdot \boldsymbol{x}) \times \boldsymbol{x}$	Vector
$\boldsymbol{x}^4$	$(\boldsymbol{x} \cdot \boldsymbol{x}) \times (\boldsymbol{x} \cdot \boldsymbol{x})$	Scalar
$\boldsymbol{x}^5$	$(\boldsymbol{x} \cdot \boldsymbol{x}) \times (\boldsymbol{x} \cdot \boldsymbol{x}) \times \boldsymbol{x}$	Vector

Note, given vector $\boldsymbol{x}$ and scalars a, b, we have:

$$a \times \boldsymbol{x} = a\boldsymbol{x} \text{ and } a \times b = ab$$

Matrix $\boldsymbol{A}$ is a table including m rows and n columns, whose cells or elements are scalar values.

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

Matrix $\boldsymbol{A}$ denoted (a_{ij}), $\boldsymbol{A}(m_x n)$, $\boldsymbol{A}_{m_x n}$, or $m_x n$ matrix $\boldsymbol{A}$ can be considered as a set of m row vectors or a set of n column vectors.

$$\boldsymbol{A}(m_x n) = \left((a_{11}, a_{12}, \dots, a_{1n}), (a_{21}, a_{22}, \dots, a_{2n}), \dots, (a_{m1}, a_{m2}, \dots, a_{mn})\right)$$
$$= \left(\begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{pmatrix}, \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{pmatrix}, \dots, \begin{pmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{pmatrix}\right)$$

Matrix $\boldsymbol{A}(n_x n)$ having the same number of rows and columns is called square matrix. Square matrix is very popular in computational application. Matrix $\boldsymbol{A}(m_x n)$ having different number of rows and columns is often called rectangle or non-square matrix. Vector can be considered as 1-row or 1-column matrix, so $\boldsymbol{X}(1_x n)$ denote row vector and $\boldsymbol{X}(n_x 1)$ denotes column vector. By default, vector $\boldsymbol{X}$ is column vector if there is no additional note. Given matrices $\boldsymbol{A}(m_x n)$ and $\boldsymbol{B}(n_x p)$ and scalar constant c, following are matrix operations including transposition, multiplication, addition, subtraction and scalar multiplication.

$\boldsymbol{A}^T = (a_{ji})$ where T is transpose operation which changes row to column and otherwise.
$\boldsymbol{A} + \boldsymbol{B} = (a_{ij}) + (b_{ij})$
$\boldsymbol{A} - \boldsymbol{B} = (a_{ij}) - (b_{ij})$
$c\boldsymbol{A} = (ca_{ij})$
$\boldsymbol{AB} = \boldsymbol{C}(m_x p) = (c_{ij}) = (\sum_{k=1}^{n} a_{ik}\, b_{jk})$ where the number of columns of $\boldsymbol{A}$ is equal to the number of

rows of $\boldsymbol{B}$. This is *multiplication condition*.

Given $n_x n$ square matrix $\boldsymbol{A}$, the n^{th} power of $\boldsymbol{A}$ is defined as $\boldsymbol{A}^k = \prod_{i=1}^{k} \boldsymbol{A}$. It is convention that $\boldsymbol{A}^0 = \boldsymbol{I}_n$ where $\boldsymbol{I}_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$ is identity matrix whose diagonal elements are 1 and remaining elements are 0. Following are properties of matrix operations.

$$(\boldsymbol{A}^T)^T = \boldsymbol{A}$$
$$(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$$
$$\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{B} + \boldsymbol{A}$$
$$\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}$$
$$\boldsymbol{A}(\boldsymbol{BC}) = (\boldsymbol{AB})\boldsymbol{C}$$

Note that matrix multiplication is not commutative. Following is the list of some special vectors and matrices (Hardle & Simar, 2013, p. 59).

Name	Definition	Notation	Example
Scalar	Scalar value is considered as 1-element vector or matrix	c	1
Column vector	$(a_1, a_2, \dots, a_n)^T$	$\boldsymbol{a}$	$\begin{pmatrix} 1 \\ 2 \end{pmatrix}$
Row vector	$(a_1, a_2, \dots, a_n)$	$\boldsymbol{a}^T$	*(1, 2)*
Vector of ones	$(1,1, \dots ,1)^T$	$\mathbf{1}_n$	$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$
Vector of zeros or zero vector	$(0,0, \dots ,0)^T$	$\mathbf{0}_n$	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$
Square matrix	Having the same number of rows and columns	$\boldsymbol{A}(n_x n)$	$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$
Diagonal matrix	Square matrix and $a_{ij} = 0, \forall i \neq j$	$diag(a_{ii})$	$\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$
Identity matrix	Diagonal matrix and $a_{ij} = 0, \forall i \neq j$ and $a_{ii} = 1$	$\boldsymbol{I}_n$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
Unit matrix	Square matrix and $a_{ij} = 1$	$\mathbf{1}_n\mathbf{1}_n{}^T$	$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$
Null matrix or zero matrix	Square matrix and $a_{ij} = 0$	$\mathbf{0}_n\mathbf{0}_n{}^T$	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$
Symmetric matrix	Square matrix and $a_{ij} = a_{ji}$		$\begin{pmatrix} 1 & 3 \\ 3 & 2 \end{pmatrix}$
Upper triangular matrix	Square matrix and $a_{ij} = 0, \forall i < j$		$\begin{pmatrix} 1 & 4 & 5 \\ 0 & 2 & 6 \\ 0 & 0 & 3 \end{pmatrix}$
Idempotent matrix	Square matrix and $\boldsymbol{AA} = \boldsymbol{A}$		$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix}$
Nilpotent matrix	Square matrix and $\underbrace{\boldsymbol{AA} \dots \boldsymbol{A}}_{k\ times} = \mathbf{0}_n\mathbf{0}_n{}^T$		$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$

Orthogonal matrix	Square matrix and $\boldsymbol{AA}^T = \boldsymbol{I}_n = \boldsymbol{A}^T\boldsymbol{A}$		$\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}$

The notation $\mathbf{0}$ can indicate both zero vector $\mathbf{0}_n$ and zero matrix $\mathbf{0}_n\mathbf{0}_n^T$ according to study context. The notation $\boldsymbol{I}$ without n-index implicates identity $n_x n$ matrix. Moreover diagonal matrix – an important kind of square matrix with regard to matrix analysis can be denoted as a composition of column vectors such as $\boldsymbol{A}(n_x n) = (a_{ij}) = diag(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_n)$ where $\boldsymbol{\lambda}_i$ are n-dimensional vector whose elements are zero except i^{th} element. Following is an example of diagonal matrix.

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} = diag(\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, \boldsymbol{\lambda_3}) \text{ where } \boldsymbol{\lambda_1} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\lambda_2} = \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix}, \boldsymbol{\lambda_3} = \begin{pmatrix} 0 \\ 0 \\ 3 \end{pmatrix}$$

We assume that all vectors and matrices have n dimensions if there is no additional explanation. Now we survey characteristics of matrix. The rank of $m_x n$ matrix $\boldsymbol{A}$ denoted $rank(\boldsymbol{A})$ is the maximum number of linear independent rows or columns, $rank(\boldsymbol{A}) \leq min(m, n)$. Non-singular or invertible matrix is square matrix and all its rows (columns) are independent; so the rank of non-singular is its number of rows (columns).

The trace of square matrix $\boldsymbol{A}$ denoted $tr(\boldsymbol{A})$ is the sum of all is diagonal, $tr(\boldsymbol{A}) = \sum a_{ii}$. Following are properties of *trace* and *rank* (Hardle & Simar, 2013, p. 62).

$tr(\boldsymbol{A} + \boldsymbol{B}) = tr(\boldsymbol{A}) + tr(\boldsymbol{B})$
$tr(c\boldsymbol{A}) = ctr(\boldsymbol{A})$
$tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$ given $\boldsymbol{A}(m_x n)$ and $\boldsymbol{B}(n_x m)$
$tr(\boldsymbol{ABC}) = tr(\boldsymbol{BCA}) = tr(\boldsymbol{CAB})$ given $\boldsymbol{A}(m_x n)$, $\boldsymbol{B}(n_x p)$ and $\boldsymbol{C}(p_x m)$
$tr(\boldsymbol{ABCD}\ldots\boldsymbol{Z}\ldots) = tr(\ldots\boldsymbol{Z}\ldots\boldsymbol{DCBA})$ when multiplication condition is satisfied.
$tr(\boldsymbol{A}) = tr(\boldsymbol{A}^T)$
$tr(\boldsymbol{A}^T\boldsymbol{B}) = tr(\boldsymbol{AB}^T)$
$\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a} = tr(\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}})$ where $\boldsymbol{a}$ is a vector
$rank(\boldsymbol{A}) \leq min(m, n)$ given $\boldsymbol{A}(m_x n)$
$rank(\boldsymbol{A}) \geq 0$
$rank(\boldsymbol{A}) = rank(\boldsymbol{A}^T)$
$rank(\boldsymbol{A}(n_x n)) = n$ if $\boldsymbol{A}$ is non-singular
$rank(\boldsymbol{A}^T\boldsymbol{A}) = rank(\boldsymbol{A}\boldsymbol{A}^T) = rank(\boldsymbol{A})$
$rank(\boldsymbol{A}+\boldsymbol{B}) = rank(\boldsymbol{A}) + rank(\boldsymbol{B})$
$rank(\boldsymbol{AB}) \leq min(rank(\boldsymbol{A}), rank(\boldsymbol{B}))$
$rank(\boldsymbol{ABC}) = rank(\boldsymbol{B})$ given $\boldsymbol{A}$ and $\boldsymbol{C}$ are non-singular.

Please pay attention to trace operator because it is important linear operator which relates to matrix derivative.

Matrix determinant and inverse

Given a square matrix $\boldsymbol{A}(n_x n)$, determinant of matrix $\boldsymbol{A}$ denoted $det(\boldsymbol{A})$ or $|\boldsymbol{A}|$ is defined as the sum over all permutations σ (s) of indexes $\{1, 2, \ldots, n\}$. The i^{th} value of permutation σ is denoted

permutations σ_i. The set of all permutations σ (s) given n indexes $\{1, 2,\ldots, n\}$ is denoted S_n. S_n has $n!$ elements (n factorial elements).

$$|\boldsymbol{A}| = \det(\boldsymbol{A}) = \sum_{\sigma \in S_n} sgn(\sigma) \prod_{i=1}^{n} a_{i\sigma_i}$$

Formula above is Leibniz formula where $sgn(\sigma)$ denotes the sign of permutation σ.

$$\begin{cases} sgn(\sigma) = 1 & \text{if } \sigma \text{ is even} \\ sgn(\sigma) = -1 & \text{if } \sigma \text{ is odd} \end{cases}$$

Permutation σ is even (or odd) when the new permutation can be obtained by the even (or odd, respectively) number of switches of numbers. For example, given indexes {1, 2, 3}, the permutation (123) is even because there is 0 number of switches and the permutation (132) is even because there is 1 number of switches. According to Leibniz formula, the determinants of 2_x2 and 3_x3 matrices are:

$$|\boldsymbol{A}(2_x2)| = a_{11}a_{22} - a_{12}a_{21}$$

$$|\boldsymbol{A}(3_x3)| = a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}$$

Leibniz formula gets a huge of operators due to $n!$ permutations. So matrix determinant is computed more effectively by applying Laplace expansion. Given element a_{ij} of square matrix $\boldsymbol{A}(n_xn)$, the algebra complement denoted $\boldsymbol{M}_{ij}$ is the sub-matrix including $n - 1$ rows and $n - 1$ columns, which is create by removing i^{th} row and j^{th} column from $\boldsymbol{A}$ (Nguyen, 1999, p. 130).

$$\boldsymbol{M}_{ij} = \begin{pmatrix} a_{11} & \cdots & a_{1(j-1)} & a_{1(j+1)} & \ldots & a_{1n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{(i-1)1} & \cdots & a_{(i-1)(j-1)} & a_{(i-1)(j+1)} & \ldots & a_{(i-1)n} \\ a_{(i+1)1} & \cdots & a_{(i+1)(j-1)} & a_{(i+1)(j+1)} & \cdots & a_{(i+1)n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{n(j-1)} & a_{n(j+1)} & \ldots & a_{nn} \end{pmatrix}$$

$\boldsymbol{M}_{ij}$ is called algebra complement or complement matrix of $\boldsymbol{A}$ at entry (i, j). By the similar way, the elementary matrix denoted $\boldsymbol{E}_{ij}$ of $\boldsymbol{A}$ at entry (i, j) is defined as the matrix having all zero elements except element a_{ij}.

$$\boldsymbol{E}_{ij} = \begin{pmatrix} 0 & 0 & \ldots & 0 \\ 0 & 0 & \vdots & 0 \\ \vdots & \vdots & a_{ij} & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

According to Laplace extension, determinant of n_xn square matrix $\boldsymbol{A}$ is computed according to determinants of algebra complements given arbitrary i^{th} row.

$$|\boldsymbol{A}| = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} |\boldsymbol{M}_{ij}|$$

Formula above is recursive formula with regard to column expansion. According to row expansion, determinant of square matrix $\boldsymbol{A}(n_xn)$ given arbitrary j^{th} column is:

$$|\boldsymbol{A}| = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} |\boldsymbol{M}_{ij}|$$

Note that the expression $(-1)^{i+j}|\boldsymbol{M}_{ij}|$ is called cofactor c_{ij} (Hardle & Simar, 2013, p. 60). Cofactor matrix of $\boldsymbol{A}$ denoted *cofactor*($\boldsymbol{A}$) is defined as matrix whose elements are cofactor c_{ij} (s). The adjoint matrix of $\boldsymbol{A}$ denoted *adj*($\boldsymbol{A}$) is the transposition of cofactor matrix of $\boldsymbol{A}$.

$$adj(\boldsymbol{A}) = (cofactor(\boldsymbol{A}))^T = (c_{ji})$$

Given square matrix $\boldsymbol{A}(n_x n)$ and $|\boldsymbol{A}| \neq 0$, the inverse of $\boldsymbol{A}$ denoted $\boldsymbol{A}^{-1}$ is the one that multiplication of $\boldsymbol{A}$ and itself is equal to identical matrix.

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_n$$

The inverse $\boldsymbol{A}^{-1}$ is determined according to adjoint matrix.

$$\boldsymbol{A}^{-1} = \frac{1}{|\boldsymbol{A}|} adj(\boldsymbol{A})$$

If there is existence of the inverse of $\boldsymbol{A}$, in other words $|\boldsymbol{A}| \neq 0$, then $\boldsymbol{A}$ is invertible or non-singular. If $\boldsymbol{A}$ is orthogonal matrix, it is very easy to determine the inverse because of $\boldsymbol{A}^{-1} = \boldsymbol{A}^T$. The generalized inverse (G-inverse) of matrix $\boldsymbol{A}$ denoted $\boldsymbol{A}^-$ is the one satisfying following condition (Hardle & Simar, 2013, p. 60).

$$\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}$$

Note that the inverse $\boldsymbol{A}^{-1}$ is a concrete case of G-inverse $\boldsymbol{A}^-$ and $\boldsymbol{A}^-$ is the general concept of the inverse $\boldsymbol{A}^{-1}$, hence, if $\boldsymbol{A}^{-1}$ exists then $\boldsymbol{A}^- = \boldsymbol{A}^{-1}$. The G-inverse is also called pseudo unique and always exists even though $\boldsymbol{A}^{-1}$ does not exist or $\boldsymbol{A}$ is not square matrix.

A typical application of matrix determinant and inverse is to solve set of linear equations. Matrix algebra is originated from research how to solve a set of linear equations but after that, theories of abstract algebra such as group, ring, field, module and vector space are applied into researching it and so matrix algebra is generalized as a large domain with a lot of applications. Given a set of n linear equations whose a_{ij} are coefficients and x_{ij} are unknowns need solved (Nguyen, 1999, pp. 136-138).

$$\begin{cases} a_{11}x_1 + a_{12}x_1 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_1 + \cdots + a_{2n}x_n = b_2 \\ \vdots \\ a_{n1}x_1 + a_{n2}x_1 + \cdots + a_{nn}x_n = b_1 \end{cases}$$

The set of equations above can be solved by Gaussian method including primary transformations such as changing order of two equations, multiplying a equation by scalar and adding to a equation by a linear combination of other equations so as to transform the set of equations into the form that is as simple as possible in order to find out unknowns.

In the report, we use another method called Cramer's rule to solve the set of linear equations. The set of n linear equations is re-written in matrix form.

$$\boldsymbol{Ax} = \boldsymbol{b}$$

Where $\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{21} & \vdots & a_{21} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n1} & a_{n1} & a_{nn} \end{pmatrix}$ is $n_x n$ matrix and

$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ is unknown vector and

$\boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$ is free − coefficient vector

We have a convention that $\boldsymbol{Ax} = \boldsymbol{b}$ is vector equation and $\boldsymbol{x}$ is unknown vector and $\boldsymbol{x}_0 = (x_{01}, x_{02}, \dots, x_{0n})$ is the solution vector and x_{0n} is called partial solution. Suppose matrix $\boldsymbol{A}$ is non-singular and so its inverse $\boldsymbol{A}^{-1}$ exists, Cramer's rule states that the number of partial solutions is the same to the number of partial equations; it means that matrix equation has unique solution $\boldsymbol{x}_0$ which has n partial solutions x_{0j} (s) and each solution x_{0j} is found out via determinants (Nguyen, 1999, p. 137):

$$x_{0j} = \frac{|\boldsymbol{A}_j|}{|\boldsymbol{A}|}$$

Where $\boldsymbol{A}_j$ is the matrix constructed by replacing j^{th} column in matrix by free-coefficient vector $\boldsymbol{b}$. For example, it is required to solve following equations (Nguyen, 1999, p. 138):

$$\begin{cases} x + y + 3z + 4t = -3 \\ x + y + 5z + 2t = 1 \\ 2x + y + 3z + 2t = -3 \\ 2x + 3y + 11z + 5t = 2 \end{cases}$$

The set of equations is written in matrix form $\boldsymbol{Ax} = \boldsymbol{b}$:

$$\begin{pmatrix} 1 & 1 & 3 & 4 \\ 1 & 1 & 5 & 2 \\ 2 & 1 & 3 & 2 \\ 2 & 3 & 11 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ t \end{pmatrix} = \begin{pmatrix} -3 \\ 1 \\ -3 \\ 2 \end{pmatrix}$$

Determinant of $\boldsymbol{A}$ is:

$$\begin{vmatrix} 1 & 1 & 3 & 4 \\ 1 & 1 & 5 & 2 \\ 2 & 1 & 3 & 2 \\ 2 & 3 & 11 & 5 \end{vmatrix} = -14$$

Solutions of these equations are totally solved as follows:

$$x = \frac{\begin{vmatrix} -3 & 1 & 3 & 4 \\ 1 & 1 & 5 & 2 \\ -3 & 1 & 3 & 2 \\ 2 & 3 & 11 & 5 \end{vmatrix}}{-14} = -2, y = \frac{\begin{vmatrix} 1 & -3 & 3 & 4 \\ 1 & 1 & 5 & 2 \\ 2 & -3 & 3 & 2 \\ 2 & 2 & 11 & 5 \end{vmatrix}}{-14} = 0,$$

$$z = \frac{\begin{vmatrix} 1 & 1 & -3 & 4 \\ 1 & 1 & 1 & 2 \\ 2 & 1 & -3 & 2 \\ 2 & 3 & 2 & 5 \end{vmatrix}}{-14} = 1, y = \frac{\begin{vmatrix} 1 & 1 & 3 & -3 \\ 1 & 1 & 5 & 1 \\ 2 & 1 & 3 & -3 \\ 2 & 3 & 11 & 2 \end{vmatrix}}{-14} = -1$$

Following are properties of matrix determinant and matrix inverse with note that determinant exists if and only if matrix is invertible.

$|\boldsymbol{I}_n| = 1$
$|\boldsymbol{A}^T| = |\boldsymbol{A}|$
$|\boldsymbol{A}^{-1}| = |\boldsymbol{A}|^{-1}$
$|\boldsymbol{A}| \neq 0 \Leftrightarrow rank(\boldsymbol{A}) = n \Leftrightarrow \boldsymbol{A}$ non $-$ singular $\Leftrightarrow \boldsymbol{A}$ invertible
$|\boldsymbol{AB}| = |\boldsymbol{A}||\boldsymbol{B}|$
$|c\boldsymbol{A}| = c^n|\boldsymbol{A}|$ where c is scalar constant.
$|\boldsymbol{A}^n| = |\boldsymbol{A}|^n$
$|\boldsymbol{A}| = \prod_{i=1}^{n} a_{ii}$ if $\boldsymbol{A}$ is triangular matrix.

$(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$
$(\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T$
$|\boldsymbol{I} + \boldsymbol{ab}^T| = 1 + \boldsymbol{a}^T\boldsymbol{b}$ where $\boldsymbol{I}$ is $n_x n$ identity matrix and $\boldsymbol{a}$, $\boldsymbol{b}$ are vectors.

2. Matrix analysis

The first section introduces basic concepts relevant to matrix algebra. Pure matrix algebra uses theories of abstract algebra such as group, ring, module, and field to explain and solve theoretical problems of matrix. Matrix analysis and matrix calculus mentioned in this report are typical applications of matrix algebra. Matrix analysis focuses on processing and analyzing multivariate data while matrix calculus focuses on derivative and differential with regard to matrix. Matrix analysis is very necessary to matrix calculus, in which spectrum decomposition is the most important. This section includes four parts such as spectrum decomposition, singular decomposition, quadratic form and partitioned matrix. Its main contents are extracted from the book "Applied Multivariate Statistical Analysis" by authors (Hardle & Simar, 2013) which is a valued document that readers should read.

Spectrum decomposition

Matrix analysis consists of decomposing matrix techniques which transform a complex matrix into a set of simple matrices because it is easy to process partially on simple matrices instead of whole complex matrix. This is problem of spectrum decomposition. Matrix analysis starts with concepts of eigenvalues and eigenvectors. Given square matrix $\boldsymbol{A}(n_x n)$, if there is a scalar value $\lambda > 0$ and a vector $\boldsymbol{u}$ such that

$$\boldsymbol{A} = \lambda \boldsymbol{u}$$

Then λ and $\boldsymbol{u}$ are eigenvalue and eigenvector of matrix $\boldsymbol{A}$, respectively. If $|\boldsymbol{A}| \neq 0$ then there are n eigenvalues and n respective eigenvectors. At that time, the vector space determined by matrix $\boldsymbol{A}$ is decomposed into n disjoint sub-spaces and each sub-space is specified by a pair of eigenvalue and eigenvector. There is a question how to find out n eigenvalues and n respective eigenvectors and so, eigenvalues are solutions of following equation:

$|\boldsymbol{A} - \lambda \boldsymbol{I}_n| = 0$ where $\boldsymbol{I}_n$ is $n_x n$ identity matrix.

Determinant $|\boldsymbol{A} - \lambda \boldsymbol{I}_n|$ is expanded as a n^{th} order polynomial which has n solutions $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ with attention that λ_i (s) should be in descending ordering, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Such polynomial is called characteristic polynomial of matrix $\boldsymbol{A}$ with respect to (w.r.t) variable scalar λ, denoted $P(\lambda) = |\boldsymbol{A} - \lambda \boldsymbol{I}_n|$. Suppose a solution of $P(\lambda) = |\boldsymbol{A} - \lambda \boldsymbol{I}_n|$ is λ_i, the respective eigenvector $\boldsymbol{u}_i$ is vector solution of following set of linear equations:

$$(\boldsymbol{A} - \lambda_i \boldsymbol{I}_n)\boldsymbol{x} = \boldsymbol{0} \Leftrightarrow \begin{pmatrix} a_{11} - \lambda_{i1} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda_{i2} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda_{in} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\Leftrightarrow \begin{cases} (a_{11} - \lambda_{i1})x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = 0 \\ a_{21}x_1 + (a_{22} - \lambda_{i2})x_2 + \cdots + a_{2n}x_n = 0 \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + (a_{nn} - \lambda_{in})x_n = 0 \end{cases}$$

In solution space of equation $(\boldsymbol{A} - \lambda \boldsymbol{I}_n)\boldsymbol{x} = \boldsymbol{0}$, we should choose eigenvectors $\boldsymbol{u}_i$ so that they are mutually orthogonal and all of eigenvectors $\boldsymbol{u}_i$ must be normalized. Maybe some eigenvectors in solution space are not mutually orthogonal but we assume that the mentioned eigenvectors are mutually orthogonal in whole report. Let $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, ..., \lambda_n)$ be the eigenvalue matrix and let $\boldsymbol{U}$ be the orthogonal matrix created by n eigenvectors. $\boldsymbol{U}$ is called eigenvector matrix. Both $\boldsymbol{\Lambda}$ and $\boldsymbol{U}$ are spectrums of $\boldsymbol{A}$.

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & ... & \lambda_n \end{pmatrix}$$

$$\boldsymbol{U} = (\boldsymbol{u_1}, \boldsymbol{u_2}, ..., \boldsymbol{u_n}) = \begin{pmatrix} u_{11} & u_{21} & ... & u_{n1} \\ u_{12} & u_{22} & ... & u_{n2} \\ \vdots & \vdots & \vdots & \vdots \\ u_{1n} & u_{2n} & ... & u_{nn} \end{pmatrix}$$

where $\boldsymbol{u_i} = \begin{pmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{in} \end{pmatrix}$ is a column normalized eigenvector

The *Jordan decomposition theorem* (Hardle & Simar, 2013, p. 63) states that

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{-1} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{T} = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$$

This theorem is the most important theorem in matrix analysis and it is a base of many decomposition techniques. Jordan decomposition exists if and only if $\boldsymbol{A}$ is square matrix and non-singular, $|\boldsymbol{A}| \neq 0$ and such matrix $\boldsymbol{A}$ is call *diagonalizable* or diagonalized matrix. In general, Jordan decomposition is essentially spectrum decomposition. Author (Hoang, 2012) gives an example for illustrating Jordan decomposition or diagonalizing matrix as follows:

Suppose $\boldsymbol{A} = \begin{pmatrix} 3 & -2 & 0 \\ -2 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix}$ and its characteristic polynomial $P(\lambda)$ is:

$$P(\lambda) = |\boldsymbol{A} - \lambda \boldsymbol{I}_3| = \begin{vmatrix} 3-\lambda & -2 & 0 \\ -2 & 3-\lambda & 0 \\ 0 & 0 & 5-\lambda \end{vmatrix} = (1-\lambda)(5-\lambda)^2$$

Solving equation $P(\lambda) = 0$ results out three solutions λ_1, λ_2 and λ_3 which are eigenvalues.

$$P(\lambda) = 0 \Leftrightarrow \begin{bmatrix} \lambda_1 = 1 \\ \lambda_2 = \lambda_3 = 5 \text{ (dual)} \end{bmatrix}$$

Substituting λ_1, λ_2 and λ_3 into the equation $(\boldsymbol{A} - \lambda \boldsymbol{I_2})\boldsymbol{x} = \boldsymbol{0}$, we get eigenvectors $\boldsymbol{u}_1$, $\boldsymbol{u}_2$ and $\boldsymbol{u}_3$ as solutions and normalize them right after.

$$(\boldsymbol{A}-\lambda_1\boldsymbol{I}_3)\boldsymbol{x}=\boldsymbol{0}\Leftrightarrow\left(\begin{pmatrix}3&-2&0\\-2&3&0\\0&0&5\end{pmatrix}-1\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}\right)\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\boldsymbol{0}$$

$$\Rightarrow\boldsymbol{x}=x_1\begin{pmatrix}1\\1\\0\end{pmatrix}\Rightarrow\boldsymbol{u}_1=\begin{pmatrix}1/\sqrt{2}\\1/\sqrt{2}\\0\end{pmatrix}$$

$$(\boldsymbol{A}-\lambda_2\boldsymbol{I}_3)\boldsymbol{x}=\boldsymbol{0}\Leftrightarrow\left(\begin{pmatrix}3&-2&0\\-2&3&0\\0&0&5\end{pmatrix}-5\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}\right)\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\boldsymbol{0}$$

$$\Rightarrow\boldsymbol{x}=x_1\begin{pmatrix}1\\-1\\0\end{pmatrix}+x_3\begin{pmatrix}0\\0\\1\end{pmatrix}\Rightarrow\begin{cases}\boldsymbol{u}_2=\begin{pmatrix}1/\sqrt{2}\\-1/\sqrt{2}\\0\end{pmatrix}\\\boldsymbol{u}_3=\begin{pmatrix}0\\0\\1\end{pmatrix}\end{cases}$$

In general, we have eigenvalue matrix $\boldsymbol{\Lambda}$ and eigenvector matrix $\boldsymbol{U}$ as following.

$$\boldsymbol{U}=\begin{pmatrix}1/\sqrt{2}&1/\sqrt{2}&0\\1/\sqrt{2}&-1/\sqrt{2}&0\\0&0&1\end{pmatrix}\text{ and }\boldsymbol{\Lambda}=\begin{pmatrix}1&0&0\\0&5&0\\0&0&5\end{pmatrix}$$

It is easy for us to validate spectrum decomposition:

$$\begin{pmatrix}3&-2&0\\-2&3&0\\0&0&5\end{pmatrix}=\boldsymbol{A}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{-1}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{T}$$

$$=\begin{pmatrix}1/\sqrt{2}&1/\sqrt{2}&0\\1/\sqrt{2}&-1/\sqrt{2}&0\\0&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&5&0\\0&0&5\end{pmatrix}\begin{pmatrix}1/\sqrt{2}&1/\sqrt{2}&0\\1/\sqrt{2}&-1/\sqrt{2}&0\\0&0&1\end{pmatrix}$$

Matrices $\boldsymbol{U}$ and $\boldsymbol{\Lambda}$ are much simpler than matrix $\boldsymbol{A}$. Moreover, $\boldsymbol{U}$ and $\boldsymbol{\Lambda}$ are orthogonal and diagonal matrices and so they have many valuable properties.

$$tr(\boldsymbol{\Lambda})=\sum_{i=1}^{n}\lambda_i$$

$$rank(\boldsymbol{\Lambda})=n$$
$$\boldsymbol{U}^{-1}=\boldsymbol{U}^{T}$$
$$\boldsymbol{U}^{-1}\boldsymbol{U}=\boldsymbol{U}\boldsymbol{U}^{-1}=\boldsymbol{U}^{T}\boldsymbol{U}=\boldsymbol{U}\boldsymbol{U}^{T}=\boldsymbol{I}_n$$
$$|\boldsymbol{A}|=|\boldsymbol{\Lambda}|=\prod_{i=1}^{n}\lambda_i$$
$$|\boldsymbol{A}|^k=|\boldsymbol{\Lambda}|^k=\prod_{i=1}^{n}\lambda_i^k$$

$$\boldsymbol{\Lambda}^k = diag\left(\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k\right) = \begin{pmatrix} \lambda_1^k & 0 & \ldots & 0 \\ 0 & \lambda_2^k & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \lambda_n^k \end{pmatrix}$$

$$\boldsymbol{A}^k = \boldsymbol{U}\boldsymbol{\Lambda}^k\boldsymbol{U}^{-1} = \boldsymbol{U}\boldsymbol{\Lambda}^k\boldsymbol{U}^T = \sum_{i=1}^{n} \lambda_i^k \boldsymbol{u}_i \boldsymbol{u}_i^T$$

Singular value decomposition

When matrix $\boldsymbol{A}(m_x n)$ is not square matrix, there is a technique called singular value decomposition (SVD), a generalization of Jordan decomposition, which is used to decompose matrix $\boldsymbol{A}(m_x n)$. Let $\boldsymbol{\Lambda} = diag(\lambda_1^{1/2}, \lambda_2^{1/2}, \ldots, \lambda_r^{1/2})$ be the eigenvalue matrix where λ_i is eigenvalue of $\boldsymbol{A}\boldsymbol{A}^T$ and $\boldsymbol{A}^T\boldsymbol{A}$. Note that λ_i (s) are in descending ordering, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r$. In other words, r eigenvalues λ_i (s) are solutions of one of two following equations:

$$|\boldsymbol{A}\boldsymbol{A}^T - \lambda \boldsymbol{I}_m| = 0$$

$$|\boldsymbol{A}^T\boldsymbol{A} - \lambda \boldsymbol{I}_n| = 0$$

The number of eigenvalues is $r = rank(\boldsymbol{A}\boldsymbol{A}^T) = rank(\boldsymbol{A}^T\boldsymbol{A}) \leq min(m, n)$. There are r eigenvectors $\boldsymbol{u}_i$ (s) corresponding to r eigenvalues, which are solutions of equation:

$$(\boldsymbol{A}\boldsymbol{A}^T - \lambda_i \boldsymbol{I}_m)\boldsymbol{x} = \boldsymbol{0}$$

There are r eigenvectors $\boldsymbol{v}_i$ (s) corresponding to r eigenvalues, which are solutions of equation:

$$(\boldsymbol{A}^T\boldsymbol{A} - \lambda_i \boldsymbol{I}_n)\boldsymbol{x} = \boldsymbol{0}$$

Let $\boldsymbol{U}(m_x r)$ and $\boldsymbol{V}(n_x r)$ be eigenvector matrices of r eigenvectors $\boldsymbol{u}_i$ (s) and r eigenvectors $\boldsymbol{v}_i$ (s). It is easy to infer that columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are mutually orthogonal eigenvectors and each eigenvector $\boldsymbol{u}_i$ ($\boldsymbol{v}_i$) has m (n) components. The *singular value decomposition* (SVD) theorem states that (Hardle & Simar, 2013, p. 64):

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$$

Where,

$$\boldsymbol{U}_{m_x r} = (\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_r}) = \begin{pmatrix} u_{11} & u_{21} & \ldots & u_{r1} \\ u_{12} & u_{22} & \ldots & u_{r2} \\ \vdots & \vdots & \ddots & \vdots \\ u_{1m} & u_{2m} & \ldots & u_{rm} \end{pmatrix}$$

where $\boldsymbol{u_i} = \begin{pmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{in} \end{pmatrix}$ is a column normalized eigenvector

$$\boldsymbol{V}_{n_x r} = (\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_r}) = \begin{pmatrix} v_{11} & v_{21} & \ldots & v_{r1} \\ v_{12} & v_{22} & \ldots & v_{r2} \\ \vdots & \vdots & \ddots & \vdots \\ v_{1n} & v_{2n} & \ldots & v_{rn} \end{pmatrix}$$

where $\boldsymbol{v_i} = \begin{pmatrix} v_{i1} \\ v_{i2} \\ \vdots \\ v_{in} \end{pmatrix}$ is a column normalized eigenvector

$$\boldsymbol{\Lambda}_{n_x r} = \begin{pmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{\lambda_r} \end{pmatrix}$$

According to SVD theorem, any matrix can be decomposed into three simpler matrices. SVD is often used to reduce vector space. For instance, when you choose $k \ll r$, which means that k is much smaller than the number r of eigenvalues, the number of columns of $\boldsymbol{U}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{V}$ are reduced to be $k \ll r$. Let $\boldsymbol{\Lambda}'(k_x k)$, $\boldsymbol{U}'(m_x k)$ and $\boldsymbol{V}'(n_x k)$ are eigenvalue matrix and eigenvector matrices with respect to k eigenvalues, we have:

$$\boldsymbol{A}' = \boldsymbol{U}'\boldsymbol{\Lambda}'\boldsymbol{V}'^T$$

Where $\boldsymbol{A}'(m_x n)$ is approximated matrix of $\boldsymbol{A}$.

In common, given matrix $\boldsymbol{A}(m_x n)$ in n-dimensional vector space with suppose that $m < n$ and it is necessary to reduce the dimension of vector space as small as possible. It means that the number k of descending-order eigenvalues ($\lambda_1 > \lambda_2 > \dots > \lambda_k$) is as small as possible and so spectrums of $\boldsymbol{A}$ such as $\boldsymbol{U}'$, $\boldsymbol{\Lambda}'$ and $\boldsymbol{V}'$ are very small while re-created matrix $\boldsymbol{A}'$ is approximated to A. In other words, the dimension of mentioned vector space is reduced to k.

In practical, there is another version of SVD (Baker, 2013, pp. 14-23). Without loss of generality, suppose $m < n$ and let $r \leq n$ be the number of eigenvalues resulting from two equations $|\boldsymbol{A}\boldsymbol{A}^T - \lambda \boldsymbol{I}_m| = 0$ and $|\boldsymbol{A}\boldsymbol{A}^T - \lambda \boldsymbol{I}_m|$. If $r < n$, it is possible to set $n - r$ eigenvalues to be zero and so we suppose that $r = n$ without loss of generality. There are m normalized eigenvectors $\boldsymbol{u}_i$ (s) resulted from $(\boldsymbol{A}\boldsymbol{A}^T - \lambda_i \boldsymbol{I}_m)\boldsymbol{x} = \boldsymbol{0}$ and n normalized eigenvectors $\boldsymbol{v}_i$ (s) resulted from $(\boldsymbol{A}^T\boldsymbol{A} - \lambda_i \boldsymbol{I}_n)\boldsymbol{x} = \boldsymbol{0}$ and so two eigenmatrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are totally determined as below.

$$\boldsymbol{U}_{m_x m} = (\boldsymbol{u_1}, \boldsymbol{u_2}, \dots, \boldsymbol{u_n}) = \begin{pmatrix} u_{11} & u_{21} & \dots & u_{m1} \\ u_{12} & u_{22} & \dots & u_{r2} \\ \vdots & \vdots & \ddots & \vdots \\ u_{1m} & u_{2m} & \dots & u_{mm} \end{pmatrix}$$

where $\boldsymbol{u_i} = \begin{pmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{im} \end{pmatrix}$ is a column normalized eigenvector

$$\boldsymbol{V}_{n_x n} = (\boldsymbol{v_1}, \boldsymbol{v_2}, \dots, \boldsymbol{v_n}) = \begin{pmatrix} v_{11} & v_{21} & \dots & v_{n1} \\ v_{12} & v_{22} & \dots & v_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ v_{1n} & v_{2n} & \dots & v_{nn} \end{pmatrix}$$

where $\boldsymbol{v_i} = \begin{pmatrix} v_{i1} \\ v_{i2} \\ \vdots \\ v_{in} \end{pmatrix}$ is a column normalized eigenvector

$$\boldsymbol{\Lambda}_{m_x n} = \begin{pmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & \ddots & 0 & 0 & \ddots & 0 \\ 0 & 0 & \dots & \sqrt{\lambda_m} & 0 & \dots & 0 \end{pmatrix}$$

Please pay attention that $\boldsymbol{U}(m_x m)$ and $\boldsymbol{V}(n_x n)$ are orthogonal matrices because they are square matrices composed of orthogonal eigenvectors. Moreover eigenvalue matrix $\boldsymbol{\Lambda}(m_x n)$ is not square matrix and it uses only m eigenvalues among n eigenvalues in its subsidiary diagonal and remaining elements are zeros. Matrix $\boldsymbol{A}$ is decomposed by these matrices.

$$\boldsymbol{A} = \boldsymbol{U\Lambda V}^{-1} = \boldsymbol{U\Lambda V}^{T}$$

In this version, although eigenvalue matrix $\boldsymbol{\Lambda}$ is not diagonal, $\boldsymbol{U}(m_x m)$ and $\boldsymbol{V}(n_x n)$ are orthogonal matrices and so this gives us many valuable properties.

By reducing technique, it is possible to decrease the number of eigenvectors in matrix $\boldsymbol{V}$ from n to m, which makes matrix $\boldsymbol{V}$ become $m_x m$ orthogonal matrix and so $\boldsymbol{\Lambda}$ is $m_x m$ diagonal matrix. Because $\boldsymbol{\Lambda}$ is invertible in this situation, let $\boldsymbol{A}^{-} = \boldsymbol{V\Lambda}^{-1}\boldsymbol{U}^{T}$ and we have:

$$\boldsymbol{AA}^{-}\boldsymbol{A} = (\boldsymbol{U\Lambda V}^{T})(\boldsymbol{V\Lambda}^{-1}\boldsymbol{U}^{T})(\boldsymbol{U\Lambda V}^{T}) = \boldsymbol{U\Lambda V}^{T} = \boldsymbol{A}$$

We infer that $\boldsymbol{AA}^{-}\boldsymbol{A}$ is G-inverse of $\boldsymbol{A}$.

Moreover, if you want to reduce vector space as small as possible, you can choose k descending-order eigenvalues ($\lambda_1 > \lambda_2 > \dots > \lambda_k$) with $k < m$ and $k \ll n$, which means that k is smaller than m and much smaller than n. So reduced eigenvalue and eigenvector matrices $\boldsymbol{U'}$, $\boldsymbol{\Lambda'}$ and $\boldsymbol{V'}$ become very small.

$$\boldsymbol{U'}_{m_x k} = (\boldsymbol{u_1}, \boldsymbol{u_2}, \dots, \boldsymbol{u_k}) = \begin{pmatrix} u_{11} & u_{21} & \dots & u_{k1} \\ u_{12} & u_{22} & \dots & u_{r2} \\ \vdots & \vdots & \ddots & \vdots \\ u_{1m} & u_{2m} & \dots & u_{km} \end{pmatrix}$$

where $\boldsymbol{u_i} = \begin{pmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{im} \end{pmatrix}$ is a column normalized eigenvector

$$\boldsymbol{U'}_{n_x k} = (\boldsymbol{v_1}, \boldsymbol{v_2}, \dots, \boldsymbol{v_k}) = \begin{pmatrix} v_{11} & v_{21} & \dots & v_{k1} \\ v_{12} & v_{22} & \dots & v_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ v_{1n} & v_{2n} & \dots & v_{kn} \end{pmatrix}$$

where $\boldsymbol{v_i} = \begin{pmatrix} v_{i1} \\ v_{i2} \\ \vdots \\ v_{in} \end{pmatrix}$ is a column normalized eigenvector

$$\boldsymbol{\Lambda}'_{k_x k} = \begin{pmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{\lambda_k} \end{pmatrix}$$

Note that re-created $\boldsymbol{A'} = \boldsymbol{U'\Lambda'V'^T}$ is approximated to matrix $\boldsymbol{A}$. Author (Baker, 2013, p. 18) gives an example for illustrating SVD with matrix $\boldsymbol{A} = \begin{pmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{pmatrix}$. In order to find out eigenvalues and eigenvectors that compose eigenmatrix $\boldsymbol{U}$, we solve following equation.

$$|\boldsymbol{AA}^T - \lambda \boldsymbol{I}_2| = 0 \Leftrightarrow \left| \begin{pmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{pmatrix} \begin{pmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = 0 \Rightarrow \begin{bmatrix} \lambda_1 = 12 \\ \lambda_2 = 10 \end{bmatrix}$$

The eigenvectors $\boldsymbol{u}_i$ creating eigenmatrix $\boldsymbol{U}$ is found out via $\lambda_1 = 12$ and $\lambda_2 = 10$ as follows:

$$(\boldsymbol{AA}^T - 12\boldsymbol{I}_2)\boldsymbol{x} = \boldsymbol{0} \Leftrightarrow \left(\begin{pmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{pmatrix} \begin{pmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{pmatrix} - 12 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \boldsymbol{0}$$

$$\Rightarrow \boldsymbol{x} = x_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} \Rightarrow \boldsymbol{u}_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$$

$$(\boldsymbol{AA}^T - 10\boldsymbol{I}_2)\boldsymbol{x} = \boldsymbol{0} \Leftrightarrow \left(\begin{pmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{pmatrix} \begin{pmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{pmatrix} - 10 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \boldsymbol{0}$$

$$\Rightarrow \boldsymbol{x} = x_1 \begin{pmatrix} 1 \\ -1 \end{pmatrix} \Rightarrow \boldsymbol{u}_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$$

So eigenmatrix $\boldsymbol{U}$ is totally determined, $\boldsymbol{U} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$. Now we calculate second eigenmatrix $\boldsymbol{V}$ by solving following equation.

$$|\boldsymbol{A}^T\boldsymbol{A} - \lambda \boldsymbol{I}_2| = 0 \Leftrightarrow \left| \begin{pmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right| = 0 \Rightarrow \begin{bmatrix} \lambda_1 = 12 \\ \lambda_2 = 10 \\ \lambda_3 = 0 \end{bmatrix}$$

The eigenvectors $\boldsymbol{v}_i$ composing eigenmatrix $\boldsymbol{V}$ is found out via $\lambda_1 = 12$, $\lambda_2 = 10$ and $\lambda_3 = 0$ as follows:

$$(\boldsymbol{A}^T\boldsymbol{A} - 12\boldsymbol{I}_2)\boldsymbol{x} = \boldsymbol{0} \Leftrightarrow \left(\begin{pmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{pmatrix} - 12 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \boldsymbol{0}$$

$$\Rightarrow \boldsymbol{x} = x_3\begin{pmatrix}1\\2\\1\end{pmatrix} \Rightarrow \boldsymbol{v}_1 = \begin{pmatrix}1/\sqrt{6}\\2/\sqrt{6}\\1/\sqrt{6}\end{pmatrix}$$

$$(\boldsymbol{A}^T\boldsymbol{A} - 10)\boldsymbol{x} = \boldsymbol{0} \Leftrightarrow \left(\begin{pmatrix}3 & -1\\1 & 3\\1 & 1\end{pmatrix}\begin{pmatrix}3 & 1 & 1\\-1 & 3 & 1\end{pmatrix} - 10\begin{pmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{pmatrix}\right)\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix} = \boldsymbol{0}$$

$$\Rightarrow \boldsymbol{x} = x_2\begin{pmatrix}2\\-1\\0\end{pmatrix} \Rightarrow \boldsymbol{v}_2 = \begin{pmatrix}2/\sqrt{5}\\-1/\sqrt{5}\\0\end{pmatrix}$$

$$(\boldsymbol{A}^T\boldsymbol{A} - 0)\boldsymbol{x} = \boldsymbol{0} \Leftrightarrow \left(\begin{pmatrix}3 & -1\\1 & 3\\1 & 1\end{pmatrix}\begin{pmatrix}3 & 1 & 1\\-1 & 3 & 1\end{pmatrix} - 0\begin{pmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{pmatrix}\right)\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix} = \boldsymbol{0}$$

$$\Rightarrow \boldsymbol{x} = \begin{pmatrix}1\\2\\-5\end{pmatrix} \Rightarrow \boldsymbol{v}_3 = \begin{pmatrix}1/\sqrt{30}\\2/\sqrt{30}\\-5/\sqrt{30}\end{pmatrix}$$

So eigenmatrix $\boldsymbol{V}$ is totally determined, in general, we have eigenvalue matrix $\boldsymbol{\Lambda}$ and eigenvector matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ as follows:

$$\boldsymbol{U} = \begin{pmatrix}1/\sqrt{2} & 1/\sqrt{2}\\1/\sqrt{2} & -1/\sqrt{2}\end{pmatrix} \text{ and } \boldsymbol{\Lambda} = \begin{pmatrix}\sqrt{12} & 0 & 0\\0 & \sqrt{10} & 0\end{pmatrix}$$

$$\text{and } \boldsymbol{V} = \begin{pmatrix}1/\sqrt{6} & 2/\sqrt{5} & 1/\sqrt{30}\\2/\sqrt{6} & -1/\sqrt{5} & 2/\sqrt{30}\\1/\sqrt{6} & 0 & -5/\sqrt{30}\end{pmatrix}$$

It is easy to validate spectrum decomposition of matrix $\boldsymbol{A}(m_x n)$.

$$\begin{pmatrix}3 & 1 & 1\\-1 & 3 & 1\end{pmatrix} = \boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$$

$$= \begin{pmatrix}1/\sqrt{2} & 1/\sqrt{2}\\1/\sqrt{2} & -1/\sqrt{2}\end{pmatrix}\begin{pmatrix}\sqrt{12} & 0 & 0\\0 & \sqrt{10} & 0\end{pmatrix}\begin{pmatrix}1/\sqrt{6} & 2/\sqrt{6} & 1/\sqrt{6}\\2/\sqrt{5} & -1/\sqrt{5} & 0\\1/\sqrt{30} & 2/\sqrt{30} & -5/\sqrt{30}\end{pmatrix}$$

When reducing matrix $\boldsymbol{A}$, let $k = 2$ be the smaller number of eigenvalues while the total number of eigenvalues are 3, we have:

$$\boldsymbol{U}' = \begin{pmatrix}1/\sqrt{2} & 1/\sqrt{2}\\1/\sqrt{2} & -1/\sqrt{2}\end{pmatrix} \text{ and } \boldsymbol{\Lambda}' = \begin{pmatrix}\sqrt{12} & 0\\0 & \sqrt{10}\end{pmatrix} \text{ and } \boldsymbol{V}' = \begin{pmatrix}1/\sqrt{6} & 2/\sqrt{5}\\2/\sqrt{6} & -1/\sqrt{5}\\1/\sqrt{6} & 0\end{pmatrix}$$

And re-created matrix $\boldsymbol{A}'$ is:

$$\boldsymbol{A'} = \boldsymbol{U'\Lambda'V'^T} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} \sqrt{12} & 0 \\ 0 & \sqrt{10} \end{pmatrix} \begin{pmatrix} 1/\sqrt{6} & 2/\sqrt{6} & 1/\sqrt{6} \\ 2/\sqrt{5} & -1/\sqrt{5} & 0 \end{pmatrix} = \begin{pmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{pmatrix} = \boldsymbol{A}$$

If $k = 1$ and we have:

$$\boldsymbol{U'} = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \text{ and } \boldsymbol{\Lambda'} = \begin{pmatrix} \sqrt{12} \end{pmatrix} \text{ and } \boldsymbol{V'} = \begin{pmatrix} 1/\sqrt{6} \\ 2/\sqrt{6} \\ 1/\sqrt{6} \end{pmatrix}$$

The re-created matrix ***A'*** becomes:

$$\boldsymbol{A'} = \boldsymbol{U'\Lambda'V'^T} = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} \sqrt{12} \end{pmatrix} \begin{pmatrix} 1/\sqrt{6} & 2/\sqrt{6} & 1/\sqrt{6} \end{pmatrix} = \begin{pmatrix} 1 & 2 & 1 \\ 1 & 2 & 1 \end{pmatrix} \approx \boldsymbol{A}$$

Quadratic form

Another important subject in matrix analysis is quadratic form (Hardle & Simar, 2013, pp. 65-67), which is described in general way. Given a invertible matrix $\boldsymbol{A}(n_x n)$ and a random vector $\boldsymbol{x}$, the quadratic form of $\boldsymbol{x}$ denoted $Q(\boldsymbol{x})$ is:

$$Q(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j$$

Quadratic form is *positive definite* or *positive semi-definite* if $Q(\boldsymbol{x}) > 0$ or $Q(\boldsymbol{x}) \geq 0$, respectively. Matrix $\boldsymbol{A}$ is called positive definite denoted $\boldsymbol{A} > 0$ or positive semi-definite denoted $\boldsymbol{A} \geq 0$ if quadratic form $Q(\boldsymbol{x})$ of any vector with regard to $\boldsymbol{A}$ is positive definite or positive semi-definite, respectively. Otherwise, if $Q(\boldsymbol{x}) < 0$ then, quadratic form and matrix $\boldsymbol{A}$ are indefinite. Because $\boldsymbol{A}$ is invertible, it is decomposed into $\boldsymbol{A} = \boldsymbol{U\Lambda U}^T$ where $\boldsymbol{\Lambda}$ and $\boldsymbol{U}$ are eigenvalue matrix and eigenvector matrix. Let vector $\boldsymbol{y} = \boldsymbol{U}^T\boldsymbol{x}$, we have (Hardle & Simar, 2013, p. 65):

$$Q(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^T(\boldsymbol{U\Lambda U}^T)\boldsymbol{x} = \boldsymbol{y}^T\boldsymbol{\Lambda}\boldsymbol{y} = \sum_{i=1}^{n} \lambda_i y_i^2$$

Where λ_i (s) are eigenvalues.

We have some properties of quadratic form:

$\boldsymbol{A} > 0$ if and only if all eigenvalues λ_i (s) > 0

If $\boldsymbol{A} > 0$ then $\boldsymbol{A}^{-1}$ exists and $|\boldsymbol{A}| > 0$

Given invertible and symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ and given a constraint $\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} = 1$, the maximum (minimum) of $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ is the largest (smallest) eigenvalue of $\boldsymbol{B}^{-1}\boldsymbol{A}$. The vector maximizing (minimizing) $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ is the eigenvector which corresponds to largest (smallest) eigenvalue of $\boldsymbol{B}^{-1}\boldsymbol{A}$. We have (Hardle & Simar, 2013, p. 66):

$$\max_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{x}=1} \quad (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}) = \max(\lambda_1, \lambda_2, \dots, \lambda_{\mathrm{n}})$$

$$\min_{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{x}=1} \quad (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}) = \min(\lambda_1, \lambda_2, \dots, \lambda_{\mathrm{n}})$$

Partitioned matrix

Now we research partitioned matrix (Hardle & Simar, 2013, pp. 68-70) which is useful in matrix analysis. Given matrix $\boldsymbol{A}(m_x n)$ are composed of other matrices $\boldsymbol{A}_{ij}(m_{i\ x}\ n_j)$.

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}$$

Where $\boldsymbol{A}_{ij}$ (s) are matrices having m_i rows and n_i columns. $\boldsymbol{A}_{ij}$ (s) are called groups or sub-matrices and $\boldsymbol{A}$ is called partitioned matrix. Partitioning matrix technique performs matrix operations and determine properties and characteristics of matrix (determinant $|\boldsymbol{A}|$, inverse $\boldsymbol{A}^{-1}$, etc.) according to sub-matrices. Given partitioned matrix $\boldsymbol{B}(m_x n) = \begin{pmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{pmatrix}$ similar to $\boldsymbol{A}$, we have (Hardle & Simar, 2013, p. 69):

$$\begin{aligned} \boldsymbol{A} + \boldsymbol{B} &= \begin{pmatrix} \boldsymbol{A}_{11} + \boldsymbol{B}_{11} & \boldsymbol{A}_{12} + \boldsymbol{B}_{12} \\ \boldsymbol{A}_{21} + \boldsymbol{B}_{21} & \boldsymbol{A}_{22} + \boldsymbol{B}_{22} \end{pmatrix} \\ \boldsymbol{B}^T &= \begin{pmatrix} \boldsymbol{B}_{11}^T & \boldsymbol{B}_{21}^T \\ \boldsymbol{B}_{12}^T & \boldsymbol{B}_{22}^T \end{pmatrix} \\ \boldsymbol{A}\boldsymbol{B}^T &= \begin{pmatrix} \boldsymbol{A}_{11}\boldsymbol{B}_{11}^T + \boldsymbol{A}_{12}\boldsymbol{B}_{12}^T & \boldsymbol{A}_{11}\boldsymbol{B}_{21}^T + \boldsymbol{A}_{12}\boldsymbol{B}_{22}^T \\ \boldsymbol{A}_{21}\boldsymbol{B}_{11}^T + \boldsymbol{A}_{22}\boldsymbol{B}_{12}^T & \boldsymbol{A}_{21}\boldsymbol{B}_{21}^T + \boldsymbol{A}_{22}\boldsymbol{B}_{22}^T \end{pmatrix} \end{aligned}$$

If $\boldsymbol{A}$ is invertible, the inverse of $\boldsymbol{A}$ is (Hardle & Simar, 2013, p. 69):

$$\boldsymbol{A}^{-1} = \begin{pmatrix} \boldsymbol{A}^{11} & \boldsymbol{A}^{12} \\ \boldsymbol{A}^{21} & \boldsymbol{A}^{22} \end{pmatrix}$$

$$\text{Where} \begin{cases} \boldsymbol{A}^{11} = (\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1} \\ \boldsymbol{A}^{12} = -\boldsymbol{A}^{11}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1} \\ \boldsymbol{A}^{21} = -\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}^{11} \\ \boldsymbol{A}^{22} = \boldsymbol{A}_{22}^{-1} + \boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}^{11}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1} \end{cases}$$

If $\boldsymbol{A}_{11}$ is invertible, the determinant of $\boldsymbol{A}$ is:

$$|\boldsymbol{A}| = |\boldsymbol{A}_{11}||\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}|$$

If $\boldsymbol{A}_{22}$ is invertible, the determinant of $\boldsymbol{A}$ is:

$$|\boldsymbol{A}| = |\boldsymbol{A}_{22}||\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}|$$

Suppose partitioned matrix $\boldsymbol{B}$ is composed of a invertible matrix $\boldsymbol{A}(n_x n)$ and two $(n_x 1)$ vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ as follows:

$$\boldsymbol{B} = \begin{pmatrix} 1 & \boldsymbol{b}^T \\ \boldsymbol{a} & \boldsymbol{A} \end{pmatrix}$$

We have (Hardle & Simar, 2013, p. 69):

$$|\boldsymbol{B}| = |\boldsymbol{A} - \boldsymbol{a}\boldsymbol{b}^T| = |\boldsymbol{A}||1 - \boldsymbol{b}^T\boldsymbol{A}^{-1}\boldsymbol{a}|$$

$$(\boldsymbol{A} - \boldsymbol{a}\boldsymbol{b}^T)^{-1} = \boldsymbol{A}^{-1} + \frac{\boldsymbol{A}^{-1}\boldsymbol{a}\boldsymbol{b}^T\boldsymbol{A}^{-1}}{1 - \boldsymbol{b}^T\boldsymbol{A}^{-1}\boldsymbol{a}}$$

3. Matrix derivative

Recall that matrix analysis and matrix calculus are typical applications of matrix algebra. Matrix analysis focuses on processing and analyzing multivariate data while matrix calculus focuses on derivative and differential with regard to matrix. Matrix derivative is the basic concept of matrix calculus and differential, which extends principles of derivative from real number space $\mathbb{R}^n$ to vector space. Vector space is a general space in which vectors and matrices are elements in vector space. Real number or scalar is 1-component vector (matrix). This section describes and classifies matrix derivatives. Its main contents are extracted from the webpage (Wikipedia, Matrix calculus, 2014) which is a valued document that readers should read. The term "scalar" refers to a real number but the study can be extended to complex number. Given a function f from n-dimension domain space $\mathbb{R}^n$ to real image space R, in other words, domain of f is vector space and image of f is scalar space. For convenience, we have convention that vector space $\mathbb{R}^n$ is based on real number field $\mathbb{R}$ and default vector space is $\mathbb{R}^n$ having n dimensions if there is no note.

f: $\mathbb{R}^n \rightarrow \mathbb{R}$

Thus, f is a scalar function with respect to (w.r.t) vector variable $\boldsymbol{x}$. Note that the domain space $\mathbb{R}^n$ can be reduced to scalar space $\mathbb{R}$ and f becomes scalar function over scalar space. The row derivative of f with respect to $\boldsymbol{x}$ is defined as below:

$$\frac{\partial f}{\partial \boldsymbol{x}} = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}\right)$$

So the row derivative of f w.r.t $\boldsymbol{x}$ is a row vector whose components are partial derivatives $\frac{\partial f}{\partial x_i}$ where x_i (s) are components of $\boldsymbol{x}$. Similarly, the column derivative of f w.r.t $\boldsymbol{x}$ is:

$$\frac{\partial f}{\partial \boldsymbol{x}^T} = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

The derivative of scalar function f w.r.t vector variable $\boldsymbol{x}$ is often called *gradient* of f denoted *gradf* or ∇f. Gradient ∇f is often known as row derivative vector. In general, we have seven ways to denote derivative as following.

$$f'(\boldsymbol{x}) = gradf = \nabla f = Df = \frac{df}{d\boldsymbol{x}} = \frac{\partial f}{\partial \boldsymbol{x}} = \left(\frac{\partial f}{\partial \boldsymbol{x}^T}\right)^T$$

Df and $\frac{df}{d\boldsymbol{x}}$ are formal notations originated from differential operation which is only applied into real number field $\mathbb{R}$ and vector space $\mathbb{R}^n$. When derivative gets value at $\boldsymbol{x}_0$, we use following notations with the same meaning:

$$f'(\boldsymbol{x}_0) = gradf(\boldsymbol{x}_0) = \nabla f(\boldsymbol{x}_0) = Df(\boldsymbol{x}_0) = \frac{df}{d\boldsymbol{x}}(\boldsymbol{x}_0) = \frac{\partial f}{\partial \boldsymbol{x}}(\boldsymbol{x}_0) = \left(\frac{\partial f}{\partial \boldsymbol{x}^T}\right)^T (\boldsymbol{x}_0)$$

The notation $f'(\boldsymbol{x})$ is the most popular in number analysis but we prefer to use notation $\frac{\partial f}{\partial \boldsymbol{x}}$ and $\frac{\partial f}{\partial \boldsymbol{x}^T}$ in matrix calculus. When the image space is extended to vector space $\mathbb{R}^m$, we have vector function $\boldsymbol{f}$.

$$f: \quad \mathbb{R}^n \quad \longrightarrow \quad \mathbb{R}^m$$

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} \longmapsto \boldsymbol{f} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{pmatrix}$$

Each elemental function $f_i \in \boldsymbol{f}$ may be the scalar function of all x_j (s). The derivative of vector function $\boldsymbol{f}$ with respective to vector variable $\boldsymbol{x}$ is a so-called Jacobian matrix.

$$\boldsymbol{f}'(\boldsymbol{x}) = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

Where $\frac{\partial f_i}{\partial x_j}$ is the partial derivative of i^{th} functional component f_i with respect to j^{th} domain component x_j. Aforementioned scalar function f and vector function $\boldsymbol{f}$ are typical examples for matrix derivative. If variable $\boldsymbol{x}$ in domain space is considered independent variable and function f in image space is considered dependent variable. We follow the same convention that "normal letters denote scalar; vectors are denoted by bold letters and matrices are denoted as bold and uppercase letters". There are nine possible combinations (Wikipedia, Matrix calculus, 2014) between independent variable (function) and dependent variable, each of combination corresponds a kind of matrix derivative when scalar and vector are simple forms of matrix. Concretely, scalar is 1_x1 matrix, column vector is n_x1 matrix and row vector is 1_xn matrix.

Dependent variable (*function*)	*Independent variable*		
	Scalar	Vector	Matrix
Scalar	$\frac{\partial f}{\partial x}$	$\frac{\partial f}{\partial \boldsymbol{x}}$	$\frac{\partial f}{\partial \boldsymbol{X}}$
Vector	$\frac{\partial \boldsymbol{f}}{\partial x}$	$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$	$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{X}}$
Matrix	$\frac{\partial \boldsymbol{F}}{\partial x}$	$\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{x}}$	$\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{X}}$

In table (Wikipedia, Matrix calculus, 2014) above, there are six descriptive kinds of derivative: scalar-by-scalar (scalar function w.r.t scalar variable), scalar-by-vector (scalar function w.r.t vector variable), scalar-by-matrix (scalar function w.r.t matrix variable), vector-by-scalar (vector function

w.r.t scalar variable), vector-by-vector (vector function w.r.t vector variable) and matrix-by-scalar (matrix function w.r.t vector variable). Three remaining kinds corresponding to shaded cells such as vector-by-matrix, matrix-by-vector and matrix-by-matrix, which relate to tensor product, are not focused in the report. We divide these kinds of function into three groups:

- Scalar function group includes scalar-by-scalar, scalar-by-vector and scalar-by-matrix.
- Vector function group includes vector-by-scalar, vector-by-vector and vector-by-matrix.
- Matrix function group includes matrix-by-scalar, matrix-by-vector and matrix-by-matrix.

Suppose that all mentioned functions (dependent variables) belong to differentiability class, which means that their k^{th} derivatives exist and continuous. Before discussing matrix derivative in more detailed, it is convenient for us to apply *numerator layout convention* (Wikipedia, Matrix calculus, 2014) into denoting multivariate derivatives. With numerator layout convention, given the vector-by-vector function from $\mathbb{R}^n$ to $\mathbb{R}^m$, the derivative $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$ is a Jacobian matrix in which the numerator $\boldsymbol{f}$ (dependent variable) lists its partial derivatives $\begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{pmatrix}$ as column vector and the independent variable $\boldsymbol{x}$ lists its components as row vector $(x_1, x_2, \ldots, x_n)$. In other words, numerator layout convention results out partial derivatives according to column alignment with $\boldsymbol{f}$ and row alignment with $\boldsymbol{x}^T$.

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

Let $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ as dependent variable, we have convention that $\boldsymbol{y}$ is identical to $\boldsymbol{f}$. Please pay attention to this convention because we will often use dependent variable $\boldsymbol{y}$ instead of $\boldsymbol{f}$.

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}$$

Now we discuss more about six kinds of derivative: scalar-by-scalar, scalar-by-vector, scalar-by-matrix, vector-by-scalar, vector-by-vector and matrix-by-scalar.

Scalar-by-scalar derivative

Let f: $\mathbb{R} \rightarrow \mathbb{R}$ be scalar-by-scalar function, the derivative of scalar y by scalar x is $f\,'(x) = \frac{\partial y}{\partial x}$, which is the simplest case and so it is not mentioned much in matrix calculus. If $x = x(t)$ is scalar function of $t \in R$, the derivative of f with respect to variable t is $f\,'(t) = f\,'(x)x'(t) = \frac{\partial y}{\partial x}\frac{\partial x}{\partial t}$. This is chain rule which is applied into all kinds of derivative. The second-order derivative of f is $f\,''(x) = \frac{\partial^2 y}{\partial x^2}$. It is necessary to distinguish between second-order derivative and second power of derivative, $\frac{\partial^2 y}{\partial x^2} \neq \left(\frac{\partial y}{\partial x}\right)^2$.

Scalar-by-vector derivative

Let f: $\mathbb{R}^n \rightarrow \mathbb{R}$ be scalar-by-vector function, the derivative of scalar y by vector $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ is:

$$\frac{\partial y}{\partial \boldsymbol{x}} = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n}\right)$$

This derivative is also called *gradient* of f, denoted *gradf* or ∇f. Scalar-by-vector function is often called multivariate function. The *directional derivative* of scalar-by-vector function is dot product between gradient ∇f and directional vector $\boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$, which is denoted $\nabla_{\boldsymbol{u}} f$.

$$\nabla_{\boldsymbol{u}} f = \nabla f(\boldsymbol{x})\boldsymbol{u} = \frac{\partial y}{\partial \boldsymbol{x}}\boldsymbol{u} = \left(\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \dots, \frac{\partial y}{\partial x_n}\right)\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

The second derivative of scalar-by-vector function is a so-called Hessian matrix whose elements are partial second-order derivatives w.r.t partial variables.

$$f''(\boldsymbol{x}) = \nabla^2 f = \frac{\partial^2 y}{\partial \boldsymbol{x}^2} = \begin{pmatrix} \frac{\partial^2 y}{\partial x_1^2} & \frac{\partial^2 y}{\partial x_1 x_2} & \cdots & \frac{\partial^2 y}{\partial x_1 x_n} \\ \frac{\partial^2 y}{\partial x_2 x_1} & \frac{\partial^2 y}{\partial x_2^2} & \cdots & \frac{\partial^2 y}{\partial x_2 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 y}{\partial x_n x_1} & \frac{\partial^2 y}{\partial x_n x_2} & \cdots & \frac{\partial^2 y}{\partial x_n^2} \end{pmatrix}$$

Where $\frac{\partial^2 y}{\partial x_i^2}$ is the second derivative w.r.t partial variable x_i and $\frac{\partial^2 y}{\partial x_i x_j} = \frac{\partial}{\partial x_j}\left(\frac{\partial y}{\partial x_i}\right)$ is the iterative derivative that y is taken derivative with respect to x_i, which in turn, is taken with respect to x_j.

Scalar-by-matrix derivative

Let f: $\mathbb{R}^{n_x m} \to \mathbb{R}$ be scalar-by-matrix function, the derivative of scalar y by matrix $\boldsymbol{X}(m_x n) = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{pmatrix}$ is:

$$\frac{\partial y}{\partial \boldsymbol{X}} = \nabla f = \begin{pmatrix} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{21}} & \dots & \frac{\partial y}{\partial x_{m1}} \\ \frac{\partial y}{\partial x_{12}} & \frac{\partial y}{\partial x_{22}} & \dots & \frac{\partial y}{\partial x_{m2}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y}{\partial x_{1n}} & \frac{\partial y}{\partial x_{2n}} & \dots & \frac{\partial y}{\partial x_{mn}} \end{pmatrix}$$

Where $\frac{\partial y}{\partial x_{ij}}$ is the partial derivative of y w.r.t partial element x_{ij} of matrix variable $\boldsymbol{X}$. While $\boldsymbol{X}$ is $m_x n$ matrix, the derivative $\frac{\partial y}{\partial \boldsymbol{X}}$ is $n_x m$ matrix. This follows the numerator layout convention and so element indices in derivative matrix $\frac{\partial y}{\partial \boldsymbol{X}}$ are transposed as compared with element indices in variable $\boldsymbol{X}$. In similar to scalar-by-vector function, the scalar-by-matrix derivative is also called *gradient matrix* ∇f of f. The directional derivative of scalar y by matrix $\boldsymbol{X}(m_x n)$ in the direction of matrix $\boldsymbol{U}(m_x n)$ is defined via the trace of matrix mentioned in previous section. So the directional derivative denoted $\nabla_{\boldsymbol{U}} f$ is a scalar and its formula is:

$$\nabla_{\boldsymbol{U}} f = \nabla f(\boldsymbol{X})\boldsymbol{U} = tr\left(\frac{\partial y}{\partial \boldsymbol{X}}\boldsymbol{U}\right)$$

Note that the above formula is also true to directional derivative of scalar-by-vector function as aforementioned when the trace of scalar is itself. We conclude that concept "gradient" always goes along with concept "directional derivative" in scalar function and the evaluation of directional derivative is always scalar.

Vector-by-scalar derivative

Let $\boldsymbol{f}$: $\mathbb{R} \to \mathbb{R}^n$ be vector-by-scalar function, the derivative of vector $\boldsymbol{y} = \boldsymbol{f}(x) = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ by scalar x is:

$$\frac{\partial \boldsymbol{y}}{\partial x} = \begin{pmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \\ \vdots \\ \frac{\partial y_n}{\partial x} \end{pmatrix}$$

Where $\frac{\partial y_i}{\partial x}$ is the derivative of partial element y_i w.r.t scalar x. When $\boldsymbol{f}$ is vector function, its derivative $\frac{\partial \boldsymbol{y}}{\partial x}$ is known as *tangent vector*. Vector function is applied into representing curve and surface in differential geometry and tangent vector is important component w.r.t geometrical object. The second-order derivative of vector-by-scalar function is:

$$\frac{\partial^2 \boldsymbol{y}}{\partial x^2} = \begin{pmatrix} \frac{\partial^2 y_1}{\partial x^2} \\ \frac{\partial^2 y_2}{\partial x^2} \\ \vdots \\ \frac{\partial^2 y_n}{\partial x^2} \end{pmatrix}$$

Vector-by-vector derivative

Given another vector function $\boldsymbol{f}$: $\mathbb{R}^n \rightarrow \mathbb{R}^m$ is vector-by-vector function from space $\mathbb{R}^n$ to space $\mathbb{R}^m$. The derivative of vector $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$ by scalar $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix}$ is:

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}$$

Where $\frac{\partial y_i}{\partial x_j}$ is the partial derivative of i^{th} functional component y_i with respect to j^{th} domain component x_j. The derivative matrix as aforementioned is known Jacobian matrix, push-forward matrix, or differential matrix. Jacobian matrix is the heart of matrix calculus. If we consider vector function $\boldsymbol{y}$ is the composition of partial scalar function y_1, y_2,…, y_m then Jacobian matrix is the column vector of gradients $\nabla f_i = \frac{\partial y_i}{\partial \boldsymbol{x}}$.

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{pmatrix} \nabla f_1 \\ \nabla f_2 \\ \vdots \\ \nabla f_m \end{pmatrix} = \begin{pmatrix} \frac{\partial y_1}{\partial \boldsymbol{x}} \\ \frac{\partial y_2}{\partial \boldsymbol{x}} \\ \vdots \\ \frac{\partial y_m}{\partial \boldsymbol{x}} \end{pmatrix} \text{ where gradient } \nabla f_i = \frac{\partial y_i}{\partial \boldsymbol{x}} = \left(\frac{\partial y_i}{\partial x_1}, \frac{\partial y_i}{\partial x_2}, \ldots, \frac{\partial y_i}{\partial x_n} \right)$$

Matrix-by-scalar derivative

Let $\boldsymbol{F}$: $\mathbb{R} \to \mathbb{R}^{n_x m}$ be matrix-by-scalar function, the derivative of matrix $\boldsymbol{Y} = \boldsymbol{F}(x) = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mn} \end{pmatrix}$ by scalar x is:

$$\frac{\partial \boldsymbol{Y}}{\partial x} = \begin{pmatrix} \frac{\partial y_{11}}{\partial x} & \frac{\partial y_{12}}{\partial x} & \cdots & \frac{\partial y_{1n}}{\partial x} \\ \frac{\partial y_{21}}{\partial x} & \frac{\partial y_{22}}{\partial x} & \cdots & \frac{\partial y_{2n}}{\partial x} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y_{m1}}{\partial x} & \frac{\partial y_{m2}}{\partial x} & \cdots & \frac{\partial y_{mn}}{\partial x} \end{pmatrix}$$

Where $\frac{\partial y_{ij}}{\partial x}$ is the derivative of partial element y_{ij} w.r.t scalar x. When $\boldsymbol{F}$ is matrix function, its derivative $\frac{\partial \boldsymbol{F}}{\partial x}$ is known as *tangent matrix*. The second-order derivative of matrix-by-scalar function is:

$$\frac{\partial^2 \boldsymbol{Y}}{\partial x^2} = \begin{pmatrix} \frac{\partial^2 y_{11}}{\partial x^2} & \frac{\partial^2 y_{12}}{\partial x^2} & \cdots & \frac{\partial^2 y_{1n}}{\partial x^2} \\ \frac{\partial^2 y_{21}}{\partial x^2} & \frac{\partial^2 y_{22}}{\partial x^2} & \cdots & \frac{\partial^2 y_{2n}}{\partial x^2} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 y_{m1}}{\partial x^2} & \frac{\partial^2 y_{m2}}{\partial x^2} & \cdots & \frac{\partial^2 y_{mn}}{\partial x^2} \end{pmatrix}$$

The derivative of matrix inverse denoted $\frac{\partial \boldsymbol{Y}^{-1}}{\partial x}$ is calculated based on tangent matrix $\frac{\partial \boldsymbol{Y}}{\partial x}$:

$$\frac{\partial \boldsymbol{Y}^{-1}}{\partial x} = -\boldsymbol{Y}^{-1} \frac{\partial \boldsymbol{Y}}{\partial x} \boldsymbol{Y}^{-1}$$

Now we researched six kinds of derivative such as scalar-by-scalar, scalar-by-vector, scalar-by-matrix, vector-by-scalar, vector-by-vector and matrix-by-scalar. There are three remaining kinds of derivative such as vector-by-matrix, matrix-by-vector and matrix-by-matrix not concerned much in the report. Moreover high-order derivatives such as second-order, third order and k^{th} order derivatives are mentioned in restriction because they relate to tensor product – a complicated subject which goes beyond this content. We will discuss tensor product, high-order derivative in detail in another report focusing on the subject of combination of tensor product and multidimensional derivative. However we should have an overview of some concepts of derivative relating to matrix such as *vector-by-matrix*, *matrix-by-vector* and *matrix-by-matrix*.

Vector-by-matrix derivative

Let $\boldsymbol{f}$: $\mathbb{R}^{m_x n} \rightarrow \mathbb{R}^p$ be vector-by-matrix function from $\mathbb{R}^{m_x n}$ space to $\mathbb{R}^p$ space, the derivative of vector $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{X}) = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}$ by matrix $\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$ is defined as follows:

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{X}} = \begin{pmatrix} \frac{\partial \boldsymbol{y}}{\partial x_{11}} & \frac{\partial \boldsymbol{y}}{\partial x_{21}} & \cdots & \frac{\partial \boldsymbol{y}}{\partial x_{m1}} \\ \frac{\partial \boldsymbol{y}}{\partial x_{12}} & \frac{\partial \boldsymbol{y}}{\partial x_{22}} & \cdots & \frac{\partial \boldsymbol{y}}{\partial x_{m2}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial \boldsymbol{y}}{\partial x_{1n}} & \frac{\partial \boldsymbol{y}}{\partial x_{2n}} & \cdots & \frac{\partial \boldsymbol{y}}{\partial x_{mn}} \end{pmatrix}$$

$$\text{where each element } \frac{\partial \boldsymbol{y}}{\partial x_{ij}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_{ij}} \\ \frac{\partial y_2}{\partial x_{ij}} \\ \vdots \\ \frac{\partial y_p}{\partial x_{ij}} \end{pmatrix} \text{ is a partial tangent vector}$$

Note that indices of matrix variable $\boldsymbol{X}$ are transposed due to numerator layout convention. The interesting discovery is that the vector-by-matrix derivative $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{X}}$ is compound gradient matrix whose components are partial tangent vectors.

Matrix-by-vector derivative

Let $\boldsymbol{F}$: $\mathbb{R}^n \rightarrow \mathbb{R}^{p_x q}$ be matrix-by-vector function from $\mathbb{R}^n$ space to $\mathbb{R}^{p_x q}$ space, the derivative of matrix $\boldsymbol{Y} = \boldsymbol{F}(\boldsymbol{x}) = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1q} \\ x_{21} & x_{22} & \cdots & x_{2q} \\ \vdots & \vdots & \vdots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pq} \end{pmatrix}$ by vector $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ is defined as follows:

$$\frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{x}} = \left(\frac{\partial \boldsymbol{Y}}{\partial x_1}, \frac{\partial \boldsymbol{Y}}{\partial x_2}, \ldots, \frac{\partial \boldsymbol{Y}}{\partial x_n} \right)$$

$$\text{where where each element } \frac{\partial \boldsymbol{Y}}{\partial x_i} = \begin{pmatrix} \frac{\partial y_{11}}{\partial x_i} & \frac{\partial y_{12}}{\partial x_i} & \cdots & \frac{\partial y_{1q}}{\partial x_i} \\ \frac{\partial y_{21}}{\partial x_i} & \frac{\partial y_{22}}{\partial x_i} & \cdots & \frac{\partial y_{2q}}{\partial x_i} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y_{p1}}{\partial x_i} & \frac{\partial y_{p2}}{\partial x_i} & \cdots & \frac{\partial y_{pq}}{\partial x_i} \end{pmatrix} \text{ is a partial tangent matrix}$$

The interesting discovery is that the matrix-by-vector derivative $\frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{x}}$ is compound gradient vector whose components are partial tangent matrices.

Matrix-by-matrix derivative

Given the most general case – matrix-by-matrix function $\boldsymbol{F}$: $\mathbb{R}^{n_x m} \rightarrow \mathbb{R}^{p_x q}$ from $\mathbb{R}^{n_x m}$ space to $\mathbb{R}^{p_x q}$ space, the derivative of matrix

$$\boldsymbol{Y} = \boldsymbol{F}(\boldsymbol{X}) = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1q} \\ x_{21} & x_{22} & \cdots & x_{2q} \\ \vdots & \vdots & \vdots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pq} \end{pmatrix} \text{ by matrix } \boldsymbol{X} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$$

is defined as follows:

$$\frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{X}} = \begin{pmatrix} \frac{\partial \boldsymbol{Y}}{\partial x_{11}} & \frac{\partial \boldsymbol{Y}}{\partial x_{21}} & \cdots & \frac{\partial \boldsymbol{Y}}{\partial x_{m1}} \\ \frac{\partial \boldsymbol{Y}}{\partial x_{12}} & \frac{\partial \boldsymbol{Y}}{\partial x_{22}} & \cdots & \frac{\partial \boldsymbol{Y}}{\partial x_{m2}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial \boldsymbol{Y}}{\partial x_{1n}} & \frac{\partial \boldsymbol{Y}}{\partial x_{2n}} & \cdots & \frac{\partial \boldsymbol{Y}}{\partial x_{mn}} \end{pmatrix} \text{ where } \frac{\partial \boldsymbol{Y}}{\partial x_{ij}} = \begin{pmatrix} \frac{\partial y_{11}}{\partial x_{ij}} & \frac{\partial y_{12}}{\partial x_{ij}} & \cdots & \frac{\partial y_{1q}}{\partial x_{ij}} \\ \frac{\partial y_{21}}{\partial x_{ij}} & \frac{\partial y_{22}}{\partial x_{ij}} & \cdots & \frac{\partial y_{2q}}{\partial x_{ij}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y_{p1}}{\partial x_{ij}} & \frac{\partial y_{p2}}{\partial x_{ij}} & \cdots & \frac{\partial y_{pq}}{\partial x_{ij}} \end{pmatrix}$$

The matrix-by-matrix derivative is $m_x n$ matrix whose elements are $p_x q$ matrices; in other words, the derivative is a fourth-rank tensor. The interesting discovery is that the matrix-by-matrix derivative $\frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{X}}$ is compound gradient matrix whose components are partial tangent matrices.

In general, following is the summary table concerning derivatives divided into 3 groups such as scalar function, vector function and matrix function with 9 kinds of functions: scalar-by-scalar, scalar-by-vector, scalar-by-matrix, vector-by-scalar, vector-by-vector, vector-by-matrix, matrix-by-scalar, matrix-by-vector and matrix-by-matrix.

Dependent variable (function)	Independent variable		
	Scalar	**Vector**	**Matrix**
Scalar	Derivative is: $\frac{\partial y}{\partial x} = f'(x)$ Second-order derivative is: $\frac{\partial^2 y}{\partial x^2} = f''(x)$	Derivative known as *gradient vector* is: $\frac{\partial y}{\partial \boldsymbol{x}} = \nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots, \frac{\partial f}{\partial x_n}\right)$ Directional derivative is: $\frac{\partial y}{\partial \boldsymbol{x}}\boldsymbol{u} = \nabla_u f = \nabla f(\boldsymbol{x})\boldsymbol{u} = \left(\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \cdots, \frac{\partial y}{\partial x_n}\right)\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$ Second-order derivative known as Hessian matrix is:	Derivative known as *gradient matrix* is: $\frac{\partial y}{\partial \boldsymbol{X}} = \nabla f = \begin{pmatrix} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{21}} & \cdots & \frac{\partial y}{\partial x_{m1}} \\ \frac{\partial y}{\partial x_{12}} & \frac{\partial y}{\partial x_{22}} & \cdots & \frac{\partial y}{\partial x_{m2}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y}{\partial x_{1n}} & \frac{\partial y}{\partial x_{2n}} & \cdots & \frac{\partial y}{\partial x_{mn}} \end{pmatrix}$ Directional derivative is: $\nabla_{\boldsymbol{U}} f = \nabla f(\boldsymbol{X})\boldsymbol{U} = tr\left(\frac{\partial y}{\partial \boldsymbol{X}}\boldsymbol{U}\right)$

		$\frac{\partial^2 y}{\partial \boldsymbol{x}^2} = \nabla^2 f = \begin{pmatrix} \frac{\partial^2 y}{\partial x_1^2} & \frac{\partial^2 y}{\partial x_1 x_2} & \cdots & \frac{\partial^2 y}{\partial x_1 x_n} \\ \frac{\partial^2 y}{\partial x_2 x_1} & \frac{\partial^2 y}{\partial x_2^2} & \cdots & \frac{\partial^2 y}{\partial x_2 x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 y}{\partial x_n x_1} & \frac{\partial^2 y}{\partial x_n x_2} & \cdots & \frac{\partial^2 y}{\partial x_n^2} \end{pmatrix}$	
Vector	Derivative known as *tangent vector* is: $\frac{\partial \boldsymbol{y}}{\partial x} = \begin{pmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \\ \vdots \\ \frac{\partial y_n}{\partial x} \end{pmatrix}$ Second-order derivative is: $\frac{\partial^2 \boldsymbol{y}}{\partial x^2} = \begin{pmatrix} \frac{\partial^2 y_1}{\partial x^2} \\ \frac{\partial^2 y_2}{\partial x^2} \\ \vdots \\ \frac{\partial^2 y_n}{\partial x^2} \end{pmatrix}$	Derivative known as *Jacobian matrix* is: $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix} =$ $\begin{pmatrix} \frac{\partial y_1}{\partial \boldsymbol{x}} \\ \frac{\partial y_2}{\partial \boldsymbol{x}} \\ \vdots \\ \frac{\partial y_m}{\partial \boldsymbol{x}} \end{pmatrix} = \begin{pmatrix} \nabla f_1 \\ \nabla f_2 \\ \vdots \\ \nabla f_m \end{pmatrix}$	Derivative known as compound gradient matrix is: $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{X}} = \begin{pmatrix} \frac{\partial \boldsymbol{y}}{\partial x_{11}} & \frac{\partial \boldsymbol{y}}{\partial x_{21}} & \cdots & \frac{\partial \boldsymbol{y}}{\partial x_{m1}} \\ \frac{\partial \boldsymbol{y}}{\partial x_{12}} & \frac{\partial \boldsymbol{y}}{\partial x_{22}} & \cdots & \frac{\partial \boldsymbol{y}}{\partial x_{m2}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial \boldsymbol{y}}{\partial x_{1n}} & \frac{\partial \boldsymbol{y}}{\partial x_{2n}} & \cdots & \frac{\partial \boldsymbol{y}}{\partial x_{mn}} \end{pmatrix}$ Where each element known as partial tangent vector is $\frac{\partial \boldsymbol{y}}{\partial x_{ij}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_{ij}} \\ \frac{\partial y_2}{\partial x_{ij}} \\ \vdots \\ \frac{\partial y_p}{\partial x_{ij}} \end{pmatrix}$
Matrix	Derivative known as *tangent matrix* is: $\frac{\partial \boldsymbol{Y}}{\partial x} = \begin{pmatrix} \frac{\partial y_{11}}{\partial x} & \frac{\partial y_{12}}{\partial x} & \cdots & \frac{\partial y_{1n}}{\partial x} \\ \frac{\partial y_{21}}{\partial x} & \frac{\partial y_{22}}{\partial x} & \cdots & \frac{\partial y_{2n}}{\partial x} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y_{m1}}{\partial x} & \frac{\partial y_{m2}}{\partial x} & \cdots & \frac{\partial y_{mn}}{\partial x} \end{pmatrix}$ Second-order derivative is: $\frac{\partial^2 \boldsymbol{Y}}{\partial x^2} = \begin{pmatrix} \frac{\partial^2 y_{11}}{\partial x^2} & \frac{\partial^2 y_{12}}{\partial x^2} & \cdots & \frac{\partial^2 y_{1n}}{\partial x^2} \\ \frac{\partial^2 y_{21}}{\partial x^2} & \frac{\partial^2 y_{22}}{\partial x^2} & \cdots & \frac{\partial^2 y_{2n}}{\partial x^2} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 y_{m1}}{\partial x^2} & \frac{\partial^2 y_{m2}}{\partial x^2} & \cdots & \frac{\partial^2 y_{mn}}{\partial x^2} \end{pmatrix}$ Derivative of inverse is:	Derivative known as compound gradient vector is: $\frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{x}} = \left(\frac{\partial \boldsymbol{Y}}{\partial x_1}, \frac{\partial \boldsymbol{Y}}{\partial x_2}, \cdots, \frac{\partial \boldsymbol{Y}}{\partial x_n} \right)$ Where each element known as partial tangent matrix is: $\frac{\partial \boldsymbol{Y}}{\partial x_i} = \begin{pmatrix} \frac{\partial y_{11}}{\partial x_i} & \frac{\partial y_{12}}{\partial x_i} & \cdots & \frac{\partial y_{1q}}{\partial x_i} \\ \frac{\partial y_{21}}{\partial x_i} & \frac{\partial y_{22}}{\partial x_i} & \cdots & \frac{\partial y_{2q}}{\partial x_i} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y_{p1}}{\partial x_i} & \frac{\partial y_{p2}}{\partial x_i} & \cdots & \frac{\partial y_{pq}}{\partial x_i} \end{pmatrix}$	Derivative known as compound gradient matrix is: $\frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{X}} = \begin{pmatrix} \frac{\partial \boldsymbol{Y}}{\partial x_{11}} & \frac{\partial \boldsymbol{Y}}{\partial x_{21}} & \cdots & \frac{\partial \boldsymbol{Y}}{\partial x_{m1}} \\ \frac{\partial \boldsymbol{Y}}{\partial x_{12}} & \frac{\partial \boldsymbol{Y}}{\partial x_{22}} & \cdots & \frac{\partial \boldsymbol{Y}}{\partial x_{m2}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial \boldsymbol{Y}}{\partial x_{1n}} & \frac{\partial \boldsymbol{Y}}{\partial x_{2n}} & \cdots & \frac{\partial \boldsymbol{Y}}{\partial x_{mn}} \end{pmatrix}$ Where each element known partial tangent matrix is $\frac{\partial \boldsymbol{Y}}{\partial x_{ij}} = \begin{pmatrix} \frac{\partial y_{11}}{\partial x_{ij}} & \frac{\partial y_{12}}{\partial x_{ij}} & \cdots & \frac{\partial y_{1q}}{\partial x_{ij}} \\ \frac{\partial y_{21}}{\partial x_{ij}} & \frac{\partial y_{22}}{\partial x_{ij}} & \cdots & \frac{\partial y_{2q}}{\partial x_{ij}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y_{p1}}{\partial x_{ij}} & \frac{\partial y_{p2}}{\partial x_{ij}} & \cdots & \frac{\partial y_{pq}}{\partial x_{ij}} \end{pmatrix}$

	$\frac{\partial \boldsymbol{Y}^{-1}}{\partial x} = -\boldsymbol{Y}^{-1}\frac{\partial \boldsymbol{Y}}{\partial x}\boldsymbol{Y}^{-1}$		

4. Composite derivative

Given $\boldsymbol{f}$ and $\boldsymbol{g}$ are vector functions (or matrix functions), there are questions "how to take derivative of composite function $\boldsymbol{g}\circ\boldsymbol{f} = \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}))$ and how to extend basic derivative-taking techniques such as chain rule, product rule and sum rule from scalar function to vector function". Let us sketch these rules before researching composite derivative in detailed.

Chain rule: Given the composite function $g\circ f = g(f(x))$ of two scalar functions f and g, the composite derivative is $(g\circ f)'(x) = g'\big(f(x)\big) = g'(y)f'(x) = \frac{\partial g\circ f}{\partial x} = \frac{\partial g}{\partial y}\frac{\partial y}{\partial x}$ where $y = f(x)$. Chain rule is applied to vector (or matrix) function in restriction. We make sense the compatibility between vector function f and g; concretely, the image space of f must be the same to the domain space of g and the product of two derivatives g' and f' must be defined. For example, given vector variables $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ x_p \end{pmatrix}$, $\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix}$ and $\boldsymbol{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{pmatrix}$ where $\boldsymbol{z}$ is function of $\boldsymbol{y}$ which in turn is function of $\boldsymbol{x}$.

By applying the chain rule, the derivative of $\boldsymbol{z}$ w.r.t $\boldsymbol{x}$ is:

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{pmatrix} \frac{\partial z_1}{\partial y_1} & \frac{\partial z_1}{\partial y_2} & \cdots & \frac{\partial z_1}{\partial y_n} \\ \frac{\partial z_2}{\partial y_1} & \frac{\partial z_2}{\partial y_2} & \cdots & \frac{\partial z_2}{\partial y_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial z_m}{\partial y_1} & \frac{\partial z_m}{\partial y_2} & \cdots & \frac{\partial z_m}{\partial y_n} \end{pmatrix} \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_p} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_p} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \frac{\partial y_n}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_p} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_1}{\partial x_2} & \cdots & \frac{\partial z_1}{\partial x_p} \\ \frac{\partial z_2}{\partial x_1} & \frac{\partial z_2}{\partial x_2} & \cdots & \frac{\partial z_2}{\partial x_p} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial z_m}{\partial x_1} & \frac{\partial z_m}{\partial x_2} & \cdots & \frac{\partial z_m}{\partial x_p} \end{pmatrix} \text{ where } \frac{\partial z_i}{\partial x_j} = \sum_{k=1}^{n} \frac{\partial z_i}{\partial y_k}\frac{\partial y_k}{\partial x_j}$$

Product rule: Given the product fg of two scalar functions f and g, the derivative of this product is $\big(f(x)g(x)\big)' = f'(x)g(x) + f(x)g'(x) = \frac{\partial f}{\partial x}g(x) + f(x)\frac{\partial g}{\partial x}$. The product rule can be applied to any vector and matrix function but it makes sense to define what *product* is and how the order of product is because there are many kinds of product such as scalar product between two vectors, matrix multiplication. Moreover matrix multiplication is not commutative and so its order is very important. For example, given two n_x1 column vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, both of them are functions of scalar x; we have $\boldsymbol{u} = \begin{pmatrix} u_1(x) \\ u_2(x) \\ \vdots \\ u_n(x) \end{pmatrix}$ and $\boldsymbol{v} = \begin{pmatrix} v_1(x) \\ v_2(x) \\ \vdots \\ v_n(x) \end{pmatrix}$. By applying product rule, the derivative of product $\boldsymbol{u}^T\boldsymbol{v}$ resulting out a scalar is different from the one of product $\boldsymbol{u}\boldsymbol{v}^T$ resulting out a matrix.

$$\frac{\partial \boldsymbol{u}^T\boldsymbol{v}}{\partial x} = \boldsymbol{u}^T\frac{\partial \boldsymbol{v}}{\partial x} + \boldsymbol{v}^T\frac{\partial \boldsymbol{u}}{\partial x} = \big(u_1(x), u_2(x), \dots, u_n(x)\big)\begin{pmatrix}\frac{\partial v_1}{\partial x}\\ \frac{\partial v_2}{\partial x}\\ \vdots \\ \frac{\partial v_n}{\partial x}\end{pmatrix} + \big(v_1(x), v_2(x), \dots, v_n(x)\big)\begin{pmatrix}\frac{\partial u_1}{\partial x}\\ \frac{\partial u_2}{\partial x}\\ \vdots \\ \frac{\partial u_n}{\partial x}\end{pmatrix}$$

$$= \sum_{i=1}^{n}\left(u_i(x)\frac{\partial v_i}{\partial x} + v_i(x)\frac{\partial u_i}{\partial x}\right)$$

$$\frac{\partial \boldsymbol{u}\boldsymbol{v}^T}{\partial x} = \boldsymbol{u}\frac{\partial \boldsymbol{v}^T}{\partial x} + \boldsymbol{v}\frac{\partial \boldsymbol{u}^T}{\partial x} = \begin{pmatrix}u_1(x)\\ u_2(x)\\ \vdots \\ u_n(x)\end{pmatrix}\left(\frac{\partial v_1}{\partial x}, \frac{\partial v_2}{\partial x}, \dots, \frac{\partial v_n}{\partial x}\right) + \begin{pmatrix}v_1(x)\\ v_2(x)\\ \vdots \\ v_n(x)\end{pmatrix}\left(\frac{\partial u_1}{\partial x}, \frac{\partial u_2}{\partial x}, \dots, \frac{\partial u_n}{\partial x}\right)$$

$$= \begin{pmatrix}z_{11} & z_{12} & \dots & z_{1n}\\ z_{21} & z_{22} & \dots & z_{2n}\\ \vdots & \vdots & \vdots & \vdots \\ z_{n1} & z_{n2} & \dots & z_{nn}\end{pmatrix} \text{ where } z_{ij} = u_i(x)\frac{\partial v_j}{\partial x} + v_j(x)\frac{\partial u_i}{\partial x}$$

Sum rule: Given the sum $f+g$ of two scalar functions f and g, the derivative of this sum is $(f+g)'(x) = f'(x) + g'(x)$. Sum rule is applied to any vector and matrix function with condition that f and g belong to the same kind. For example, given two m_xn matrices:

$$\boldsymbol{U} = \begin{pmatrix}u_{11}(x) & u_{12}(x) & \dots & u_{1n}(x)\\ u_{21}(x) & u_{22}(x) & \dots & u_{2n}(x)\\ \vdots & \vdots & \vdots & \vdots \\ u_{m1}(x) & u_{m2}(x) & \dots & u_{mn}(x)\end{pmatrix} \text{ and } \boldsymbol{V} = \begin{pmatrix}v_{11}(x) & v_{12}(x) & \dots & v_{1n}(x)\\ v_{21}(x) & v_{22}(x) & \dots & v_{2n}(x)\\ \vdots & \vdots & \vdots & \vdots \\ v_{m1}(x) & v_{m2}(x) & \dots & v_{mn}(x)\end{pmatrix}$$

which are matrix functions of scalar x, the derivative of their sum is:

$$\frac{\partial(\boldsymbol{U}+\boldsymbol{V})}{\partial x} = \frac{\partial \boldsymbol{U}}{\partial x} + \frac{\partial \boldsymbol{V}}{\partial x} = \begin{pmatrix}\frac{\partial u_{11}}{\partial x} & \frac{\partial u_{12}}{\partial x} & \dots & \frac{\partial u_{1n}}{\partial x}\\ \frac{\partial u_{21}}{\partial x} & \frac{\partial u_{22}}{\partial x} & \dots & \frac{\partial u_{2n}}{\partial x}\\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial u_{m1}}{\partial x} & \frac{\partial u_{m2}}{\partial x} & \dots & \frac{\partial u_{mn}}{\partial x}\end{pmatrix} + \begin{pmatrix}\frac{\partial v_{11}}{\partial x} & \frac{\partial v_{12}}{\partial x} & \dots & \frac{\partial v_{1n}}{\partial x}\\ \frac{\partial v_{21}}{\partial x} & \frac{\partial v_{22}}{\partial x} & \dots & \frac{\partial v_{2n}}{\partial x}\\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial v_{m1}}{\partial x} & \frac{\partial v_{m2}}{\partial x} & \dots & \frac{\partial v_{mn}}{\partial x}\end{pmatrix}$$

$$= \begin{pmatrix}\frac{\partial u_{11}}{\partial x} + \frac{\partial v_{11}}{\partial x} & \frac{\partial u_{12}}{\partial x} + \frac{\partial v_{12}}{\partial x} & \dots & \frac{\partial u_{1n}}{\partial x} + \frac{\partial v_{1n}}{\partial x}\\ \frac{\partial u_{21}}{\partial x} + \frac{\partial v_{21}}{\partial x} & \frac{\partial u_{22}}{\partial x} + \frac{\partial v_{22}}{\partial x} & \dots & \frac{\partial u_{2n}}{\partial x} + \frac{\partial v_{2n}}{\partial x}\\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial u_{m1}}{\partial x} + \frac{\partial v_{m1}}{\partial x} & \frac{\partial u_{m2}}{\partial x} + \frac{\partial v_{m2}}{\partial x} & \dots & \frac{\partial u_{mn}}{\partial x} + \frac{\partial v_{mn}}{\partial x}\end{pmatrix}$$

By combination of these rules, it is easy for us to take derivative of any composite function if some conditions w.r.t vector (matrix) functions are satisfied. Please pay attention that numerator layout convention for derivative notation is followed. Moreover, given a function and one of its variables so-called x, the constant w.r.t variable x is defined as a scalar, vector or matrix whose evaluation is

not a function of x but such constant can be a function of other variables. There are three kinds of constant such as scalar constant, vector constant and matrix constant.

Next section will discusses composite derivative of vector and matrix function, hence, derivative of simple scalar function is ignored but its variants related to vector and matrix will be surveyed later.

Scalar-by-vector composite derivative

Following are composite derivatives of typical scalar-by-vector functions (Wikipedia, Matrix calculus, 2014). Note that independent variable $\boldsymbol{x}$ is always column vector and derivative of scalar-by-vector function results out gradient vector.

Condition	Scalar-by-vector derivative
a is a scalar constant	$\frac{\partial a}{\partial \boldsymbol{x}} = \boldsymbol{0}^T$
a is a scalar constant $u = u(\boldsymbol{x})$ is a scalar function of $\boldsymbol{x}$	$\frac{\partial au}{\partial \boldsymbol{x}} = a\frac{\partial u}{\partial \boldsymbol{x}}$
$u = u(\boldsymbol{x})$, $v = v(\boldsymbol{x})$ are scalar functions of $\boldsymbol{x}$	$\frac{\partial (u+v)}{\partial \boldsymbol{x}} = \frac{\partial u}{\partial \boldsymbol{x}} + \frac{\partial v}{\partial \boldsymbol{x}}$
$u = u(\boldsymbol{x})$, $v = v(\boldsymbol{x})$ are scalar functions of $\boldsymbol{x}$	$\frac{\partial uv}{\partial \boldsymbol{x}} = u\frac{\partial v}{\partial \boldsymbol{x}} + v\frac{\partial u}{\partial \boldsymbol{x}}$
$u = u(\boldsymbol{x})$ is a scalar function of $\boldsymbol{x}$ f is a scalar-by-scalar function	$\frac{\partial f(u)}{\partial \boldsymbol{x}} = \frac{\partial f}{\partial u}\frac{\partial u}{\partial \boldsymbol{x}}$
$u = u(\boldsymbol{x})$ is a scalar function of $\boldsymbol{x}$ f, g are scalar-by-scalar functions	$\frac{\partial g(f(u))}{\partial \boldsymbol{x}} = \frac{\partial g}{\partial f}\frac{\partial f}{\partial u}\frac{\partial u}{\partial \boldsymbol{x}}$
a is a vector constant	$\frac{\partial a^T \boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{x}^T a}{\partial \boldsymbol{x}} = a^T$
	$\frac{\partial \boldsymbol{x}^T \boldsymbol{x}}{\partial \boldsymbol{x}} = 2\boldsymbol{x}^T$
$\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$, $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x})$ are vector functions of $\boldsymbol{x}$	$\frac{\partial \boldsymbol{u}^T \boldsymbol{v}}{\partial \boldsymbol{x}} = \boldsymbol{u}^T\frac{\partial \boldsymbol{v}}{\partial \boldsymbol{x}} + \boldsymbol{v}^T\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}}$ Note that $\frac{\partial \boldsymbol{v}}{\partial \boldsymbol{x}}$ and $\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}}$ are derivatives of vector-by-scalar functions, which results out Jacobian matrix, which implies that $\boldsymbol{u}^T\frac{\partial \boldsymbol{v}}{\partial \boldsymbol{x}}$ and $\boldsymbol{v}^T\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}}$ are vectors.

$\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$ is a vector function of $\boldsymbol{x}$	$\frac{\partial a^T \boldsymbol{u}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{u}^T a}{\partial \boldsymbol{x}} = a^T \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}}$
a and b are vector constants	$\frac{\partial a^T \boldsymbol{x} \boldsymbol{x}^T b}{\partial \boldsymbol{x}} = \boldsymbol{x}^T (ab^T + ba^T)$
b is a vector constant. A is a matrix constant	$\frac{\partial b^T A \boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{x}^T A^T b}{\partial \boldsymbol{x}} = b^T A$
A is a square matrix constant	$\frac{\partial \boldsymbol{x}^T A \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{x}^T (A + A^T)$
A is a symmetric matrix constant	$\frac{\partial \boldsymbol{x}^T A \boldsymbol{x}}{\partial \boldsymbol{x}} = 2\boldsymbol{x}^T A$
A is a square matrix constant	$\frac{\partial^2 \boldsymbol{x}^T A \boldsymbol{x}}{\partial \boldsymbol{x}^2} = A + A^T$
A is a symmetric matrix constant	$\frac{\partial^2 \boldsymbol{x}^T A \boldsymbol{x}}{\partial \boldsymbol{x}^2} = 2A$
$\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$, $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x})$ are vector functions of $\boldsymbol{x}$ and A is matrix constant.	$\frac{\partial \boldsymbol{u}^T A \boldsymbol{v}}{\partial \boldsymbol{x}} = \boldsymbol{u}^T A \frac{\partial \boldsymbol{v}}{\partial \boldsymbol{x}} + \boldsymbol{v}^T A^T \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}}$
A, C and D are matrix constants b and e are vector constants	$\frac{\partial (A\boldsymbol{x} + b)^T C (D\boldsymbol{x} + e)}{\partial \boldsymbol{x}} =$ $(D\boldsymbol{x} + e)^T C^T A + (A\boldsymbol{x} + b)^T CD$
a and is are vector constants	$\frac{\partial \lvert \boldsymbol{x} - a \rvert}{\partial \boldsymbol{x}} = \frac{(\boldsymbol{x} - a)^T}{\lvert x - a \rvert}$ Note that $\lvert \boldsymbol{x} - a \rvert$ is the length or norm or module of vector $\boldsymbol{x} - a$

Vector-by-scalar composite derivative

Following are composite derivatives of typical vector-by-scalar functions (Wikipedia, Matrix calculus, 2014). Note that independent variable x is always scalar and derivative of scalar-by-vector function results out tangent vector.

Condition	Vector-by-scalar derivative
a is a vector constant	$\frac{\partial a}{\partial x} = \boldsymbol{0}$
a is a scalar constant $\boldsymbol{u} = \boldsymbol{u}(x)$ is a vector function of x	$\frac{\partial a\boldsymbol{u}}{\partial x} = a \frac{\partial \boldsymbol{u}}{\partial x}$
A is a matrix constant $\boldsymbol{u} = \boldsymbol{u}(x)$ is a vector function of x	$\frac{\partial A\boldsymbol{u}}{\partial x} = A \frac{\partial \boldsymbol{u}}{\partial x}$

$\boldsymbol{u} = \boldsymbol{u}(x)$ is a vector function of x	$\frac{\partial \boldsymbol{u}^T}{\partial x} = \left(\frac{\partial \boldsymbol{u}}{\partial x}\right)^T$
$\boldsymbol{u} = \boldsymbol{u}(x)$, $\boldsymbol{v} = \boldsymbol{v}(x)$ are vector functions of x	$\frac{\partial (\boldsymbol{u} + \boldsymbol{v})}{\partial x} = \frac{\partial \boldsymbol{u}}{\partial x} + \frac{\partial \boldsymbol{v}}{\partial x}$
$\boldsymbol{u} = \boldsymbol{u}(x)$, $\boldsymbol{v} = \boldsymbol{v}(x)$ are vector functions of x	$\frac{\partial (\boldsymbol{u} \times \boldsymbol{v})}{\partial x} = \boldsymbol{u} \times \frac{\partial \boldsymbol{v}}{\partial x} + \frac{\partial \boldsymbol{u}}{\partial x} \times \boldsymbol{v}$ Note that sign × represents cross product between two vectors, mentioned in previous section.
$\boldsymbol{u} = \boldsymbol{u}(x)$ is a vector function of x $\boldsymbol{f}$ is a vector-by-vector function	$\frac{\partial \boldsymbol{f}(\boldsymbol{u})}{\partial x} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}} \frac{\partial \boldsymbol{u}}{\partial x}$ Note that vector-by-vector derivative gives Jacobian matrix, which refers that the product $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}} \frac{\partial \boldsymbol{u}}{\partial x}$ results out a vector.
$\boldsymbol{u} = \boldsymbol{u}(x)$ is a scalar function of x $\boldsymbol{f}, \boldsymbol{g}$ are vector-by-vector functions	$\frac{\partial \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{u}))}{\partial x} = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{f}} \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}} \frac{\partial \boldsymbol{u}}{\partial x}$

Vector-by-vector composite derivative

Following are composite derivatives of typical vector-by-vector functions (Wikipedia, Matrix calculus, 2014). Note that and independent variable $\boldsymbol{x}$ is always column vector and derivative of vector-by-vector function results out Jacobian matrix.

Condition	Vector-by-vector derivative
a is a vector constant	$\frac{\partial a}{\partial \boldsymbol{x}} = \boldsymbol{0}$
	$\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{I}$
A is a matrix constant	$\frac{\partial A\boldsymbol{x}}{\partial \boldsymbol{x}} = A$
A is a matrix constant	$\frac{\partial \boldsymbol{x}^T A}{\partial \boldsymbol{x}} = A^T$
a is a scalar constant $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$ is a vector function of $\boldsymbol{x}$	$\frac{\partial a\boldsymbol{u}}{\partial \boldsymbol{x}} = a \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}}$
$a = a(\boldsymbol{x})$ is a scalar function of $\boldsymbol{x}$ $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$ is a vector function of $\boldsymbol{x}$	$\frac{\partial a\boldsymbol{u}}{\partial \boldsymbol{x}} = a \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} + \boldsymbol{u} \frac{\partial a}{\partial \boldsymbol{x}}$

	Note that scalar-by-scalar derivative $\frac{\partial a}{\partial \boldsymbol{x}}$ gives out gradient vector (row vector) and so the expression $a\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} + \boldsymbol{u}\frac{\partial a}{\partial \boldsymbol{x}}$ results out a matrix
A is matrix constant $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$ is a vector function of $\boldsymbol{x}$	$\frac{\partial A\boldsymbol{u}}{\partial \boldsymbol{x}} = A\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}}$
$\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$, $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x})$ are vector functions of $\boldsymbol{x}$	$\frac{\partial(\boldsymbol{u}+\boldsymbol{v})}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} + \frac{\partial \boldsymbol{v}}{\partial \boldsymbol{x}}$
$\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$ is a vector function of x $\boldsymbol{f}$ is a vector-by-vector function	$\frac{\partial \boldsymbol{f}(\boldsymbol{u})}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}}$ Note that vector-by-vector derivative gives Jacobian matrix, which refers that the matrix multiplication $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}}$ results out a matrix.
$\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$ is a vector function of $\boldsymbol{x}$ $\boldsymbol{f}$, $\boldsymbol{g}$ are vector-by-vector functions	$\frac{\partial \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{u}))}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{f}}\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}}$

Matrix function and differential

Before researching matrix derivative, we should have a preparation of concepts of matrix function and differential. Among scalar-by-matrix functions, determinant $|\boldsymbol{X}|$ and operator trace $tr(\boldsymbol{X})$ are the most important. Given square matrix $\boldsymbol{X}$, the derivative of determinant $|\boldsymbol{X}|$ is adjoint matrix $adj(\boldsymbol{X})$; please see section "Basic concepts" for adjoint matrix definition.

$$\frac{\partial|\boldsymbol{X}|}{\partial \boldsymbol{X}} = adj(\boldsymbol{X}) = |\boldsymbol{X}|\boldsymbol{X}^{-1}$$

Given square matrix, the derivative of operator trace $tr(\boldsymbol{X})$ is identity matrix $\boldsymbol{I}$.

$$\frac{\partial tr(\boldsymbol{X})}{\partial \boldsymbol{X}} = \boldsymbol{I}$$

Differential is independent concept but it is often defined via derivative. Given matrix variable $\boldsymbol{X}$, differential of $\boldsymbol{X}$ denoted $d\boldsymbol{X}$ is an infinitesimal according to pre-defined metric. If $\boldsymbol{X}$ is defined in real number field $\mathbb{R}$, then differential $d\boldsymbol{X}$ is

$$d\boldsymbol{X} = \begin{pmatrix} dx_{11} & dx_{12} & \dots & dx_{1n} \\ dx_{21} & dx_{22} & \dots & dx_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ dx_{m1} & dx_{m2} & \dots & dx_{mn} \end{pmatrix} \text{ where } dx_{ij} \text{ is differential of real number } x_{ij}$$

Given matrix function $\boldsymbol{Y} = \boldsymbol{F}(\boldsymbol{X})$, differential of $\boldsymbol{Y}$ denoted $d\boldsymbol{Y}$ is defined as multiplication of derivative of $\boldsymbol{Y}$ w.r.t $\boldsymbol{X}$ by differential of $\boldsymbol{X}$.

$$d\boldsymbol{Y} = d\boldsymbol{F} = \boldsymbol{F}'(\boldsymbol{X})d\boldsymbol{X} = \frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{X}} d\boldsymbol{X}$$

For convenience, we use derivative notation for denoting differential. It means that differentials of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are denoted $\partial\boldsymbol{X}$ and $\partial\boldsymbol{Y}$, respectively instead of $d\boldsymbol{X}$ and $d\boldsymbol{Y}$. Note that vector is reduced form of matrix and so, differential of vector is defined by the same way. For example, given vector-by-vector function $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) = A\boldsymbol{x}$ where A is matrix constant, derivative of vector $\boldsymbol{y}$ w.r.t vector $\boldsymbol{x}$ is:

$$\frac{\partial A\boldsymbol{x}}{\partial \boldsymbol{x}} = A$$

The differential of $\boldsymbol{y}$ is:

$$\partial \boldsymbol{y} = A\partial \boldsymbol{x}$$

Therefore, the concept of differential totally coincides with the concept of derivative. The semantic of differential is deviation and differential is used to approximate a function or dependent variable. The deviation of $\boldsymbol{y}$ denoted $\partial\boldsymbol{y}$ is approximated by the deviation of $\boldsymbol{x}$ denoted $\partial\boldsymbol{x}$ and so the approximation expression is $\partial \boldsymbol{y} = A\partial \boldsymbol{x}$ in which A is result of derivative of $\boldsymbol{y}$. Two expressions $\frac{\partial A\boldsymbol{x}}{\partial \boldsymbol{x}} = A$ and $\partial \boldsymbol{y} = A\partial \boldsymbol{x}$ are the same. As a result, all aforementioned derivatives can be written in differential form. If you want to convert a complicated differential to a derivative and vice versa, please find out the canonical form of such differential and look up the respective derivative of canonical form in following table (Wikipedia, Matrix calculus, 2014).

Canonical differential form	Equivalent derivative form	Note
$\partial y = a\partial x$	$\frac{\partial y}{\partial x} = a$	y is dependent scalar variable. x is independent scalar variable. a is scalar.
$\partial y = \boldsymbol{a}\partial \boldsymbol{x}$	$\frac{\partial y}{\partial \boldsymbol{x}} = \boldsymbol{a}$	y is dependent scalar variable. $\boldsymbol{x}$ is independent vector variable. $\boldsymbol{a}^T$ is gradient (row) vector.
$\partial y = tr(\boldsymbol{A}\partial \boldsymbol{X})$	$\frac{\partial y}{\partial \boldsymbol{X}} = \boldsymbol{A}$	y is dependent scalar variable. $\boldsymbol{X}$ is independent matrix variable. $\boldsymbol{A}$ is gradient matrix.
$\partial \boldsymbol{y} = \boldsymbol{a}\partial x$	$\frac{\partial \boldsymbol{y}}{\partial x} = \boldsymbol{a}$	$\boldsymbol{y}$ is dependent vector variable. x is independent scalar variable. $\boldsymbol{a}$ is tangent (column) vector.
$\partial \boldsymbol{y} = \boldsymbol{A}\partial \boldsymbol{x}$	$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \boldsymbol{A}$	$\boldsymbol{y}$ is dependent vector variable. $\boldsymbol{x}$ is independent vector variable. $\boldsymbol{A}$ is Jacobian matrix.
$\partial \boldsymbol{Y} = \boldsymbol{A}\partial x$	$\frac{\partial \boldsymbol{Y}}{\partial x} = \boldsymbol{A}$	$\boldsymbol{Y}$ is dependent matrix variable. x is independent scalar variable. $\boldsymbol{A}$ is tangent matrix.

All of principles and techniques in matrix derivative such as chain rule, product rule, sum rule, and etc are applied in matrix differential. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be dependent square matrix variables, following are some basic properties of differential inferred directly from derivatives (Petersen & Pedersen, 2012, p. 8).

$\partial A = 0$ where A is constant
$\partial \boldsymbol{X}^T = (\partial \boldsymbol{X})^T$
$\partial a\boldsymbol{X} = a\partial \boldsymbol{X}$ where a is constant
$\partial tr(\boldsymbol{X}) = tr(\partial \boldsymbol{X})$
$\partial(\boldsymbol{X} + \boldsymbol{Y}) = \partial \boldsymbol{X} + \partial \boldsymbol{Y}$
$\partial \boldsymbol{XY} = (\partial \boldsymbol{X})\boldsymbol{Y} + \boldsymbol{X}(\partial \boldsymbol{Y})$
$\partial \boldsymbol{X}^{-1} = -\boldsymbol{X}^{-1}\partial \boldsymbol{X}\boldsymbol{X}^{-1}$
$\partial|\boldsymbol{X}| = tr(adj(\boldsymbol{X})\partial \boldsymbol{X}) = |\boldsymbol{X}|tr(\boldsymbol{X}^{-1}\partial \boldsymbol{X})$
where $adj(\boldsymbol{X})$ is adjoint matrix of $\boldsymbol{X}$
$\partial(ln|\boldsymbol{X}|) = tr(\boldsymbol{X}^{-1}\partial \boldsymbol{X})$

We have an example of trace differential for illustrating the mutual relationship of differential and derivative. Let $\boldsymbol{Y} = A\boldsymbol{X}B\boldsymbol{X}^TC$ be matrix function by matrix variable $\boldsymbol{X}$, all of them are square matrices. The differential of $\boldsymbol{Y}$ (Wikipedia, Matrix calculus, 2014) is:

$$\begin{aligned}
\partial tr(A\boldsymbol{X}B\boldsymbol{X}^TC) &= \partial tr(CA\boldsymbol{X}B\boldsymbol{X}^T) = tr\big(\partial(CA\boldsymbol{X}B\boldsymbol{X}^T)\big) = tr\big(\partial(CA\boldsymbol{X})B\boldsymbol{X}^T + CA\boldsymbol{X}\partial(B\boldsymbol{X}^T)\big) \\
&= tr(CA\partial(\boldsymbol{X})B\boldsymbol{X}^T) + tr(CA\boldsymbol{X}B\partial\boldsymbol{X}^T) = tr\big(B\boldsymbol{X}^TCA\partial(\boldsymbol{X})\big) + tr(CA\boldsymbol{X}B(\partial\boldsymbol{X})^T) \\
&= tr\big(B\boldsymbol{X}^TCA\partial(\boldsymbol{X})\big) + tr((CA\boldsymbol{X}B(\partial\boldsymbol{X})^T)^T) \\
&= tr\big(B\boldsymbol{X}^TCA\partial(\boldsymbol{X})\big) + tr((CA\boldsymbol{X}B(\partial\boldsymbol{X})^T)^T) \\
&= tr\big(B\boldsymbol{X}^TCA\partial(\boldsymbol{X})\big) + tr(\partial\boldsymbol{X}B^T\boldsymbol{X}^TA^TC^T) = tr\big(B\boldsymbol{X}^TCA\partial(\boldsymbol{X})\big) + tr(B^T\boldsymbol{X}^TA^TC^T\partial\boldsymbol{X}) \\
&= tr\big((B\boldsymbol{X}^TCA + B^T\boldsymbol{X}^TA^TC^T)\partial\boldsymbol{X}\big)
\end{aligned}$$

$$\Rightarrow \frac{\partial tr(A\boldsymbol{X}B\boldsymbol{X}^TC)}{\partial \boldsymbol{X}} = B\boldsymbol{X}^TCA + B^T\boldsymbol{X}^TA^TC^T$$

We will survey matrix-by-matrix functions. Aforementioned in section "Matrix analysis", given non-singular matrix $\boldsymbol{A}(n_x n)$, its eigenvalues are solutions of following equation:

$|\boldsymbol{A} - \lambda\boldsymbol{I}_n| = 0$ where $\boldsymbol{I}_n$ is $n_x n$ identity matrix

Determinant $|\boldsymbol{A} - \lambda\boldsymbol{I}_n|$ is expanded as n^{th} order polynomial called *characteristic polynomial* in which λ is scalar variable, we re-write this polynomial.

$P(\lambda) = |\boldsymbol{A} - \lambda\boldsymbol{I}_n|$

Cayley–Hamilton theorem (Wikipedia, Cayley–Hamilton theorem, 2014) states that if $P(\lambda)$ has solution(s) then, when substituting λ by matrix $\boldsymbol{A}$, we get a zero matrix **0.**

$\boldsymbol{P}(\boldsymbol{A}) = \boldsymbol{0}$

For example let $\boldsymbol{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, the characteristic polynomial is expressed as below:

$$P(\lambda) = |A - \lambda I_2| = \begin{vmatrix} 1-\lambda & 2 \\ 3 & 4-\lambda \end{vmatrix} = \lambda^2 - 5\lambda - 2$$

Because $P(\lambda) = 0$ has solutions $\lambda = \frac{5 \pm \sqrt{33}}{2}$, we have:

$$\boldsymbol{P(A)} = \boldsymbol{A^2 - 5A - 2I_2} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} - 5\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} - 2\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

Given any square matrix constant C, matrix polynomial $2C = \boldsymbol{P(A)} = \boldsymbol{A^2 - 5A - 2C}$ and its derivative always exist.

$$\frac{\partial \boldsymbol{P}}{\partial \boldsymbol{A}} = 2\boldsymbol{A} - 5\boldsymbol{I}_2$$

In generalization, given matrix variable $\boldsymbol{X}$, the **matrix polynomial** known as matrix-by-matrix function always exists together its derivative if $\boldsymbol{X}$ is non-singular square matrix. Note that matrix operations within matrix polynomial include scalar multiplication, addition, subtraction, multiplication and power.

$$\boldsymbol{P(X)} = a_0\boldsymbol{I} + a_1\boldsymbol{X} + a_2\boldsymbol{X}^2 + \cdots + a_n\boldsymbol{X}^n = \sum_{k=0}^{n} a_k\boldsymbol{X}^k$$

$$\boldsymbol{P'(X)} = \frac{\partial \boldsymbol{P}}{\partial \boldsymbol{X}} = a_1\boldsymbol{I} + 2a_2\boldsymbol{X} + 3a_3\boldsymbol{X}^2 + \cdots + na_n\boldsymbol{X}^{n-1} = \sum_{k=1}^{n} ka_k\boldsymbol{X}^{k-1}$$

Where a_i (s) are scalar and $\boldsymbol{I}$ is identity matrix.

Given real scalar function f has Taylor expansion or Taylor series at entry 0.

$$f(x) = f(0) + \frac{f'(0)x}{1!} + \frac{f''(0)x^2}{2!} + \cdots + \frac{f^{(n)}(0)x^n}{n!} + \cdots = \sum_{k=0}^{\infty} \frac{f^{(k)}(x)x^k}{k!}$$

$$\text{where } f^{(k)} \text{denotes } k^{th} \text{ derivative of } f \text{ and } f^{(0)} = f$$

When substituting x by square matrix $\boldsymbol{X}$, we get an matrix-by-matrix function called **matrix-power-series function** in which arithmetic operations are extended as matrix operations such as scalar multiplication, addition, subtraction, multiplication and power. Following is Taylor series of matrix-power-series function $\boldsymbol{S}$ at matrix entry $\boldsymbol{0}$.

$$\boldsymbol{S(X)} = \boldsymbol{S(0)} + \frac{\boldsymbol{S'(0)X}}{1!} + \frac{\boldsymbol{S''(0)X}^2}{2!} + \cdots + \frac{\boldsymbol{S}^{(n)}\boldsymbol{(0)X}^n}{n!} + \cdots = \sum_{k=0}^{\infty} \frac{\boldsymbol{S}^{(k)}\boldsymbol{(0)X}^k}{k!}$$

$$\text{where } \boldsymbol{S}^{(k)} \text{denotes } k^{th} \text{ derivative of } \boldsymbol{S} \text{ and } \boldsymbol{S}^{(0)} = \boldsymbol{S}$$

There are some popular matrix-power-series functions such as matrix exponential $\boldsymbol{E^X}$, natural logarithm of matrix ***LNX***, ***SINX***, ***COSX***. Because all matrix-power-series functions $\boldsymbol{S(X)}$ are

defined by Taylor series, it is very complicated to evaluate them and so there is a practical method to calculate them. If $\boldsymbol{X} = \begin{pmatrix} x_{11} & 0 & \dots & 0 \\ 0 & x_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & x_{nn} \end{pmatrix}$ is diagonal matrix then

$$\boldsymbol{S}(\boldsymbol{X}) = \begin{pmatrix} s(x_{11}) & 0 & \dots & 0 \\ 0 & s(x_{22}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s(x_{nn}) \end{pmatrix}$$

Where $\boldsymbol{S}(\boldsymbol{X})$ represents matrix-power-series functions such as matrix exponential $\boldsymbol{E}^{\boldsymbol{X}}$, natural logarithm of matrix ***LNX***, ***SINX***, ***COSX***; *s* represents scalar functions such as *e*, *ln*, *sin*, *cos*. If diagonalizable matrix $\boldsymbol{X}$ is decomposed into spectrums $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{-1}$ then

$$\boldsymbol{S}(\boldsymbol{X}) = \boldsymbol{U}\boldsymbol{S}(\boldsymbol{\Lambda})\boldsymbol{U}^{-1}$$

The formula above is very important for calculating matrix-power-series function; which implies it is possible to compute practically any diagonalizable matrix instead of using complicated Taylor series. For example, given 2_x2 matrix $\boldsymbol{X} = \begin{pmatrix} 3/2 & 1/2 \\ 1/2 & 3/2 \end{pmatrix}$ is diagonalized as follows:

$$\boldsymbol{X} = \begin{pmatrix} 3/2 & 1/2 \\ 1/2 & 3/2 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$

Matrix-power-series functions $\boldsymbol{E}^{\boldsymbol{X}}$, ***LNX***, ***SINX*** and ***COSX*** are totally evaluated.

$$\begin{aligned} \boldsymbol{E}^{\boldsymbol{X}} &= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \boldsymbol{E}^{\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \\ &= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} e^1 & 0 \\ 0 & e^2 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} = \frac{e}{2} \begin{pmatrix} e+1 & e-1 \\ e-1 & e+1 \end{pmatrix} \end{aligned}$$

$$\begin{aligned} \boldsymbol{LNX} &= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \boldsymbol{LN}^{\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \\ &= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} ln1 & 0 \\ 0 & ln2 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} = \frac{ln2}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \end{aligned}$$

$$\begin{aligned} \boldsymbol{SINX} &= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \boldsymbol{SIN}^{\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \\ &= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} sin1 & 0 \\ 0 & sin2 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} sin2 + sin1 & sin2 - sin1 \\ sin2 - sin1 & sin2 + sin1 \end{pmatrix} \end{aligned}$$

$$\begin{aligned} \boldsymbol{COSX} &= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \boldsymbol{COS}^{\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \\ &= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} cos1 & 0 \\ 0 & cos2 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} cos2 + cos1 & cos2 - cos1 \\ cos2 - cos1 & cos2 + cos1 \end{pmatrix} \end{aligned}$$

Now we research popular matrix-power-series functions such as matrix exponential $\boldsymbol{E^X}$, natural logarithm of matrix ***LNX***, ***SINX*** and ***COSX***. The exponential of square matrix $\boldsymbol{X}$ denoted $\boldsymbol{E^X}$ or ***EXP***($\boldsymbol{X}$) is invertible matrix which defined as Taylor series (Wikipedia, Matrix exponential, 2014):

$$\boldsymbol{E^X} = \boldsymbol{I} + \frac{\boldsymbol{X}}{1!} + \frac{\boldsymbol{X}^2}{2!} + \cdots + \frac{\boldsymbol{X}^n}{n!} + \cdots = \sum_{k=0}^{\infty} \frac{\boldsymbol{X}^k}{k!}$$

and its derivative is

$$(\boldsymbol{E^X})' = \frac{\partial \boldsymbol{E^X}}{\partial \boldsymbol{X}} = \boldsymbol{E^X}$$

Given $x \in \mathbb{R}$ is real scalar and A is matrix constant, we have

$$\frac{\partial \boldsymbol{E}^{xA}}{\partial x} = A\boldsymbol{E}^{tA} = \boldsymbol{E}^{tA}A$$

Given square matrix $\boldsymbol{U}(x) = \begin{pmatrix} u_{11}(x) & u_{12}(x) & \dots & u_{1n}(x) \\ u_{21}(x) & u_{22}(x) & \dots & u_{2n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ u_{m1}(x) & u_{m2}(x) & \dots & u_{mn}(x) \end{pmatrix}$ is matrix-by-scalar function, the derivative of $\boldsymbol{E}^{\boldsymbol{U}(x)}$ w.r.t scalar variable x is (Wikipedia, Matrix exponential, 2014):

$$\frac{\partial \boldsymbol{E}^{\boldsymbol{U}(x)}}{\partial x} = \int_0^1 \boldsymbol{E}^{t\boldsymbol{U}(x)} \frac{\partial \boldsymbol{U}}{\partial x} \boldsymbol{E}^{(1-t)\boldsymbol{U}(x)} dt$$

Following are other properties of $\boldsymbol{E^X}$ (Wikipedia, Matrix exponential, 2014):

$\boldsymbol{E^X}$ is always invertible matrix
$\boldsymbol{E^0} = \boldsymbol{I}$
$\boldsymbol{E}^{aX}\boldsymbol{E}^{bX} = \boldsymbol{E}^{(a+b)X}$
$\boldsymbol{E^X E^{-X}} = \boldsymbol{I}$
If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are mutually commutative $\boldsymbol{XY} = \boldsymbol{YX}$ then $\boldsymbol{E^X E^Y} = \boldsymbol{E^Y E^X} = \boldsymbol{E}^{(X+Y)}$
If $\boldsymbol{Y}$ is invertible then $\boldsymbol{E}^{\boldsymbol{YXY}^{-1}} = \boldsymbol{YE^X Y}^{-1}$
$\boldsymbol{EXP}(\boldsymbol{X}^T) = (\boldsymbol{EXP}(\boldsymbol{X}))^T$
$|\boldsymbol{E}^{\boldsymbol{X}}| = \boldsymbol{E}^{tr(\boldsymbol{X})}$
If $\boldsymbol{X}$ is idempotent matrix ($\boldsymbol{XX} = \boldsymbol{X}$) then $\boldsymbol{E^X} = \boldsymbol{I} + (e-1)\boldsymbol{X}$

The natural logarithm (Wikipedia, Logarithm of a matrix, 2014) of invertible matrix $\boldsymbol{X}$ denoted ***LNX*** is defined as the matrix $\boldsymbol{Y}$ such that $\boldsymbol{Y} = \boldsymbol{E^X}$. Note that ***LNX*** exists if and only if $\boldsymbol{X}$ is invertible matrix. ***LNX*** is defined as Taylor series:

$$\boldsymbol{LN}(\boldsymbol{X}) = \boldsymbol{LN}(\boldsymbol{I} + \boldsymbol{A}) = \boldsymbol{A} - \frac{\boldsymbol{A}^2}{2} + \frac{\boldsymbol{A}^3}{3} - \frac{\boldsymbol{A}^4}{4} + \cdots = \sum_{k=0}^{\infty} \frac{(-1)^k \boldsymbol{A}^{k+1}}{k+1}$$

Following are properties of ***LNX*** (Wikipedia, Logarithm of a matrix, 2014):

$\boldsymbol{X}^{-1} = \boldsymbol{e}^{-\boldsymbol{ln}(\boldsymbol{X})}$

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are both positive-definite and commutative matrices, then

$$\boldsymbol{XY} = \boldsymbol{E}^{\boldsymbol{LN}\,\boldsymbol{X}+\boldsymbol{LN}\,\boldsymbol{Y}}$$

Given square matrix $\boldsymbol{X}$, sine and cosine of $\boldsymbol{X}$ is defined by Taylor series.

$$\boldsymbol{SINX} = \boldsymbol{X} - \frac{\boldsymbol{X}^3}{3!} + \frac{\boldsymbol{X}^5}{5!} - \frac{\boldsymbol{X}^7}{7!} + \cdots + \frac{(-1)^n \boldsymbol{X}^{2n+1}}{(2n+1)!} + \cdots = \sum_{k=0}^{\infty} \frac{(-1)^k \boldsymbol{X}^{2k+1}}{(2k+1)!}$$

$$\boldsymbol{COSX} = \boldsymbol{I} - \frac{\boldsymbol{X}^2}{2!} + \frac{\boldsymbol{X}^4}{4!} - \frac{\boldsymbol{X}^6}{6!} + \cdots + \frac{(-1)^n \boldsymbol{X}^{2n}}{(2n)!} + \cdots = \sum_{k=0}^{\infty} \frac{(-1)^k \boldsymbol{X}^{2k}}{(2k)!}$$

The derivatives of sine and cosine of $\boldsymbol{X}$ are:

$$\boldsymbol{SIN'X} = \frac{\partial \boldsymbol{SINX}}{\partial \boldsymbol{X}} = \boldsymbol{COSX}$$

$$\boldsymbol{COS'X} = \frac{\partial \boldsymbol{COSX}}{\partial \boldsymbol{X}} = -\boldsymbol{SINX}$$

Following table is sketchy summarization of matrix functions, suppose that independent variable $\boldsymbol{X}$ is square matrix.

Dependent variable	Independent variable		
	Scalar	**Vector**	**Matrix**
Scalar			Determinant $\lvert\boldsymbol{X}\rvert$ with $\boldsymbol{X}$ is square matrix $\frac{\partial \lvert\boldsymbol{X}\rvert}{\partial \boldsymbol{X}} = adj(\boldsymbol{X}) = \lvert\boldsymbol{X}\rvert \boldsymbol{X}^{-1}$
			Trace operator $tr(\boldsymbol{X})$ $\frac{\partial tr(\boldsymbol{X})}{\partial \boldsymbol{X}} = \boldsymbol{I}$ $\partial tr(\boldsymbol{X}) = tr(\partial \boldsymbol{X})$ $\partial \lvert\boldsymbol{X}\rvert = tr(adj(\boldsymbol{X})\partial \boldsymbol{X}) = \lvert\boldsymbol{X}\rvert tr(\boldsymbol{X}^{-1}\partial \boldsymbol{X})$
Vector			
Matrix			Matrix polynomial $\boldsymbol{P}(\boldsymbol{X}) = \sum_{k=0}^{n} a_k \boldsymbol{X}^k$ $\boldsymbol{P}'(\boldsymbol{X}) = \frac{\partial \boldsymbol{P}}{\partial \boldsymbol{X}} = \sum_{k=1}^{n} k a_k \boldsymbol{X}^{k-1}$
			Matrix-power-series function such as $\boldsymbol{E^X}$, $\boldsymbol{LNX}$, $\boldsymbol{SINX}$ and $\boldsymbol{COSX}$ $\boldsymbol{E^X} = \sum_{k=0}^{\infty} \frac{\boldsymbol{X}^k}{k!}$ $(\boldsymbol{E^X})' = \frac{\partial \boldsymbol{E^X}}{\partial \boldsymbol{X}} = \boldsymbol{E^X}$ $\boldsymbol{SINX} = \sum_{k=0}^{\infty} \frac{(-1)^k \boldsymbol{X}^{2k+1}}{(2k+1)!}$ $\boldsymbol{COSX} = \sum_{k=0}^{\infty} \frac{(-1)^k \boldsymbol{X}^{2k}}{(2k)!}$ $\boldsymbol{SIN'X} = \boldsymbol{COSX}$

			$\boldsymbol{COS'X = -SINX}$

Other composite derivatives relevant to matrix will be described below.

Scalar-by-matrix composite derivative

Determinant and trace operators are important to scalar-by-matrix functions and hence, scalar-by-matrix derivatives often relate to determinant and trace operators. Following are composite derivatives of typical scalar-by-matrix functions. Note that independent variable $\boldsymbol{X}$ is always matrix and derivative of scalar-by-matrix function results out gradient matrix.

Condition	Scalar-by-matrix derivative
a is a scalar constant	$\frac{\partial a}{\partial \boldsymbol{X}} = \boldsymbol{0}^T$
a is a scalar constant $u = u(\boldsymbol{X})$ is a scalar function of $\boldsymbol{X}$	$\frac{\partial au}{\partial \boldsymbol{X}} = a\frac{\partial u}{\partial \boldsymbol{X}}$
$u = u(\boldsymbol{X})$, $v = v(\boldsymbol{X})$ are scalar functions of $\boldsymbol{X}$	$\frac{\partial (u+v)}{\partial \boldsymbol{X}} = \frac{\partial u}{\partial \boldsymbol{X}} + \frac{\partial v}{\partial \boldsymbol{X}}$
$u = u(\boldsymbol{X})$, $v = v(\boldsymbol{X})$ are scalar functions of $\boldsymbol{X}$	$\frac{\partial uv}{\partial \boldsymbol{X}} = u\frac{\partial v}{\partial \boldsymbol{X}} + v\frac{\partial u}{\partial \boldsymbol{X}}$
$u = u(\boldsymbol{X})$ is a scalar function of $\boldsymbol{X}$ f is a scalar-by-scalar function	$\frac{\partial f(u)}{\partial \boldsymbol{X}} = \frac{\partial f}{\partial u}\frac{\partial u}{\partial \boldsymbol{X}}$
$u = u(\boldsymbol{X})$ is a scalar function of $\boldsymbol{X}$ f, g are scalar-by-scalar functions	$\frac{\partial g(f(u))}{\partial \boldsymbol{X}} = \frac{\partial g}{\partial f}\frac{\partial f}{\partial u}\frac{\partial u}{\partial \boldsymbol{X}}$
$\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{X})$ is a matrix function of $\boldsymbol{X}$ f is scalar-by-matrix function	$\frac{\partial f(\boldsymbol{U})}{\partial x_{ij}} = tr\left(\frac{\partial f(\boldsymbol{U})}{\partial \boldsymbol{U}}\frac{\partial \boldsymbol{U}}{\partial x_{ij}}\right)$ Note that $\frac{\partial f(\boldsymbol{U})}{\partial x_{ij}}$ is derivative of scalar-by-scalar function, which results out a scalar
Trace related derivative	
	$\frac{\partial tr(\boldsymbol{X})}{\partial \boldsymbol{X}} = \boldsymbol{I}$ Note that the derivative of trace operator is mentioned in previous section
$\boldsymbol{G}(\boldsymbol{X})$ is any matrix polynomial $\boldsymbol{P}(\boldsymbol{X})$ or matrix-power-series function such as $\boldsymbol{E}^{\boldsymbol{X}}$,	$\frac{\partial tr(\boldsymbol{G}(\boldsymbol{X}))}{\partial \boldsymbol{X}} = \boldsymbol{G}'(\boldsymbol{X})$

LNX, ***SINX***, ***COSX*** and etc	Following are some popular derivatives $\frac{\partial tr(\boldsymbol{P}(\boldsymbol{X}))}{\partial \boldsymbol{X}} = \frac{\partial tr(\sum_{k=0}^{n} a_k \boldsymbol{X}^k)}{\partial \boldsymbol{X}} = \boldsymbol{P}'(\boldsymbol{X}) = \sum_{k=1}^{n} k a_k \boldsymbol{X}^{k-1}$ $\frac{\partial tr(\boldsymbol{X}^n)}{\partial \boldsymbol{X}} = n\boldsymbol{X}^{n-1}$ where n is positive integer $\frac{\partial tr(\boldsymbol{E}^{\boldsymbol{X}})}{\partial \boldsymbol{X}} = \boldsymbol{E}^{\boldsymbol{X}}$ $\frac{\partial tr(\boldsymbol{SINX})}{\partial \boldsymbol{X}} = \boldsymbol{COSX}$ $\frac{\partial tr(\boldsymbol{COSX})}{\partial \boldsymbol{X}} = -\boldsymbol{SINX}$
$\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{X})$ and $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{X})$ are matrix functions of $\boldsymbol{X}$	$\frac{\partial tr(\boldsymbol{U}+\boldsymbol{V})}{\partial \boldsymbol{X}} = \frac{\partial tr(\boldsymbol{U})}{\partial \boldsymbol{X}} + \frac{\partial tr(\boldsymbol{V})}{\partial \boldsymbol{X}}$
a is a scalar constant $\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{X})$ is a matrix function of $\boldsymbol{X}$	$\frac{\partial tr(a\boldsymbol{U})}{\partial \boldsymbol{X}} = a\frac{\partial tr(\boldsymbol{U})}{\partial \boldsymbol{X}}$
A, B are matrix constants	$\frac{\partial tr(A\boldsymbol{X})}{\partial \boldsymbol{X}} = \frac{\partial tr(\boldsymbol{X}A)}{\partial \boldsymbol{X}} = A$ $\frac{\partial tr(A\boldsymbol{X}B)}{\partial \boldsymbol{X}} = \frac{\partial tr(BA\boldsymbol{X})}{\partial \boldsymbol{X}} = BA$ $\frac{\partial tr(A\boldsymbol{X}^T)}{\partial \boldsymbol{X}} = \frac{\partial tr(\boldsymbol{X}^T A)}{\partial \boldsymbol{X}} = A^T$ $\frac{\partial tr(A\boldsymbol{X}^T B)}{\partial \boldsymbol{X}} = \frac{\partial tr(BA\boldsymbol{X}^T)}{\partial \boldsymbol{X}} = A^T B^T$
A is a matrix constant	$\frac{\partial tr(\boldsymbol{X}^T A\boldsymbol{X})}{\partial \boldsymbol{X}} = \frac{\partial tr(\boldsymbol{X}A\boldsymbol{X}^T)}{\partial \boldsymbol{X}} = \boldsymbol{X}^T(A + A^T)$
A, B are matrix constants	$\frac{\partial tr(\boldsymbol{X}^{-1}B)}{\partial \boldsymbol{X}} = \frac{\partial tr(B\boldsymbol{X}^{-1})}{\partial \boldsymbol{X}} = -(\boldsymbol{X}^{-1})^T B(\boldsymbol{X}^{-1})^T$ $\frac{\partial tr(A\boldsymbol{X}^{-1}B)}{\partial \boldsymbol{X}} = \frac{\partial tr(BA\boldsymbol{X}^{-1})}{\partial \boldsymbol{X}} = -(\boldsymbol{X}^{-1})^T BA(\boldsymbol{X}^{-1})^T$
A, B, C are matrix constants	$\frac{\partial tr(A\boldsymbol{X}B\boldsymbol{X}^T C)}{\partial \boldsymbol{X}} = B\boldsymbol{X}^T CA + B^T\boldsymbol{X}^T A^T C^T$

A is matrix constant	$\frac{\partial tr(A\boldsymbol{X}^n)}{\partial \boldsymbol{X}} = \sum_{k=0}^{n-1} \boldsymbol{X}^k A \boldsymbol{X}^{n-k-1}$
Determinant related derivative	
	$\frac{\partial \lvert\boldsymbol{X}\rvert}{\partial \boldsymbol{X}} = adj(\boldsymbol{X}) = \lvert\boldsymbol{X}\rvert \boldsymbol{X}^{-1}$
a is scalar constant $\boldsymbol{X}^-$ is G-inverse of $\boldsymbol{X}$, please see section "Basic concepts"	$\frac{\partial ln\lvert a\boldsymbol{X}\rvert}{\partial \boldsymbol{X}} = \boldsymbol{X}^{-1}$ $\frac{\partial ln\lvert \boldsymbol{X}^T\boldsymbol{X}\rvert}{\partial \boldsymbol{X}} = 2\boldsymbol{X}^-$ $\frac{\partial ln\lvert \boldsymbol{X}^T\boldsymbol{X}\rvert}{\partial \boldsymbol{X}^-} = -2\boldsymbol{X}$
A, B are matrix constants	$\frac{\partial \lvert A\boldsymbol{X}B\rvert}{\partial \boldsymbol{X}} = \lvert A\boldsymbol{X}B\rvert \boldsymbol{X}^{-1}$
n is positive integer	$\frac{\partial \lvert\boldsymbol{X}^n\rvert}{\partial \boldsymbol{X}} = n\lvert\boldsymbol{X}^n\rvert \boldsymbol{X}^{-1}$
A is a matrix constant $\boldsymbol{X}$ is invertible	$\frac{\partial \lvert\boldsymbol{X}^T A\boldsymbol{X}\rvert}{\partial \boldsymbol{X}} = 2\lvert\boldsymbol{X}^T A\boldsymbol{X}\rvert \boldsymbol{X}^{-1}$
A is a matrix constant and symmetric $\boldsymbol{X}$ is not square matrix	$\frac{\partial \lvert\boldsymbol{X}^T A\boldsymbol{X}\rvert}{\partial \boldsymbol{X}} = 2\lvert\boldsymbol{X}^T A\boldsymbol{X}\rvert (\boldsymbol{X}^T A^T \boldsymbol{X})^{-1} \boldsymbol{X}^T A^T$
A is a matrix constant and non-symmetric $\boldsymbol{X}$ is not square matrix	$\frac{\partial \lvert\boldsymbol{X}^T A\boldsymbol{X}\rvert}{\partial \boldsymbol{X}} = 2\lvert\boldsymbol{X}^T A\boldsymbol{X}\rvert ((\boldsymbol{X}^T A\boldsymbol{X})^{-1} \boldsymbol{X}^T A + (\boldsymbol{X}^T A^T \boldsymbol{X})^{-1} \boldsymbol{X}^T A^T)$

Matrix-by-scalar composite derivative

Following are composite derivatives of typical scalar-by-matrix functions. Note that independent variable x is always scalar and derivative of matrix-by-scalar function results out tangent matrix.

Condition	Matrix-by-scalar derivative
A is a matrix constant	$\frac{\partial A}{\partial x} = \boldsymbol{0}$
a is a scalar constant $\boldsymbol{U} = \boldsymbol{U}(x)$ is a matrix function of x	$\frac{\partial a\boldsymbol{U}}{\partial x} = a\frac{\partial \boldsymbol{U}}{\partial x}$

A, B are a matrix constants $\boldsymbol{U} = \boldsymbol{U}(x)$ is a matrix function of x	$\frac{\partial A\boldsymbol{U}B}{\partial x} = A\frac{\partial \boldsymbol{U}}{\partial x}B$
$\boldsymbol{U} = \boldsymbol{U}(x)$, $\boldsymbol{V} = \boldsymbol{V}(x)$ are matrix functions of x	$\frac{\partial(\boldsymbol{U}+\boldsymbol{V})}{\partial x} = \frac{\partial \boldsymbol{U}}{\partial x} + \frac{\partial \boldsymbol{V}}{\partial x}$
$\boldsymbol{U} = \boldsymbol{U}(x)$, $\boldsymbol{V} = \boldsymbol{V}(x)$ are matrix functions of x	$\frac{\partial(\boldsymbol{UV})}{\partial x} = \boldsymbol{U}\frac{\partial \boldsymbol{V}}{\partial x} + \frac{\partial \boldsymbol{U}}{\partial x}\boldsymbol{V}$
$\boldsymbol{U} = \boldsymbol{U}(x)$ is a matrix function of x	$\frac{\partial \boldsymbol{U}^{-1}}{\partial x} = -\boldsymbol{U}^{-1}\frac{\partial \boldsymbol{U}}{\partial x}\boldsymbol{U}^{-1}$
$\boldsymbol{U} = \boldsymbol{U}(x, y)$ is a matrix function of x and y. Note that $\boldsymbol{U}$ is not scalar-by-matrix function, it is really vector-by-matrix function	$\frac{\partial \boldsymbol{U}^{-1}}{\partial x \partial y} = -\boldsymbol{U}^{-1}\left(\frac{\partial \boldsymbol{U}}{\partial x}\boldsymbol{U}^{-1}\frac{\partial \boldsymbol{U}}{\partial y} - \frac{\partial^2 \boldsymbol{U}}{\partial x \partial y} + \frac{\partial \boldsymbol{U}}{\partial y}\boldsymbol{U}^{-1}\frac{\partial \boldsymbol{U}}{\partial x}\right)\boldsymbol{U}^{-1}$
A is square matrix constant. $\boldsymbol{G}(.)$ is any matrix polynomial $\boldsymbol{P}(\boldsymbol{X})$ or matrix-power-series function such as $\boldsymbol{E}^{\boldsymbol{X}}$, $\boldsymbol{LNX}$, $\boldsymbol{SINX}$, $\boldsymbol{COSX}$.	$\frac{\partial \boldsymbol{G}(xA)}{\partial x} = A\boldsymbol{G}'(xA) = \boldsymbol{G}'(xA)A$ Following are some popular derivatives $\frac{\partial \boldsymbol{P}(xA)}{\partial x} = \frac{\partial \sum_{k=0}^{n} a_k (xA)^k}{\partial x} = A\boldsymbol{P}'(xA) = A\sum_{k=1}^{n} k a_k (xA)^{k-1} = \left(\sum_{k=1}^{n} k a_k (xA)^{k-1}\right)A$ $\frac{\partial (xA)^n}{\partial x} = nA(xA)^{n-1}$ $= n(xA)^{n-1}A$ where n is positive integer $\frac{\partial \boldsymbol{E}^{xA}}{\partial x} = A\boldsymbol{E}^{xA} = \boldsymbol{E}^{xA}A$ $\frac{\partial \boldsymbol{SIN}(xA)}{\partial x} = A\boldsymbol{COS}(xA) = \boldsymbol{COS}(xA)A$ $\frac{\partial \boldsymbol{COS}(xA)}{\partial x} = -A\boldsymbol{SIN}(xA) = -\boldsymbol{SIN}(xA)A$
$\boldsymbol{U} = \boldsymbol{U}(x)$ is a matrix function of x and $\boldsymbol{U}$ is square matrix. $\boldsymbol{E}^{\boldsymbol{X}}$ is matrix exponential function.	$\frac{\partial \boldsymbol{E}^{\boldsymbol{U}(x)}}{\partial x} = \int_0^1 \boldsymbol{E}^{t\boldsymbol{U}(x)}\frac{\partial \boldsymbol{U}}{\partial x}\boldsymbol{E}^{(1-t)\boldsymbol{U}(x)}dt$

Scalar-by-scalar composite derivative with vector and matrix involved

There are many cases in which a scalar-by-scalar function is composed of vector and matrix functions, which requires some efforts to solve some new problems although scalar-by-scalar derivative is much simpler than vector and matrix derivative. Following are composite derivatives of typical scalar-by-scalar functions involved vector and matrix. Note that independent variable x is always scalar and derivative of scalar-by-scalar function results out scalar.

Condition	Vector-by-vector derivative
$\boldsymbol{u} = \boldsymbol{u}(x)$ is a vector-by-scalar function. f is scalar-by-vector function	$\frac{\partial f(\boldsymbol{u})}{\partial x} = \frac{\partial f(\boldsymbol{u})}{\partial \boldsymbol{u}} \frac{\partial \boldsymbol{u}}{\partial x}$ Note that $\frac{\partial f(\boldsymbol{u})}{\partial \boldsymbol{u}}$ is row gradient vector and $\frac{\partial \boldsymbol{u}}{\partial x}$ is tangent vector. Therefore, the result $\frac{\partial f(\boldsymbol{u})}{\partial \boldsymbol{u}} \frac{\partial \boldsymbol{u}}{\partial x}$ is really a scalar product of two vectors, which produces a scalar.
$\boldsymbol{u} = \boldsymbol{u}(x)$, $\boldsymbol{v} = \boldsymbol{v}(x)$ are vector-by-scalar functions	$\frac{\partial \boldsymbol{u}^T \boldsymbol{v}}{\partial x} = \boldsymbol{u}^T \frac{\partial \boldsymbol{v}}{\partial x} + \left(\frac{\partial \boldsymbol{u}}{\partial x}\right)^T \boldsymbol{v}$ This is the derivative of scalar product of two vectors when $\frac{\partial \boldsymbol{u}}{\partial x}$ and $\frac{\partial \boldsymbol{v}}{\partial x}$ are tangent vectors.
$\boldsymbol{U} = \boldsymbol{U}(x)$ is matrix-by-scalar function	$\frac{\partial \lvert \boldsymbol{U} \rvert}{\partial x} = \lvert \boldsymbol{U} \rvert tr\left(\boldsymbol{U}^{-1} \frac{\partial \boldsymbol{U}}{\partial x}\right)$
$\boldsymbol{U} = \boldsymbol{U}(x)$ is matrix-by-scalar function	$\frac{\partial ln \lvert \boldsymbol{U} \rvert}{\partial x} = tr\left(\boldsymbol{U}^{-1} \frac{\partial \boldsymbol{U}}{\partial x}\right)$
$\boldsymbol{U} = \boldsymbol{U}(x)$ is matrix-by-scalar function	$\frac{\partial^2 \lvert \boldsymbol{U} \rvert}{\partial x^2} = \lvert \boldsymbol{U} \rvert \left(tr\left(\boldsymbol{U}^{-1} \frac{\partial^2 \boldsymbol{U}}{\partial x^2}\right) + \left(tr\left(\boldsymbol{U}^{-1} \frac{\partial \boldsymbol{U}}{\partial x}\right)\right)^2 - tr\left(\left(\boldsymbol{U}^{-1} \frac{\partial \boldsymbol{U}}{\partial x}\right)^2\right)\right)$
$\boldsymbol{U} = \boldsymbol{U}(x)$ is matrix-by-scalar function. f is scalar-by-matrix function	$\frac{\partial g(\boldsymbol{U})}{\partial x} = tr\left(\frac{\partial g(\boldsymbol{U})}{\partial \boldsymbol{U}} \frac{\partial \boldsymbol{U}}{\partial x}\right)$
A is square matrix constant. $\boldsymbol{G}(.)$ is any matrix polynomial $\boldsymbol{P}(\boldsymbol{X})$ or matrix-power-series function such as $\boldsymbol{E}^{\boldsymbol{X}}$, $\boldsymbol{LNX}$, $\boldsymbol{SINX}$, $\boldsymbol{COSX}$.	$\frac{\partial tr\left(\boldsymbol{G}(xA)\right)}{\partial x} = tr\left(A\boldsymbol{G}'(xA)\right)$ Following are some popular derivatives $\frac{\partial tr\left(\boldsymbol{P}(xA)\right)}{\partial x} = \frac{\partial tr\left(\sum_{k=0}^{n} a_k (xA)^k\right)}{\partial x} = tr\left(A\boldsymbol{P}'(xA)\right) = tr\left(A \sum_{k=1}^{n} k a_k (xA)^{k-1}\right)$

	$\frac{\partial tr((xA)^n)}{\partial x}$ $= tr(nA(xA)^{n-1})$ where n is positive integer $\frac{\partial tr(\boldsymbol{E}^{xA})}{\partial x} = tr(A\boldsymbol{E}^{xA})$ $\frac{\partial tr(\boldsymbol{SIN}(xA))}{\partial x} = tr(A\boldsymbol{COS}(xA))$ $\frac{\partial tr(\boldsymbol{COS}(xA))}{\partial x} = -tr(A\boldsymbol{SIN}(xA))$

5. Some applications of matrix analysis and calculus

There are a lot of applications of matrix analysis and calculus, for example, statistics, data analysis, optimization, geometry, kinematics, and dynamics. These applications are huge and there are many subjects related them. In the short report, we only introduce typical examples of matrix analysis and calculus. Some aforementioned examples such as Jordan decomposition and singular value decomposition are also good examples for interesting matrix applications. This section has two examples:

- Euclidean distance with general metric matrix (Hardle & Simar, 2013, pp. 71-74) illustrates applications of matrix analysis.
- Optimization with Lagrangian duality illustrates applications of matrix calculus.

Euclidean distance with general metric matrix

Distance d is defined as the function from $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}_+$:

$$\mathbb{R}^n \times \mathbb{R}^n \longmapsto \mathbb{R}_+$$

$$d: (\boldsymbol{x}, \boldsymbol{y}) \longrightarrow r \geq 0$$

Where $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors in vector space over real field $\mathbb{R}^n$.

Distance d is always greater than or equal to 0, which satisfies three following axioms (Hardle & Simar, 2013, p. 71).

$$\begin{cases} d(\boldsymbol{x}, \boldsymbol{y}) > 0 \; \forall \boldsymbol{x} \neq \boldsymbol{y} \\ d(\boldsymbol{x}, \boldsymbol{y}) = 0 \text{ if and only if } \boldsymbol{x} = \boldsymbol{y} \\ d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x}) \\ d(\boldsymbol{x}, \boldsymbol{y}) \leq d(\boldsymbol{x}, \boldsymbol{z}) + d(\boldsymbol{y}, \boldsymbol{z}) \; \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \end{cases}$$

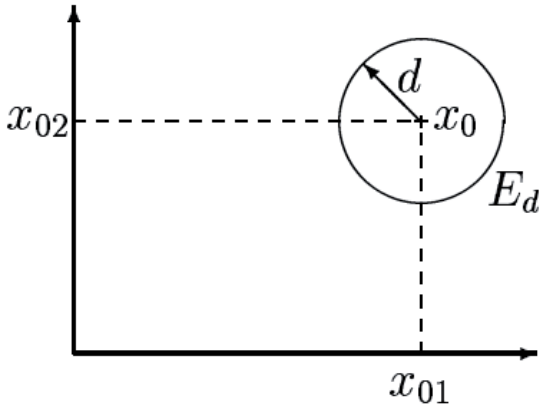
The Euclidean distance between two points is defined as following equation (Hardle & Simar, 2013, p. 71).

$$d^2(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})^T \boldsymbol{A} (\boldsymbol{x} - \boldsymbol{y})$$

Where $\boldsymbol{A}$ is positive definite matrix ($\boldsymbol{A} > 0$) and $\boldsymbol{A}$ is called a metric and the space on which $\boldsymbol{A}$ is defined is called metric space. Euclidean distance is concerned. If $\boldsymbol{A}$ is identical matrix $\boldsymbol{A} = \boldsymbol{I}_n$, Euclidean distance defined by $d^2(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} (x_i - y_i)^2$. Given a point $\boldsymbol{x}_0$ and a scalar constant d, a n-dimension sphere with radius d is defined as following equation.

$$(\boldsymbol{x} - \boldsymbol{x}_0)^T (\boldsymbol{x} - \boldsymbol{x}_0) = d^2$$

Following is figure of d-radius sphere (Hardle & Simar, 2013, p. 72):

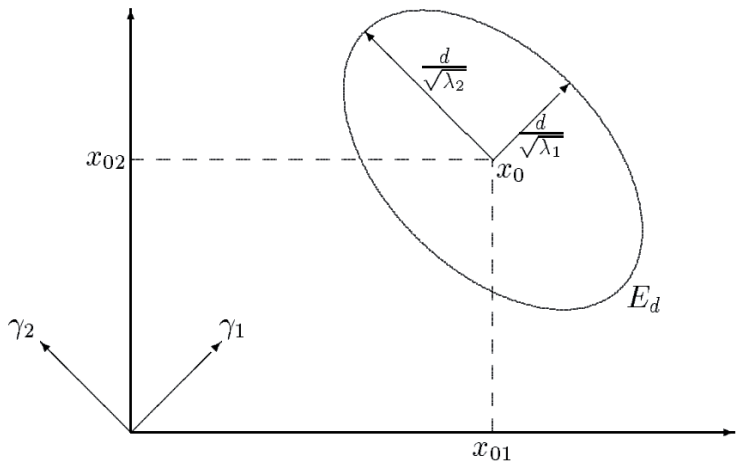


Where E_d denotes Euclidean metric space.

This sphere is a set of points which are far from $\boldsymbol{x}_0$ a distance d. Given a point $\boldsymbol{x}_0$, a matrix $\boldsymbol{A}(n_x n)$ and a scalar d, a n-dimension ellipsoid is defined as following equation.

$$(\boldsymbol{x} - \boldsymbol{x}_0)^T \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_0) = d^2$$

Following is figure of Ellipsoid with center $\boldsymbol{x}_0$, matrix $\boldsymbol{A}$ and constant d (Hardle & Simar, 2013, p. 72):



Note that $\boldsymbol{A}$ is invertible matrix. Suppose matrix $\boldsymbol{A}$ has n eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ and n respective orthogonal eigenvectors $\gamma_1, \gamma_2, \ldots, \gamma_n$. This ellipsoid has following properties (Hardle & Simar, 2013, p. 73):

- The principle axes of ellipsoid have the same direction to eigenvectors $\gamma_1, \gamma_2, \ldots, \gamma_n$. Of course, the number of axes is equal to the number of eigenvectors.

- The half of length of each axes is equal to $\sqrt{\frac{d^2}{\lambda_i}}$ where λ_i (s) are eigenvalues.
- Let x_{0i} be the element i of center $\boldsymbol{x}_0 = \{x_{01}, x_{02}, \ldots, x_{0n}\}$. The n-dimension rectangle surrounding ellipsoid is determined by equation: $x_{0i} - \sqrt{d^2 a^{ii}} \leq x_i \leq x_{0i} + \sqrt{d^2 a^{ii}}$ where a^{ii} is the element (i, i) of $\boldsymbol{A}^{-1}$.
- The coordinate of tangency point (x_i) between ellipsoid and its surrounding rectangle in the positive direction of j^{th} axis is $x_i = \left(a^{ij}\sqrt{\frac{d^2}{a^{jj}}}\right)$ where a^{ij} and a^{jj} is the elements (i, j) and (j, j) of $\boldsymbol{A}^{-1}$.

Optimization with Lagrangian function

Given a scalar-by-vector function $f(\boldsymbol{x})$, there is a requirement "how to minimize f with a scalar-by-vector constraint $h(\boldsymbol{x})$" This is optimization problem also called non-linear programming, specialized by following extreme-value expression.

$$\begin{aligned} &\text{minimize} \quad f(\boldsymbol{x}) \\ &\text{subject to} \quad h(\boldsymbol{x}) = 0 \end{aligned}$$

Where $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$ and f, h have continuous second partial derivatives. Function f is called target function and h is called constraint function.

Suppose $\boldsymbol{x}^*$ is extreme point of target function f satisfying constraint $h(\boldsymbol{x}^*) = 0$. Given the hyper-surface (surface in $\mathbb{R}^3$) determined equation $h(\boldsymbol{x}) = 0$, for each curve $\boldsymbol{\gamma}(t)$ lying on this hyper-surface together with condition that $\gamma(t)$ goes through $\boldsymbol{x}^*$ at $t = 0$, we have (Dinh, Pham, & Ta, 2002, pp. 59-61):

$$\boldsymbol{\gamma}: \mathbb{R} \to \mathbb{R}^n$$

$$\boldsymbol{\gamma}(0) = \boldsymbol{x}^*$$

$$h\big(\boldsymbol{\gamma}(0)\big) = h(\boldsymbol{x}^*) = 0$$

Note that $\boldsymbol{\gamma}(t)$ is vector-by-scalar function. The tangent vector of $\boldsymbol{\gamma}(t)$ denoted $\boldsymbol{\tau}$ at point $\boldsymbol{x}^*$ is:

$$\boldsymbol{\tau} = \boldsymbol{\gamma}'(0) = \frac{\partial \boldsymbol{\gamma}}{\partial t}(0)$$

The tangent space (tangent plane in $\mathbb{R}^3$) denoted $\mathbb{T}$ of hyper-surface $h(\boldsymbol{x}) = 0$ at point $\boldsymbol{x}^*$ is composed of all tangent vectors of curves $\boldsymbol{\gamma}(t)$.

$$\mathbb{T} = \{\boldsymbol{\tau} = \boldsymbol{\gamma}'(0) \text{ for all curves } \boldsymbol{\gamma}(t)\}$$

The derivative of $h(\boldsymbol{x})$ which is the gradient ∇h is orthogonal to tangent space $\mathbb{T}$ at $\boldsymbol{x}^*$; in other words it lies in the orthogonal complement $\mathbb{T}^\perp$ of tangent space $\mathbb{T}$.

$$\nabla h(\boldsymbol{x}^*)\boldsymbol{\tau} = 0, \forall \boldsymbol{\tau} \in \mathbb{T}$$

Function f gets extreme value at $\boldsymbol{x}^*$ and so its first-order derivative at $\boldsymbol{x}^*$ is 0. According to chain rule (Dinh, Pham, & Ta, 2002, pp. 49-50), we have:

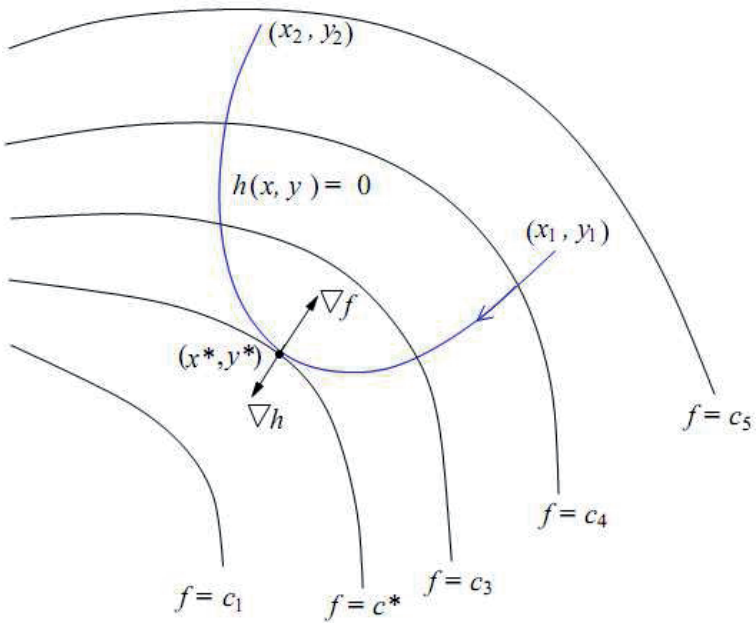
$$0 = f'(\boldsymbol{x}^*) = f'\big(\boldsymbol{\gamma}(0)\big) = \nabla f(\boldsymbol{x}^*)\boldsymbol{\gamma}'(0) = \nabla f(\boldsymbol{x}^*)\boldsymbol{\tau}, \forall \boldsymbol{\tau} \in \mathbb{T}$$

Where ∇f is gradient of f.

Hence, both gradients ∇f and ∇h are orthogonal to all tangent vector $\boldsymbol{\tau}$ at extreme point $\boldsymbol{x}^*$, which leads to conclusion that gradient ∇f is parallel with the gradient ∇h at $\boldsymbol{x}^*$. Consequently, it is reasoned out that there exists a scalar λ such that:

$$\nabla f(\boldsymbol{x}^*) + \lambda \nabla h(\boldsymbol{x}^*) = 0$$

For illustrating the assertion "gradient ∇f is parallel with gradient ∇h at extreme point $\boldsymbol{x}^*$", we suppose that $f(x, y)$ and $h(x, y)$ are 2-variable functions. The constraint function $h(x, y)$ is reduced to a curve superposed by a set of contours of target function f. Suppose we have 5 contours $f = c_1 < f = c^* < f = c_3 < f = c_4 < c_5$ where each value c_i is the value of contour of f and c^* is extreme value of f. Following is the figure (Jia, 2013, p. 2) illustrating constraint function $h(x, y)$ and contours of target function $f(x, y)$.



Please pay attention that when projecting the gradient vector of function $y = f(x, y)$ onto the plane containing its contour $f(x, y) = c$ then the gradient is orthogonal to the tangent vector of the contour. According to (Jia, 2013, pp. 2-3), it is easy to recognize that when moving along the curve $h(x, y) = 0$, the value of contour is increased and decreased and function f gets extreme value so that $h(x, y) =$

0 at point (x^*, y^*). Although $f = c_1$ is minimum but $h(x, y)$ does not intersect with contour $f(x, y) = c_1$. It means that gradient ∇f is parallel with gradient ∇h.

In general, the extreme points $\boldsymbol{x}^*$ are solutions of equations:

$$\begin{cases} h(\boldsymbol{x}) = 0 \\ \nabla f + \lambda \nabla h = 0 \end{cases}$$

Let $l(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda h(\boldsymbol{x})$ be Lagrangian scalar-by-vector function, we have:

$$\frac{\partial l(\boldsymbol{x}, \lambda)}{\partial (\boldsymbol{x}, \lambda)} = \nabla l = (\nabla f + \lambda \nabla h, h)$$

Note that both $\boldsymbol{x}$ and λ are independent variable in Lagrangian function and so the notation $\frac{\partial l(\boldsymbol{x}, \lambda)}{\partial (\boldsymbol{x}, \lambda)}$ indicates derivative of Lagrangian function w.r.t vector $(\boldsymbol{x}^T, \lambda) = (x_1, x_2, \ldots, x_n, \lambda)$.

Derived from conclusion "extreme points $\boldsymbol{x}^*$ are solutions of equations $h(\boldsymbol{x}) = 0$ and $\nabla f + \lambda \nabla h = 0$", principle Lagrange states that if $\boldsymbol{x}^*$ is extreme point, it always exists the real scalar λ^* so that the gradient ∇l is equal to $\mathbf{0}$ when substituting $(\boldsymbol{x}^*, \lambda^*)$ into it.

$$\frac{\partial l(\boldsymbol{x}, \lambda)}{\partial (\boldsymbol{x}, \lambda)} (\boldsymbol{x}^*, \lambda^*) = \nabla L(\boldsymbol{x}^*, \lambda^*) = \begin{pmatrix} \nabla f(\boldsymbol{x}^*) + \lambda^* \nabla h(\boldsymbol{x}^*) \\ h(\boldsymbol{x}^*) \end{pmatrix}^{\mathrm{T}} = (\mathbf{0}_n^T, 0)$$

Note that $(\mathbf{0}_n^T, 0)$ is zero $1_x(n+1)$ row vector. In brief, we have:

$$\nabla l(\boldsymbol{x}^*, \lambda^*) = \mathbf{0}_{n+1}^T$$

Note that Lagrangian function is scalar-by-vector function and so its derivative is really row gradient vector. Variables $\boldsymbol{x}^*$ and λ^* are always co-existent, hence, λ is called Lagrangian multiplier or Lagrangian dual variable. Lagrange principle is known as Lagrangian duality. If $x \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$ then, Lagrangian function has $n + 1$ partial elements and so it has $n + 1$ partial derivatives. Therefore, in practice, we set $n + 1$ these partial derivatives to be 0 so as to compose a set of $n + 1$ equations. After that Lagrangian multiplier λ^* and extreme point $\boldsymbol{x}^*$ are solution of such equations, as follows:

$$\begin{cases} \dfrac{\partial f}{\partial x_1} + \lambda \dfrac{\partial h}{\partial x_1} = 0 \\ \dfrac{\partial f}{\partial x_2} + \lambda \dfrac{\partial h}{\partial x_2} = 0 \\ \vdots \\ \dfrac{\partial f}{\partial x_n} + \lambda \dfrac{\partial h}{\partial x_n} = 0 \\ h(\boldsymbol{x}) = 0 \end{cases}$$

If it is too difficult to solve these equations by primary transformation, there are many methods to solve them, for example, Newton's method.

For example, let us find out extreme value of $f(x, y) = xy$ with constraint

$$h(x, y) = x^2/8 + y^2/2 - 1 = 0$$

The Lagrangian function and its gradient are constructed as below:

$$l(x, y, \lambda) = xy + \lambda(x^2/8 + y^2/2 - 1)$$

$$\nabla l = \frac{\partial l(\boldsymbol{x}, \lambda)}{\partial(\boldsymbol{x}, \lambda)} = \left(y + \frac{\lambda x}{4}, x + \lambda y, \frac{x^2}{8} + \frac{y^2}{2} - 1\right)$$

Lagrangian multiplier λ^* and extreme point $\boldsymbol{x}^*$ are solutions of equations constructed from setting ∇l to be $\mathbf{0}$.

$$\nabla l = \mathbf{0} \Leftrightarrow \begin{cases} y + \dfrac{\lambda x}{4} = 0 \\ x + \lambda y = 0 \\ \dfrac{x^2}{8} + \dfrac{y^2}{2} - 1 = 0 \end{cases}$$

$$\Rightarrow \begin{bmatrix} \lambda = 2, \text{two extreme points } (x = -2, y = 1) \text{ or } (x = 2, y = -1) \\ \lambda = -2, \text{two extreme points } (x = 2, y = 1) \text{ or } (x = -2, y = -1) \end{bmatrix}$$

We have:

$$\begin{bmatrix} f(-2,1) = -2 \\ f(2,-1) = -2 \\ f(2,1) = 2 \\ f(-2,-1) = 2 \end{bmatrix}$$

Therefore, f gets local maximum at (2, 1), (–2, –1) and local minimum at (–2, 1), (2, –1).

Now the optimization problem is extended with multi-constraints

$$\begin{aligned} \text{Minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & h_1(\boldsymbol{x}) = 0 \\ & h_2(\boldsymbol{x}) = 0 \\ & \vdots \\ & h_m(\boldsymbol{x}) = 0 \end{aligned}$$

Or

$$\begin{aligned} \text{Minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0} \end{aligned}$$

Hence, the constraint function becomes vector-by-vector function $\boldsymbol{h}(\boldsymbol{x}) = \begin{pmatrix} h_1(\boldsymbol{x}) \\ h_2(\boldsymbol{x}) \\ \vdots \\ h_m(\boldsymbol{x}) \end{pmatrix}$. For convenience, the derivative of function $\boldsymbol{h}(\boldsymbol{x})$ which is Jacobian matrix is denoted by gradient notation.

$$\nabla \boldsymbol{h} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}} = \begin{pmatrix} \dfrac{\partial h_1}{\partial \boldsymbol{x}} \\ \dfrac{\partial h_2}{\partial \boldsymbol{x}} \\ \vdots \\ \dfrac{\partial h_m}{\partial \boldsymbol{x}} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial h_1}{\partial x_1} & \dfrac{\partial h_1}{\partial x_2} & \cdots & \dfrac{\partial h_1}{\partial x_n} \\ \dfrac{\partial h_2}{\partial x_1} & \dfrac{\partial h_2}{\partial x_2} & \cdots & \dfrac{\partial h_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{\partial h_m}{\partial x_1} & \dfrac{\partial h_m}{\partial x_2} & \cdots & \dfrac{\partial h_m}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \nabla h_1 \\ \nabla h_2 \\ \vdots \\ \nabla h_m \end{pmatrix}$$

$$\text{Where } \nabla h_i = \left(\frac{\partial h_i}{\partial x_1}, \frac{\partial h_i}{\partial x_2}, \ldots, \frac{\partial h_i}{\partial x_n}\right), \forall i = \overline{1, m}$$

Given extreme point $\boldsymbol{x}^*$ satisfying $\boldsymbol{h}(\boldsymbol{x}^*) = \boldsymbol{0}$ is regular point if Jacobian matrix $\nabla \boldsymbol{h}$ is non-singular at $\boldsymbol{x}^*$. The tangent space $\mathbb{T}$ at $\boldsymbol{x}^*$ exists if and only if $\boldsymbol{x}^*$ is regular point and so the Lagrangian function exists if and only if Jacobian matrix $\nabla \boldsymbol{h}$ is non-singular at $\boldsymbol{x}^*$. The tangent space at $\boldsymbol{x}^*$ is null space of $\nabla \boldsymbol{h}(\boldsymbol{x}^*)$, it means:

$$\forall \boldsymbol{\tau} \in \mathbb{T}, \nabla \boldsymbol{h}(\boldsymbol{x}^*)\boldsymbol{\tau} = \boldsymbol{0}$$

Note that $\boldsymbol{0}$ is zero column vector and we still have:

$$\nabla f(\boldsymbol{x}^*)\boldsymbol{\tau} = 0, \forall \boldsymbol{\tau} \in \mathbb{T}$$

Both gradient ∇f and Jacobian matrix $\nabla \boldsymbol{h}$ lie in complement $\mathbb{T}^\perp$ of tangent space $\mathbb{T}$ at $\boldsymbol{x}^*$, so there is a real number vector $\boldsymbol{\lambda} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_m \end{pmatrix}$ such that:

$$\nabla f(\boldsymbol{x}^*) + \boldsymbol{\lambda}^T \nabla \boldsymbol{h}(\boldsymbol{x}^*) = 0$$

and

$$\boldsymbol{h}(\boldsymbol{x}^*) = \boldsymbol{0}$$

Lagrangian function is re-constructed:

$$l(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^T \boldsymbol{h}(\boldsymbol{x})$$

According to principle Lagrange, if $\boldsymbol{x}^*$ is extreme point, it always exists the m_x1 vector $\boldsymbol{\lambda}^*$ such that

$$\nabla l(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) = \begin{pmatrix} \nabla f(\boldsymbol{x}^*) + \boldsymbol{\lambda}^{*T} \nabla \boldsymbol{h}(\boldsymbol{x}^*) \\ \left(\boldsymbol{h}(\boldsymbol{x}^*)\right)^T \end{pmatrix}^T = (\boldsymbol{0}_n^T, \boldsymbol{0}_m^T)$$

Note that $(\boldsymbol{0}_n^T, \boldsymbol{0}_m^T)$ is zero $1_x(n+m)$ row vector. In brief, we have:

$$\nabla l(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) = \boldsymbol{0}_{n+m}^T$$

Lagrangian function with m constraints has $n + m$ partial elements and so it has $n + m$ partial derivatives. Therefore, in practice, we set $n + m$ these partial derivatives to be 0 so as to construct a set of $n + m$ equations whose solutions are extreme point $\boldsymbol{x}^*$ and Lagrangian multipliers vector $\boldsymbol{\lambda}^*$, as follows:

$$\begin{cases} \dfrac{\partial f}{\partial x_1} + \sum_{i=1}^{m} \lambda_i \dfrac{\partial h_i}{\partial x_1} = 0 \\ \dfrac{\partial f}{\partial x_2} + \sum_{i=1}^{m} \lambda_i \dfrac{\partial h_i}{\partial x_2} = 0 \\ \vdots \\ \dfrac{\partial f}{\partial x_n} + \sum_{i=1}^{m} \lambda_i \dfrac{\partial h_i}{\partial x_n} = 0 \\ h_1(\boldsymbol{x}) = 0 \\ h_2(\boldsymbol{x}) = 0 \\ \vdots \\ h_m(\boldsymbol{x}) = 0 \end{cases}$$

Lagrangian duality can be extended to solve optimization problem with inequality constraints. This is the most general case of non-linear programming, as follows:

$$\begin{aligned} &\text{Minimize} && f(\boldsymbol{x}) \\ &\text{subject to} && \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0} \\ &&& \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0} \end{aligned}$$

Where $\boldsymbol{h}(\boldsymbol{x}) = \begin{pmatrix} h_1(\boldsymbol{x}) \\ h_2(\boldsymbol{x}) \\ \vdots \\ h_m(\boldsymbol{x}) \end{pmatrix}$ is equality constraint vector-by-vector function and $\boldsymbol{g}(\boldsymbol{x}) = \begin{pmatrix} g_1(\boldsymbol{x}) \\ g_2(\boldsymbol{x}) \\ \vdots \\ g_p(\boldsymbol{x}) \end{pmatrix}$ is inequality constraint vector-by-vector function. Lagrangian function is re-constructed:

$$l(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^T \boldsymbol{h}(\boldsymbol{x}) + \boldsymbol{\mu}^T \boldsymbol{g}(\boldsymbol{x})$$

Where $\boldsymbol{\lambda} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_m \end{pmatrix}$ is real number vector and $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$ with attention that $\mu_i \geq 0, \forall i$. According to principle Lagrange, if $\boldsymbol{x}^*$ is extreme point, it always exists $m_x 1$ vector $\boldsymbol{\lambda}^*$ and $p_x 1$ vector $\boldsymbol{\mu}^*$ such that

$$\nabla l(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = \begin{pmatrix} \nabla f(\boldsymbol{x}^*) + \boldsymbol{\lambda}^{*T} \nabla \boldsymbol{h}(\boldsymbol{x}^*) + \boldsymbol{\mu}^{*T} \nabla \boldsymbol{g}(\boldsymbol{x}^*) \\ \left(\boldsymbol{h}(\boldsymbol{x}^*)\right)^T \\ \left(\boldsymbol{g}(\boldsymbol{x}^*)\right)^T \end{pmatrix}^T = (\boldsymbol{0}_n^T, \boldsymbol{0}_m^T, \boldsymbol{0}_p^T)$$

Where,

$$\nabla \boldsymbol{h} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}} = \begin{pmatrix} \dfrac{\partial h_1}{\partial \boldsymbol{x}} \\ \dfrac{\partial h_2}{\partial \boldsymbol{x}} \\ \vdots \\ \dfrac{\partial h_m}{\partial \boldsymbol{x}} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial h_1}{\partial x_1} & \dfrac{\partial h_1}{\partial x_2} & \cdots & \dfrac{\partial h_1}{\partial x_n} \\ \dfrac{\partial h_2}{\partial x_1} & \dfrac{\partial h_2}{\partial x_2} & \cdots & \dfrac{\partial h_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{\partial h_m}{\partial x_1} & \dfrac{\partial h_m}{\partial x_2} & \cdots & \dfrac{\partial h_m}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \nabla h_1 \\ \nabla h_2 \\ \vdots \\ \nabla h_m \end{pmatrix}$$

$$\nabla h_i = \left(\frac{\partial h_i}{\partial x_1}, \frac{\partial h_i}{\partial x_2}, \ldots, \frac{\partial h_i}{\partial x_n}\right), \forall i = \overline{1,m}$$

$$\nabla \boldsymbol{g} = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}} = \begin{pmatrix} \frac{\partial g_1}{\partial \boldsymbol{x}} \\ \frac{\partial g_2}{\partial \boldsymbol{x}} \\ \vdots \\ \frac{\partial g_p}{\partial \boldsymbol{x}} \end{pmatrix} = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial g_p}{\partial x_1} & \frac{\partial g_p}{\partial x_2} & \cdots & \frac{\partial g_p}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \nabla g_1 \\ \nabla g_2 \\ \vdots \\ \nabla g_p \end{pmatrix}$$

$$\nabla g_i = \left(\frac{\partial g_i}{\partial x_1}, \frac{\partial g_i}{\partial x_2}, \ldots, \frac{\partial g_i}{\partial x_n}\right), \forall i = \overline{1,p}$$

Note that $(\mathbf{0}_n^T, \mathbf{0}_m^T, \mathbf{0}_p^T)$ is zero $1_x(n+m+p)$ row vector. In brief, we have:

$$\nabla l(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = \mathbf{0}_{n+m+p}^T$$

In practice, we set $n + m + p$ these partial derivatives to be 0 so as to construct a set of $n + m$ equations whose solutions are extreme point $\boldsymbol{x}^*$ and Lagrangian multipliers vectors $\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*$ as follows:

$$\begin{cases} \frac{\partial f}{\partial x_1} + \sum_{i=1}^{m} \lambda_i \frac{\partial h_i}{\partial x_1} + \sum_{j=1}^{p} \mu_j \frac{\partial g_j}{\partial x_1} = 0 \\ \frac{\partial f}{\partial x_2} + \sum_{i=1}^{m} \lambda_i \frac{\partial h_i}{\partial x_2} + \sum_{j=1}^{p} \mu_j \frac{\partial g_j}{\partial x_2} = 0 \\ \vdots \\ \frac{\partial f}{\partial x_n} + \sum_{i=1}^{m} \lambda_i \frac{\partial h_i}{\partial x_n} + + \sum_{j=1}^{p} \mu_j \frac{\partial g_j}{\partial x_n} = 0 \\ h_1(\boldsymbol{x}) = 0 \\ h_2(\boldsymbol{x}) = 0 \\ \vdots \\ h_m(\boldsymbol{x}) = 0 \\ g_1(\boldsymbol{x}) = 0 \\ g_2(\boldsymbol{x}) = 0 \\ \vdots \\ g_p(\boldsymbol{x}) = 0 \end{cases}$$

In general, we need to solve these equations above to find out optimal solutions of optimization problem.

6. Conclusion

Now we researched main subjects of matrix analysis and calculus over the report in which sections "Matrix analysis", "Matrix derivative" and "Composite derivative" are important ones. Matrix analysis focuses on data analyzing and data processing whereas matrix calculus focuses on real-time process, differential geometry, and optimization. Some domains apply both matrix analysis and matrix calculus. For example, multivariate statistics uses not only data analysis techniques but also other studies such as differential, calculus, graphics, combinatorics, and probability for analyzing data and testing hypothesis. Matrix analysis is very necessary to matrix calculus and so we should research it before studying matrix calculus and differential. Some enhanced subjects such as matrix-by-matrix functions and derivatives, Kronecker product, Hadamard product, complex matrix, and multivariate distribution are not mentioned in the report. We will discuss them in next research.

Acknowledgement

The report expresses my deep gratitude to Professor Ho, Minh T. at Vietnam Institute of Mathematics who taught me a lot of valuable knowledge and gave me the motivation to complete both our knowledge and the report.

Bibliography

Baker, K. (2013). *Singular Value Decomposition Tutorial.* The College of Humanities and Social Sciences, Linguistics Department Advisory Board. Montclair State University.

Baker, M. J. (n.d.). *Maths - Powers of Vectors*. Retrieved 2014, from EuclideanSpace - Mathematics and Computing: http://www.euclideanspace.com/maths/algebra/vectors/vecAlgebra/powers/index.htm

Dinh, L. T., Pham, D. H., & Ta, P. D. (2002). *Multivariate Analysis - Principles and Practices* (Vol. II). (K. H. Ha, T. V. Ngo, & D. H. Pham, Eds.) Hanoi, Vietnam: Hanoi National University Publisher.

Hardle, W., & Simar, L. (2013). *Applied Multivariate Statistical Analysis.* Berlin, Germany: Research Data Center, School of Business and Economics, Humboldt University.

Hoang, V. H. (2012, December 24). *Linear mapping - Diagonalizing matrix.* (V. H. Hoang, Performer) YouTube, Ho Chi Minh, Vietnam. Retrieved from https://www.youtube.com/watch?v=NkSDymM-qPg

Jia, Y.-B. (2013). *Lagrange Multipliers.* Lecture notes on course "Problem Solving Techniques for Applied Computer Science", Iowa State University of Science and Technology, USA.

Nguyen, V. H. (1999). *Linear Algebra.* Hanoi, Vietnam: Hanoi National University Publishing House.

Petersen, K. B., & Pedersen, M. S. (2012). *The Matrix Cookbook.* Copenhagen, Denmark: SAS Institute. Retrieved from http://orion.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Wikipedia. (2014, March 9). *Cayley–Hamilton theorem*. (Wikimedia Foundation) Retrieved from Wikipedia website: http://en.wikipedia.org/wiki/Cayley%E2%80%93Hamilton_theorem

Wikipedia. (2014, February 25). *Logarithm of a matrix*. (Wikimedia Foundation) Retrieved from Wikipedia website: http://en.wikipedia.org/wiki/Matrix_logarithm

Wikipedia. (2014, March 3). *Matrix calculus*. (Wikimedia Foundation) Retrieved from Wikipedia website: http://en.wikipedia.org/wiki/Matrix_calculus

Wikipedia. (2014, February 20). *Matrix exponential*. (Wikimedia Foundation) Retrieved from Wikipedia website: http://en.wikipedia.org/wiki/Matrix_exponential

Wikipedia. (2014, February 12). *Matrix function*. (Wikimedia Foundation) Retrieved from Wikipedia website: http://en.wikipedia.org/wiki/Matrix_function

Buy your books fast and straightforward online - at one of the world's fastest growing online book stores! Environmentally sound due to Print-on-Demand technologies.

Buy your books online at

www.get-morebooks.com

Kaufen Sie Ihre Bücher schnell und unkompliziert online – auf einer der am schnellsten wachsenden Buchhandelsplattformen weltweit!
Dank Print-On-Demand umwelt- und ressourcenschonend produziert.

Bücher schneller online kaufen

www.morebooks.de

OmniScriptum Marketing DEU GmbH
Heinrich-Böcking-Str. 6-8
D - 66121 Saarbrücken
Telefax: +49 681 93 81 567-9

info@omniscriptum.com
www.omniscriptum.com

OMNIScriptum