Matrix Calculus:
Derivation and Simple Application

HU, Pili*

March 30, 2012†

Abstract

Matrix Calculus[3] is a very useful tool in many engineering problems. Basic rules of matrix calculus are nothing more than ordinary calculus rules covered in undergraduate courses. However, using matrix calculus, the derivation process is more compact. This document is adapted from the notes of a course the author recently attends. It builds matrix calculus from scratch. Only prerequisites are basic calculus notions and linear algebra operation.

To get a quick executive guide, please refer to the cheat sheet in section(4).

To see how matrix calculus simplify the process of derivation, please refer to the application in section(3.4).

*hupili [at] ie [dot] cuhk [dot] edu [dot] hk

†Last compile:April 24, 2012

Contents

1 Introductory Example

We start with an one variable linear function:

$$f(x) = ax \tag{1}$$

To be coherent, we abuse the partial derivative notation:

$$\frac{\partial f}{\partial x} = a \tag{2}$$

Extending this function to be multivariate, we have:

$$f(x) = \sum_i a_i x_i = a^{\mathrm{T}} x \tag{3}$$

Where $a = [a_1, a_2, \ldots, a_n]^{\mathrm{T}}$ and $x = [x_1, x_2, \ldots, x_n]^{\mathrm{T}}$. We first compute partial derivatives directly:

$$\frac{\partial f}{\partial x_k} = \frac{\partial(\sum_i a_i x_i)}{\partial x_k} = a_k \tag{4}$$

for all $k = 1, 2, \ldots, n$. Then we organize n partial derivatives in the following way:

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = a \tag{5}$$

The first equality is by proper definition and the rest roots from ordinary calculus rules.

Eqn(5) is analogous to eqn(2), except the variable changes from a scalar to a vector. Thus we want to directly claim the result of eqn(5) without those intermediate steps solving for partial derivatives separately. Actually, we'll see soon that eqn(5) plays a core role in matrix calculus.

Following sections are organized as follows:

- Section(2) builds commonly used matrix calculus rules from ordinary calculus and linear algebra. Necessary and important properties of linear algebra is also proved along the way. This section is not organized afterhand. All results are proved when we need them.
- Section(3) shows some applications using matrix calculus. Table(1) shows the relation between Section(2) and Section(3).
- Section(4) concludes a cheat sheet of matrix calculus. Note that this cheat sheet may be different from others. Users need to figure out some basic definitions before applying the rules.

Table 1: Derivation and Application Correspondance

Derivation	Application
2.1-2.7	3.1
2.9,2.10	3.2
2.8,2.11	3.3

2 Derivation

2.1 Organization of Elements

From the introductary example, we already see that matrix calculus does not distinguish from ordinary calculus by fundamental rules. However, with better organization of elements and proving useful properties, we can simplify the derivation process in real problems.

The author would like to adopt the following definition:

Definition 1. *For a scalar valued function $f(x)$, the result $\frac{\partial f}{\partial x}$ has the same size with x. That is*

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{12}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \frac{\partial f}{\partial x_{21}} & \frac{\partial f}{\partial x_{22}} & \cdots & \frac{\partial f}{\partial x_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \frac{\partial f}{\partial x_{m2}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix} \tag{6}$$

In eqn(2), x is a 1-by-1 matrix and the result $\frac{\partial f}{\partial x} = a$ is also a 1-by-1 matrix. In eqn(5), x is a column vector(known as n-by-1 matrix) and the result $\frac{\partial f}{\partial x} = a$ has the same size.

Example 1. *By this definition, we have:*

$$\frac{\partial f}{\partial x^{\mathrm{T}}} = (\frac{\partial f}{\partial x})^{\mathrm{T}} = a^{\mathrm{T}} \tag{7}$$

Note that we only use the organization definition in this example. Later we'll show that with some matrix properties, this formula can be derived without using $\frac{\partial f}{\partial x}$ as a bridge.

2.2 Deal with Inner Product

Theorem 1. *If there's a multivariate scalar function $f(x) = a^{\mathrm{T}}x$, we have $\frac{\partial f}{\partial x} = a$.*

Proof. See introductary example. □

Since $a^{\mathrm{T}}x$ is scalar, we can write it equivalently as the trace of its own. Thus,

Proposition 2. *If there's a multivariate scalar function* $f(x) = \mathrm{Tr}\left[a^{\mathrm{T}}x\right]$*, we have* $\frac{\partial f}{\partial x} = a$*.*

$\mathrm{Tr}\left[\bullet\right]$ is the operator to sum up diagonal elements of a matrix. In the next section, we'll explore more properties of trace. As long as we can transform our target function into the form of theorem(1) or proposition(2), the result can be written out directly. Notice in proposition(2), a and x are both vectors. We'll show later as long as their sizes agree, it holds for matrix a and x.

2.3 Properties of Trace

Definition 2. *Trace of square matrix is defined as:* $\mathrm{Tr}\left[A\right] = \sum_i A_{ii}$

Example 2. *Using definition(1,2), it is very easy to show:*

$$\frac{\partial \mathrm{Tr}\left[A\right]}{\partial A} = I \tag{8}$$

since only diagonal elements are kept by the trace operator.

Theorem 3. *Matrix trace has the following properties:*

- *(1)* $\mathrm{Tr}\left[A + B\right] = \mathrm{Tr}\left[A\right] + \mathrm{Tr}\left[B\right]$
- *(2)* $\mathrm{Tr}\left[cA\right] = c\mathrm{Tr}\left[A\right]$
- *(3)* $\mathrm{Tr}\left[AB\right] = \mathrm{Tr}\left[BA\right]$
- *(4)* $\mathrm{Tr}\left[A_1A_2\ldots A_n\right] = \mathrm{Tr}\left[A_nA_1\ldots A_{n-1}\right]$
- *(5)* $\mathrm{Tr}\left[A^{\mathrm{T}}B\right] = \sum_i\sum_j A_{ij}B_{ij}$
- *(6)* $\mathrm{Tr}\left[A\right] = \mathrm{Tr}\left[A^{\mathrm{T}}\right]$

where A, B *are matrices with proper sizes, and* c *is a scalar value.*

Proof. See wikipedia [5] for the proof. □

Here we explain the intuitions behind each property to make it easier to remenber. Property(1) and property(2) shows the linearity of trace. Property(3) means two matrices' multiplication inside a the trace operator is commutative. Note that the matrix multiplication without trace is not commutative and the commutative property inside the trace does not hold for more than 2 matrices. Property (4) is the proposition of property (3) by considering $A_1 A_2 \dots A_{n-1}$ as a whole. It is known as cyclic property, so that you can rotate the matrices inside a trace operator. Property (5) shows a way to express the sum of element by element product using matrix product and trace. Note that inner product of two vectors is also the sum of element by element product. Property (5) resembles the vector inner product by form($A^{\mathrm{T}}B$). The author regards property (5) as the extension of inner product to matrices(Generalized Inner Product).

2.4 Deal with Generalized Inner Product

Theorem 4. *If there's a multivariate scalar function $f(x) = \mathrm{Tr}\left[A^{\mathrm{T}}x\right]$, we have $\frac{\partial f}{\partial x} = A$. ($A, x$ can be matrices).*

Proof. Using property (5) of trace, we can write f as:

$$f(x) = \mathrm{Tr}\left[A^{\mathrm{T}}x\right] = \sum_{ij} A_{ij}x_{ij} \tag{9}$$

It's easy to show:

$$\frac{\partial f}{\partial x_{ij}} = \frac{\partial(\sum_{ij} A_{ij}x_{ij})}{\partial x_{ij}} = A_{ij} \tag{10}$$

Organize elements using definition(1), it is proved. □

With this theorem and properties of trace we revisit example(1).

Example 3. *For vector a, x and function $f(x) = a^{\mathrm{T}}x$*

$$\begin{aligned}
& & \frac{\partial f}{\partial x^{\mathrm{T}}} & \quad (11)\\
& = & \frac{\partial (a^{\mathrm{T}}x)}{\partial x^{\mathrm{T}}} & \quad (12)\\
\textit{(f is scalar)} & = & \frac{\partial (\mathrm{Tr}\left[a^{\mathrm{T}}x\right])}{\partial x^{\mathrm{T}}} & \quad (13)\\
\textit{(property(3))} & = & \frac{\partial (\mathrm{Tr}\left[xa^{\mathrm{T}}\right])}{\partial x^{\mathrm{T}}} & \quad (14)\\
\textit{(property(6))} & = & \frac{\partial (\mathrm{Tr}\left[ax^{\mathrm{T}}\right])}{\partial x^{\mathrm{T}}} & \quad (15)\\
\textit{(property of transpose)} & = & \frac{\partial (\mathrm{Tr}\left[(a^{\mathrm{T}})^{\mathrm{T}}x^{\mathrm{T}}\right])}{\partial x^{\mathrm{T}}} & \quad (16)\\
\textit{(theorem(4))} & = & a^{\mathrm{T}} & \quad (17)
\end{aligned}$$

The result is the same with example(1), where we used the basic definition.

The above example actually demonstrates the usual way of handling a matrix derivative problem.

2.5 Define Matrix Differential

Although we want matrix derivative at most time, it turns out matrix differential is easier to operate due to the form invariance property of differential. Matrix differential inherit this property as a natural consequence of the following definition.

Definition 3. *Define matrix differential:*

$$\mathrm{d}A = \begin{bmatrix} \mathrm{d}A_{11} & \mathrm{d}A_{12} & \dots & \mathrm{d}A_{1n} \\ \mathrm{d}A_{21} & \mathrm{d}A_{22} & \dots & \mathrm{d}A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{d}A_{m1} & \mathrm{d}A_{m2} & \dots & \mathrm{d}A_{mn} \end{bmatrix} \tag{18}$$

Theorem 5. *Differential operator is distributive through trace operator:* $\mathrm{dTr}\left[A\right] = \mathrm{Tr}\left[\mathrm{d}A\right]$

Proof.

$$\mathrm{LHS} = \mathrm{d}(\sum_i A_{ii}) = \sum_i \mathrm{d}A_{ii} \tag{19}$$

$$\mathrm{RHS} = \mathrm{Tr}\begin{bmatrix} \mathrm{d}A_{11} & \mathrm{d}A_{12} & \dots & \mathrm{d}A_{1n} \\ \mathrm{d}A_{21} & \mathrm{d}A_{22} & \dots & \mathrm{d}A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{d}A_{m1} & \mathrm{d}A_{m2} & \dots & \mathrm{d}A_{mn} \end{bmatrix} \tag{20}$$

$$= \sum_i \mathrm{d}A_{ii} = \mathrm{LHS} \tag{21}$$

□

Now that matrix differential is well defined, we want to relate it back to matrix derivative. The scalar version differential and derivative can be related as follows:

$$\mathrm{d}f = \frac{\partial f}{\partial x}\mathrm{d}x \tag{22}$$

So far, we're dealing with scalar function f and matrix variable x. $\frac{\partial f}{\partial x}$ and $\mathrm{d}x$ are both matrix according to definition. In order to make the quantities in eqn(22) equal, we must figure out a way to make the RHS a scalar. It's not surprising that trace is what we want.

Theorem 6.

$$\mathrm{d}f = \mathrm{Tr}\left[(\frac{\partial f}{\partial x})^{\mathrm{T}}\mathrm{d}x\right] \tag{23}$$

for scalar function f and arbitrarily sized x.

Proof.

$$
\begin{aligned}
\text{LHS} &= \mathrm{d}f && (24)\\
\text{(definition of scalar differential)} &= \sum_{ij} \frac{\partial f}{\partial x_{ij}} \mathrm{d}x_{ij} && (25)\\
\text{RHS} &= \mathrm{Tr}\left[(\frac{\partial f}{\partial x})^{\mathrm{T}} \mathrm{d}x \right] && (26)\\
\text{(trace property (5))} &= \sum_{ij} (\frac{\partial f}{\partial x})_{ij} (\mathrm{d}x)_{ij} && (27)\\
\text{(definition(3))} &= \sum_{ij} (\frac{\partial f}{\partial x})_{ij} \mathrm{d}x_{ij} && (28)\\
\text{(definition(1))} &= \sum_{ij} \frac{\partial f}{\partial x_{ij}} \mathrm{d}x_{ij} && (29)\\
&= \text{LHS} && (30)
\end{aligned}
$$

□

Theorem(6) is the bridge between matrix derivative and matrix differential. We'll see in later applications that matrix differential is more convenient to manipulate. After certain manipulation we can get the form of theorem(6). Then we can directly write out matrix derivative using this theorem.

2.6 Matrix Differential Properties

Theorem 7. *We claim the following properties of matrix differential:*

- $\mathrm{d}(cA) = c\mathrm{d}A$
- $\mathrm{d}(A + B) = \mathrm{d}A + \mathrm{d}B$
- $\mathrm{d}(AB) = \mathrm{d}AB + A\mathrm{d}B$

Proof. They're all natural consequences given the definition(3). We only show the 3rd one in this document. Note that the equivalence holds if LHS and RHS are equivalent element by element. We consider the (ij)-th element.

$$
\begin{aligned}
\text{LHS}_{ij} &= \mathrm{d}(\sum_k A_{ik} B_{kj}) && (31)\\
&= \sum_k (\mathrm{d}A_{ik} B_{kj} + A_{ik} \mathrm{d}B_{kj}) && (32)\\
\text{RHS}_{ij} &= (\mathrm{d}AB)_{ij} + (A\mathrm{d}B)_{ij} && (33)\\
&= \sum_k \mathrm{d}A_{ik} B_{kj} + \sum_k A_{ik} \mathrm{d}B_{kj} && (34)\\
&= \text{LHS}_{ij} && (35)
\end{aligned}
$$

$\square$

Example 4. *Given the function $f(x) = x^{\mathrm{T}}Ax$, where A is square and x is a column vector, we can compute:*

$$\begin{aligned}
\mathrm{d}f &= \mathrm{dTr}\left[x^{\mathrm{T}}Ax\right] && (36)\\
&= \mathrm{Tr}\left[\mathrm{d}(x^{\mathrm{T}}Ax)\right] && (37)\\
&= \mathrm{Tr}\left[\mathrm{d}(x^{\mathrm{T}})Ax + x^{\mathrm{T}}\mathrm{d}(Ax)\right] && (38)\\
&= \mathrm{Tr}\left[\mathrm{d}(x^{\mathrm{T}})Ax + x^{\mathrm{T}}\mathrm{d}Ax + x^{\mathrm{T}}A\mathrm{d}x\right] && (39)\\
\textit{(A is constant)} &= \mathrm{Tr}\left[\mathrm{d}x^{\mathrm{T}}Ax + x^{\mathrm{T}}A\mathrm{d}x\right] && (40)\\
&= \mathrm{Tr}\left[\mathrm{d}x^{\mathrm{T}}Ax\right] + \mathrm{Tr}\left[x^{\mathrm{T}}A\mathrm{d}x\right] && (41)\\
&= \mathrm{Tr}\left[x^{\mathrm{T}}A^{\mathrm{T}}\mathrm{d}x\right] + \mathrm{Tr}\left[x^{\mathrm{T}}A\mathrm{d}x\right] && (42)\\
&= \mathrm{Tr}\left[x^{\mathrm{T}}A^{\mathrm{T}}\mathrm{d}x + x^{\mathrm{T}}A\mathrm{d}x\right] && (43)\\
&= \mathrm{Tr}\left[(x^{\mathrm{T}}A^{\mathrm{T}} + x^{\mathrm{T}}A)\mathrm{d}x\right] && (44)
\end{aligned}$$

Using theorem(6), we obtain the derivative:

$$\frac{\partial f}{\partial x} = (x^{\mathrm{T}}A^{\mathrm{T}} + x^{\mathrm{T}}A)^{\mathrm{T}} = Ax + A^{\mathrm{T}}x \tag{45}$$

When A is symmetric, it simplifies to:

$$\frac{\partial f}{\partial x} = 2Ax \tag{46}$$

Let $A = I$, we have:

$$\frac{\partial (x^{\mathrm{T}}x)}{\partial x} = 2x \tag{47}$$

Example 5. *For a non-singular square matrix X, we have $XX^{-1} = I$. Take matrix differentials at both sides:*

$$0 = \mathrm{d}I = \mathrm{d}(XX^{-1}) = \mathrm{d}XX^{-1} + X\mathrm{d}(X^{-1}) \tag{48}$$

Rearrange terms:

$$\mathrm{d}(X^{-1}) = -X^{-1}\mathrm{d}XX^{-1} \tag{49}$$

2.7 Schema of Hanlding Scalar Function

The above example already demonstrates the general schema. Here we conclude the process:

1. $\mathrm{d}f = \mathrm{dTr}\,[f] = \mathrm{Tr}\,[\mathrm{d}f]$
2. Apply trace properties(see theorem(3)) and matrix differential properties(see theorem(7)) to get the following form:

$$\mathrm{d}f = \mathrm{Tr}\left[A^{\mathrm{T}}x\right] \tag{50}$$

3. Apply theorem(6) to get:
$$\frac{\partial f}{\partial x} = A \tag{51}$$

To this point, you can handle many problems. In this schema, matrix differential and trace play crucial roles. Later we'll deduce some widely used formula to facilitate potential applications. As you will see, although we rely on matrix differential in the schema, the deduction of certain formula may be more easily done using matrix derivatives.

2.8 Determinant

For a background of determinant, please refer to [6]. We first quote some definitions and properties without proof:

Theorem 8. *Let A be a square matrix:*

- *The minor M_{ij} is obtained by remove i-th row and j-th column of A and then take determinant of the resulting (n-1) by (n-1) matrix.*
- *The ij-th cofactor is defined as $C_{ij} = (-1)^{i+j} M_{ij}$.*
- *If we expand determinant with respect to the i-th row,* $\det(A) = \sum_j A_{ij} C_{ij}$.
- *The adjugate of A is defined as* $\mathrm{adj}(A)_{ij} = (-1)^{i+j} M_{ji} = C_{ji}$. *So* $\mathrm{adj}(A) = C^{\mathrm{T}}$
- *For non-singular matrix A, we have:* $A^{-1} = \frac{\mathrm{adj}(A)}{\det(A)} = \frac{C^{\mathrm{T}}}{\det(A)}$

Now we're ready to show the derivative of determinant. Note that determinant is just a scalar function, so all techniques discussed above is applicable. We first write the derivative element by element. Expanding determinant on the i-th row, we have:

$$\frac{\partial \det(A)}{\partial A_{ij}} = \frac{\partial(\sum_j A_{ij} C_{ij})}{\partial A_{ij}} = C_{ij} \tag{52}$$

First equality is from determinant definition and second equality is by the observation that only coefficient of A_{ij} is left. Grouping all elements using definition(1), we have:

$$\frac{\partial \det(A)}{\partial A} = C = \mathrm{adj}(A)^{\mathrm{T}} \tag{53}$$

If A is non-singular, we have:

$$\frac{\partial \det(A)}{\partial A} = (\det(A) A^{-1})^{\mathrm{T}} = \det(A)(A^{-1})^{\mathrm{T}} \tag{54}$$

Next, we use theorem(6) to give the differential relationship:

$$\begin{aligned}
\mathrm{d}\det(A) &= \mathrm{Tr}\left[(\frac{\partial \det(A)}{\partial A})^{\mathrm{T}}\mathrm{d}A\right] && (55)\\
&= \mathrm{Tr}\left[(\det(A)(A^{-1})^{\mathrm{T}})^{\mathrm{T}}\mathrm{d}A\right] && (56)\\
&= \mathrm{Tr}\left[\det(A)A^{-1}\mathrm{d}A\right] && (57)
\end{aligned}$$

In many practical problem, the log determinant is more widely used:

$$\frac{\partial \ln\det(A)}{\partial A} = \frac{1}{\det(A)}\frac{\partial \det(A)}{\partial A} = (A^{-1})^{\mathrm{T}} \tag{58}$$

The first equality comes from chain rule of ordinary calculus($\ln\det(A)$ and $\det(A)$ are both scalars). Similarly, we derive for differential:

$$\mathrm{d}\ln\det(A) = \mathrm{Tr}\left[A^{-1}\mathrm{d}A\right] \tag{59}$$

2.9 Vector Function and Vector Variable

The above sections show how to deal with scalar functions. In order to deal with vector function, we should restrict our attention to vector variables. It's no surprising that the tractable forms in matrix calculus is so scarse. If we allow matrix functions and matrix variables, given the fact that fully specification of all partial derivatives calls for a tensor, it will be difficult to visualize the result on a paper. An alternative is to stretch functions and variables such that they appear as vectors.

An annoying fact of matrix calculus is that, when you try to find reference materials, there are always two kinds of people. One group calculates as the transpose of another. Many online resources are not coherent, which mislead people.

We borrow the following definitions of Hessian matrix[7] and Jacobian matrix[8] from Wikipedia:

$$H(f) = \begin{bmatrix}
\frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1\,\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1\,\partial x_n} \\
\frac{\partial^2 f}{\partial x_2\,\partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2\,\partial x_n} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 f}{\partial x_n\,\partial x_1} & \frac{\partial^2 f}{\partial x_n\,\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2}
\end{bmatrix} \tag{60}$$

$$J = \begin{bmatrix}
\frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\
\vdots & \ddots & \vdots \\
\frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n}
\end{bmatrix} \tag{61}$$

Note two things about second order derivative:

- By writting the abbreviation $\frac{\partial^2 f}{\partial x_2 \, \partial x_1}$, people mean $\frac{\partial}{\partial x_2}\left(\frac{\partial f}{\partial x_1}\right)$ by convention. That is, first take derivative with respect to x_1 and then take derivative with respect to x_2.

- The Hessian matrix can be regarded as first compute the $\frac{\partial f}{\partial x}$ (using definition(1) to organize), and then compute a vector-function-to-vector-variable derivative treating $\frac{\partial f}{\partial x}$ as the function.

Bearing this in mind, we find Hessian matrix and Jacobian matrix actually have contradictory notion of organization. In order to be coherent in this document, we adopt the Hessian style. That is, each row corresponds to a variable, and each column corresponds to a function. To be concrete:

Definition 4. *For a vector function $f = [f_1, f_2, \ldots, f_n]^{\mathrm{T}}$, and $f_i = f_i(x)$ where $x = [x_1, x_2, \ldots, x_m]^{\mathrm{T}}$, we have the following definition:*

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_m} & \frac{\partial f_2}{\partial x_m} & \cdots & \frac{\partial f_n}{\partial x_m} \end{bmatrix} \tag{62}$$

Example 6. *According to definition(4), we revisit the definition of Hessian and Jacobian.*

Given twice differentiable function $f(x)$, the Hessian is defined as:

$$H(f) = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x^{\mathrm{T}}}\right) \tag{63}$$

Given two variables x and y, if we want to transform x into y, the Jacobian is defined as:

$$J = \det\left((\frac{\partial x}{\partial y})^{\mathrm{T}} \right) = \det\left(\frac{\partial x}{\partial y}\right) \tag{64}$$

Note "Jacobian" is the shorthand name of "Jacobian determinant", which is the determinant of "Jacobian matrix". Due to the transpose invariance of determinant, the second equality shows that it does not matter which organization method we use if we only want to do compute the Jacobian, rather than Jacobin matrix. However, if we're to write out the Jacobin matrix, this may be a pitfall depending on what organization of vector-to-vector derivative we define.

2.10 Vector Function Differential

In the previous sections, we relate matrix differential with matrix derivatvie using theorem(6). This theorem bridges the two quantities using trace. Thus we can handle the problem using preferred form intechangeably. (As is seen: calculating derivative of determinant is more direct; calculating differential of inverse is tractable.)

In this section, we provide another theorem to relate vector function differential to vector-to-vector derivatives. Amazingly, it takes a cleaner form.

Theorem 9. *Consider the definition(4), we have:* $\mathrm{d}f = (\frac{\partial f}{\partial x})^{\mathrm{T}}\mathrm{d}x$

Proof. Apparently, $\mathrm{d}f$ has n components, so we prove element by element. Consider the j-th component:

$$\begin{aligned}
\mathrm{LHS}_j &= \mathrm{d}f_j && (65)\\
&= \sum_{i=1}^{m} \frac{\partial f_j}{\partial x_i}\mathrm{d}x_i && (66)\\
\mathrm{RHS}_j &= ((\frac{\partial f}{\partial x})^{\mathrm{T}}x)_j && (67)\\
&= \sum_{i=1}^{m} (\frac{\partial f}{\partial x})^{\mathrm{T}}_{ji}x_i && (68)\\
&= \sum_{i=1}^{m} (\frac{\partial f}{\partial x})_{ij}x_i && (69)\\
&= \sum_{i=1}^{m} (\frac{\partial f}{\partial x})_{ij}x_i && (70)\\
&= \sum_{i=1}^{m} (\frac{\partial f_j}{\partial x_i})x_i = \mathrm{LHS}_j && (71)
\end{aligned}$$

□

Note that the trace operator is gone compared with theorem(6) due to the nice way of defining matrix vector multiplication. We can have a similar schema of handling vector-to-vector derivatives using this scheme. We don't bother to list the schema again. Instead, we provide an example.

Example 7. *Consider the variable transformation:*$x = \sigma\Lambda^{-0.5}W^{\mathrm{T}}\xi$, *where* σ *is a real value,* Λ *is full rank diagonal matrix, and* W *is orthonormal square matrix (namely*$WW^{\mathrm{T}} = W^{\mathrm{T}}W = I$*). Compute the absolute value of Jacobian.*

First, we want to find $\frac{\partial x}{\partial \xi}$*. This can be easily done by computing the differential:*

$$\mathrm{d}x = \mathrm{d}(\sigma\Lambda^{-0.5}W^{\mathrm{T}}\xi) = \sigma\Lambda^{-0.5}W^{\mathrm{T}}\mathrm{d}\xi \tag{72}$$

Applying theorem(9), we have:

$$\frac{\partial x}{\partial \xi} = (\sigma\Lambda^{-0.5}W^{\mathrm{T}})^{\mathrm{T}} \tag{73}$$

Thus,

$$J_m = (\frac{\partial x}{\partial \xi})^{\mathrm{T}} = ((\sigma\Lambda^{-0.5}W^{\mathrm{T}})^{\mathrm{T}})^{\mathrm{T}} = \sigma\Lambda^{-0.5}W^{\mathrm{T}} \tag{74}$$

where J_m *is the Jacobian matrix and* $J = \det(J_m)$*. Then we use some property of determinant to calculate the absolute value of Jacobian:*

$$\begin{aligned}
|J| &= |\det(J_m)| && (75)\\
&= \sqrt{|\det(J_m)|\,|\det(J_m)|} && (76)\\
&= \sqrt{|\det(J_m^{\mathrm{T}})|\,|\det(J_m)|} && (77)\\
&= \sqrt{|\det(J_m^{\mathrm{T}}J_m)|} && (78)\\
&= \sqrt{|\det(J_mJ_m^{\mathrm{T}})|} && (79)\\
&= \sqrt{|\det(W\Lambda^{-0.5}\sigma\sigma\Lambda^{-0.5}W^{\mathrm{T}})|} && (80)\\
&= \sqrt{|\det(\sigma^2W\Lambda^{-1}W^{\mathrm{T}})|} && (81)
\end{aligned}$$

If the dimension of x *and* ξ *is* d *and we define* $\Sigma = W\Lambda W^{\mathrm{T}}$*. A nice result is calculated:*

$$|J| = \sigma^d \det(\Sigma)^{-1/2} \tag{82}$$

which we'll see an application of generalizing the multivariate Gaussian distribution.

Theorem 10. *When* f *and* x *are of the same size, with definition(4):*

$$(\frac{\partial f}{\partial x})^{-1} = \frac{\partial x}{\partial f} \tag{83}$$

Proof. Using theorem(9):

$$\begin{aligned}
\mathrm{d}f &= (\frac{\partial f}{\partial x})^{\mathrm{T}}\mathrm{d}x && (84)\\
((\frac{\partial f}{\partial x})^{\mathrm{T}})^{-1}\mathrm{d}f &= \mathrm{d}x && (85)\\
\mathrm{d}x &= ((\frac{\partial f}{\partial x})^{-1})^{\mathrm{T}}\mathrm{d}f && (86)\\
& && (87)
\end{aligned}$$

Using theorem(9 in the reverse direction:

$$\frac{\partial x}{\partial f} = (\frac{\partial f}{\partial x})^{-1} \tag{88}$$

□

This result is consistent with scalar derivative. It is useful in manipulating variable substitutions(the inverse of Jacobian is the Jacobian of reverse substitution).

2.11 Chain Rule

In the schemas we conclude above, differential is convenient in many problems. For derivative, the nice aspect is that we have chain rule, which is an analogy to the chain rule in ordinary calculus. However, one should be very careful applying this chain rule, for the multiplication of matrix requires dimension agreement.

Theorem 11. *Suppose we have n column vectors $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$, each is of length $l_1, l_2, \ldots, l_n$. We know $x^{(i)}$ is a function of $x^{(i-1)}$, for all $i = 2, 3, \ldots, n$. The following relationship holds:*

$$\frac{\partial x^{(n)}}{\partial x^{(1)}} = \frac{\partial x^{(2)}}{\partial x^{(1)}} \frac{\partial x^{(3)}}{\partial x^{(2)}} \cdots \frac{\partial x^{(n)}}{\partial x^{(n-1)}} \tag{89}$$

Proof. Under definition(4), theorem(9) holds. We apply this theorem to each consecutive vectors:

$$\mathrm{d}x^{(2)} = (\frac{\partial x^{(2)}}{\partial x^{(1)}})^{\mathrm{T}} \mathrm{d}x^{(1)} \tag{90}$$

$$\mathrm{d}x^{(3)} = (\frac{\partial x^{(3)}}{\partial x^{(2)}})^{\mathrm{T}} \mathrm{d}x^{(2)} \tag{91}$$

$$\cdots \tag{92}$$

$$\mathrm{d}x^{(n)} = (\frac{\partial x^{(n)}}{\partial x^{(n-1)}})^{\mathrm{T}} \mathrm{d}x^{(n-1)} \tag{93}$$

Plug previous one into next one, we have:

$$\mathrm{d}x^{(n)} = (\frac{\partial x^{(n)}}{\partial x^{(n-1)}})^{\mathrm{T}} \cdots (\frac{\partial x^{(3)}}{\partial x^{(2)}})^{\mathrm{T}} (\frac{\partial x^{(2)}}{\partial x^{(1)}})^{\mathrm{T}} \mathrm{d}x^{(1)} \tag{94}$$

$$= (\frac{\partial x^{(2)}}{\partial x^{(1)}} \frac{\partial x^{(3)}}{\partial x^{(2)}} \cdots \frac{\partial x^{(n)}}{\partial x^{(n-1)}})^{\mathrm{T}} \mathrm{d}x^{(1)} \tag{95}$$

Applying theorem(9) again in the reverse direction, we have:

$$\frac{\partial x^{(n)}}{\partial x^{(1)}} = \frac{\partial x^{(2)}}{\partial x^{(1)}} \frac{\partial x^{(3)}}{\partial x^{(2)}} \cdots \frac{\partial x^{(n)}}{\partial x^{(n-1)}} \tag{96}$$

□

Proposition 12. *Consider a chain of: x a scalar, y a column vector, z a scalar. $z = z(y), y_i = y_i(x), i = 1, 2 \dots, n$. Apply the chain rule, we have*

$$\frac{\partial z}{\partial x} = \frac{\partial y}{\partial x}\frac{\partial z}{\partial y} = \sum_{i=1}^{n} \frac{\partial y_i}{\partial x}\frac{\partial z}{\partial y_i} \tag{97}$$

Now we explain the intuition behind. x, z are both scalar, so we're basically calculating the derivative in ordinary calculus. Besides, we have a group of "bridge variables", y_i. $\frac{\partial y_i}{\partial x}\frac{\partial z}{\partial y_i}$ is just the result of applying scalar chain rule on the chain:$x \to y_i \to z$. The separate results of different bridge variables are additive! To see why I make this proposition, interested readers can refer to the comments in the corresponding LaTeX source file.

Example 8. *Show the derivative of $(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)$ to μ (for symmetric Σ^{-1}).*

$$\begin{aligned}
& \frac{\partial[(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)]}{\partial\mu} && (98)\\
&= \frac{\partial[x-\mu]}{\partial\mu}\frac{\partial[(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)]}{\partial[x-\mu]} && (99)\\
\textit{(example(4))} \quad &= \frac{\partial[x-\mu]}{\partial\mu}\, 2\,\Sigma^{-1}(x-\mu) && (100)\\
(\mathrm{d}[x-\mu] = -I\,\mathrm{d}\mu) \quad &= -I\,2\,\Sigma^{-1}(x-\mu) && (101)\\
&= -2\Sigma^{-1}(x-\mu) && (102)
\end{aligned}$$

3 Application

3.1 The 2nd Induced Norm of Matrix

The induced norm of matrix is defined as [4]:

$$||A||_p = \max_x \frac{||Ax||_p}{||x||_p} \tag{103}$$

where $||\bullet||_p$ denotes the p-norm of vectors. Now we solve for $p = 2$. (By default, $||\bullet||$ means $||\bullet||_2$)

The problem can be restated as:

$$||A||^2 = \max_x \frac{||Ax||^2}{||x||^2} \tag{104}$$

since all quantities involved are non-negative. Then we consider a scaling of vector $x' = tx$, thus:

$$||A||^2 = \max_{x'} \frac{||Ax'||^2}{||x'||^2} = \max_x \frac{||tAx||^2}{||tx||^2} = \max_x \frac{t^2||Ax||^2}{t^2||x||^2} = \max_x \frac{||Ax||^2}{||x||^2} \tag{105}$$

This shows the invariance under scaling. Now we can restrict our attention to those x with $||x|| = 1$, and reach the following formulation:

$$\begin{aligned} \text{Maximize} \quad & f(x) = ||Ax||^2 && (106) \\ s.t. \quad & ||x||^2 = 1 && (107) \end{aligned}$$

The standard way to handle this constrained optimization is using Lagrange relaxation:

$$L(x) = f(x) - \lambda(||x||^2 - 1) \tag{108}$$

Then we apply the general schema of handling scalar function on $L(x)$. First take differential:

$$\begin{aligned} \mathrm{d}L(x) &= \mathrm{d}\mathrm{Tr}\left[L(x)\right] && (109) \\ &= \mathrm{Tr}\left[\mathrm{d}(L(x))\right] && (110) \\ &= \mathrm{Tr}\left[\mathrm{d}(x^{\mathrm{T}}A^{\mathrm{T}}Ax - \lambda(x^{\mathrm{T}}x - 1))\right] && (111) \\ &= \mathrm{Tr}\left[2x^{\mathrm{T}}A^{\mathrm{T}}A\mathrm{d}x - \lambda(2x^{\mathrm{T}}\mathrm{d}x)\right] && (112) \\ &= \mathrm{Tr}\left[(2x^{\mathrm{T}}A^{\mathrm{T}}A - 2\lambda x^{\mathrm{T}})\mathrm{d}x\right] && (113) \end{aligned}$$

Next write out derivative:

$$\frac{\partial L}{\partial x} = 2A^{\mathrm{T}}Ax - 2\lambda x \tag{114}$$

Let $\dfrac{\partial L}{\partial x} = 0$, we have:

$$(A^{\mathrm{T}}A)x = \lambda x \tag{115}$$

That means x is the eigen vector of $(A^{\mathrm{T}}A)$ (normalized to $||x|| = 1$), and λ is corresponding eigen value. We plug this result back to objective function:

$$f(x) = x^{\mathrm{T}}(A^{\mathrm{T}}Ax) = x^{\mathrm{T}}(\lambda x) = \lambda \tag{116}$$

which means, to maximize $f(x)$, we should pick the maximum eigen value:

$$||A||^2 = \max f(x) = \lambda_{\max}(A^{\mathrm{T}}A) \tag{117}$$

That is:

$$||A|| = \sqrt{\lambda_{\max}(A^{\mathrm{T}}A)} = \sigma_{\max}(A) \tag{118}$$

where $\sigma_{\max}$ denotes the maximum singular value. If A is real symmetric, $\sigma_{\max}(A) = \lambda_{\max}(A)$.

Now we consider a real symmetric A and check whether:

$$\lambda^2_{\max}(A) = \max_x \frac{||Ax||^2}{||x||^2} = \max_x \frac{x^{\mathrm{T}}A^{\mathrm{T}}Ax}{x^{\mathrm{T}}x} \tag{119}$$

Proof. Since $A^{\mathrm{T}}A$ is real symmetric, it has an orthnormal basis formed by n eigen vectors, $v_1, v_2, \ldots, v_n$, with eigen values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. We can write $x = \sum_i c_i v_i$, where $c_i =< x, v_i >$. Then,

$$\begin{aligned}
& & \frac{x^{\mathrm{T}}A^{\mathrm{T}}Ax}{x^{\mathrm{T}}x} & \qquad (120) \\
(v_k \text{ is an orthnormal set}) & = & \frac{\sum_i \lambda_i c_i^2}{\sum_i c_i^2} & \qquad (121) \\
& \leq & \frac{\sum_i \lambda_1 c_i^2}{\sum_i c_i^2} & \qquad (122) \\
& = & \lambda_1 & \qquad (123)
\end{aligned}$$

Now we have proved an upper bound for $||A||^2$. We show this bound is achievable by assigning $x = v_1$. □

3.2 General Multivaraite Gaussian Distribution

The first time I came across multivariate Gaussian distribution is in my sophomore year. However, at that time, the multivariate version is optional, and the text book only gives us the formula rather than explaining any intuition behind. I had trouble remembering the formula, since I don't know why it is the case.

During the Machine Learning course[12] I take recently, the rationale becomes clear to me. We'll start from the basic notion, generalize it to multivariate isotropic Gaussian, and then use matrix calculus to derive the most general version. The following content is adapted from the corresponding course notes and homework exercises.

We start from the univariate Gaussian[9]:

$$G(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (124)$$

The intuition behind is:

- Use one centroid to represent a set of samples coming from the same Gaussian mixture, which is denoted by μ.
- Allow the existence of error. We express the reverse notion of "error" by "likelihood". We want the density function describing the distribution that the farther away, the less likely a sample is drawn from the mixture. We want this likelihood to decrease in a accelerated manner. Negative exponential is one proper candidate, say $\exp\{-D(x, \mu)\}$, where D measures the distance between sample and centroid.
- To fit human's intuition, the "best" choice of distance measure is Euclidean distance, since we live in this Euclidean space. So far, we have $\exp\{-(x - \mu)^2\}$

- How fast the likelihood decreases should be controlled by a parameter, say $\exp\{\frac{-(x-\mu)^2}{\sigma}\}$. σ can also be thought to control the uncertainty.

Now we already get Gaussian distribution. The division by 2 in the exponent is only to simplify deductions. Writting σ^2 instead of σ is to align with some statistic quantities. The rest term is basically used to do normalization, which can be derived by extending the 1-D Gaussian integral to a 2-D area integral using Fubini's thorem[11]. Interested reader can refer to [10].

Now we're ready to extend the univariate version to multivariate version. The above four characteristics appear as simple interpretation to everyone who learned Gaussian distribution before. However, they're the real "axioms" behind. Now consider an isotropic Gaussian distribution and we start with perpendicular axes. The uncertainty along each axis is the same, so the distance measure is now $||x-\mu||_2^2 = (x-\mu)^{\mathrm{T}}(x-\mu)$. The exponent is $\exp\{-\frac{1}{2\sigma^2}(x-\mu)^{\mathrm{T}}(x-\mu)\}$. Integrating over the volume of x can be done by transforming it into iterated form using Fubini's theorem. Then we simply apply the method that we deal with univariate Gaussian integral. Now we have multivariate isotropic Gaussian distribution:

$$G(x|\mu,\sigma^2) = \frac{1}{(2\pi)^{d/2}\sigma^d}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^{\mathrm{T}}(x-\mu)\} \tag{125}$$

where d is the dimension of x, μ.

We are just one step away from a general Gaussian. Suppose we have a general Gaussian, whose covariance is not isotropic nor does it all along the standard Euclidean axes. Denote the sample from this Gaussian by ξ. We can first apply a rotation on ξ to bring it back to the standard axes, which can be done by left multiplying an orthongonal matrix W^T. Then we scale each component of $W\xi$ by different ratio to make them isotropic, which can be done by left multiplying a diagonal matrix $\Lambda^{-0.5}$. The exponent -0.5 is simply to make later discussion convenient. Finally, we multiply it by σ to be able to control the uncertainty at each direction again. The transform is given by $x = \sigma\Lambda^{-0.5}W^{\mathrm{T}}\xi$. Plugging this back to eqn(125), we derive the following exponent:

$$\begin{aligned}
& \exp\{-\frac{1}{2\sigma^2}(x-\mu)^{\mathrm{T}}(x-\mu)\} && (126)\\
= \; & \exp\{-\frac{1}{2\sigma^2}(\sigma\Lambda^{-0.5}W^{\mathrm{T}}\xi-\mu)^{\mathrm{T}}(\sigma\Lambda^{-0.5}W^{\mathrm{T}}\xi-\mu)\} && (127)\\
= \; & \exp\{-\frac{1}{2\sigma^2}(\xi-\mu_\xi)^{\mathrm{T}}\sigma^2 W\Lambda^{-1}W^{\mathrm{T}}(\xi-\mu_\xi)\} && (128)\\
= \; & \exp\{-\frac{1}{2}(\xi-\mu_\xi)^{\mathrm{T}}\Sigma^{-1}(\xi-\mu_\xi)\} && (129)
\end{aligned}$$

where we let $\mu_\xi = \mathbb{E}\xi$, and we plugged in the following result:

$$\mu = \mathbb{E}x = \mathbb{E}[\sigma\Lambda^{-0.5}W^{\mathrm{T}}\xi] = \sigma\Lambda^{-0.5}W^{\mathrm{T}}\mathbb{E}\xi = \sigma\Lambda^{-0.5}W^{\mathrm{T}}\mu_\xi \tag{130}$$

Now we only need to find the normalization factor to make it a probability distribution. Note eqn(125) integrates to 1, which is:

$$\int G(x|\mu,\sigma^2)\mathrm{d}x = \int \frac{1}{(2\pi)^{d/2}\sigma^d}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^{\mathrm{T}}(x-\mu)\}\mathrm{d}x = 1 \quad (131)$$

Transforming variable from x to ξ causes a scaling by absolute Jacobian, which we already calculated in example(7). That is:

$$\int \frac{1}{(2\pi)^{d/2}\sigma^d}\exp\{-\frac{1}{2}(\xi-\mu_\xi)^{\mathrm{T}}\Sigma^{-1}(\xi-\mu_\xi)\}|J|\mathrm{d}\xi = 1 \quad (132)$$

$$\int \frac{\sigma^d \det(\Sigma)^{-1/2}}{(2\pi)^{d/2}\sigma^d}\exp\{-\frac{1}{2}(\xi-\mu_\xi)^{\mathrm{T}}\Sigma^{-1}(\xi-\mu_\xi)\}\mathrm{d}\xi = 1 \quad (133)$$

$$\int \frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}}\exp\{-\frac{1}{2}(\xi-\mu_\xi)^{\mathrm{T}}\Sigma^{-1}(\xi-\mu_\xi)\}\mathrm{d}\xi = 1 \quad (134)$$

The term inside integral is just the general Gaussian density we want to find:

$$G(x|\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}}\exp\{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\} \quad (135)$$

3.3 Maximum Likelihood Estimation of Gaussian

Given N samples, $x_t, t = 1, 2, \ldots, N$, independently indentically distributed that are drawn from the following Gaussian distribution:

$$G(x_t|\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}}\exp\{-\frac{1}{2}(x_t-\mu)^{\mathrm{T}}\Sigma^{-1}(x_t-\mu)\} \quad (136)$$

solve the paramters $\theta = \{\mu, \Sigma\}$, that maximize:

$$p(X|\theta) = \prod_{t=1}^{N} G(x_t|\theta) \quad (137)$$

It's more conveninent to handle the log likelihood as defined below:

$$L = \ln p(X|\theta) = \sum_{t=1}^{N} \ln G(x_t|\theta) \quad (138)$$

We write each term of L out to facilitate further processing:

$$L = -\frac{Nd}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_t (x_t-\mu)^{\mathrm{T}}\Sigma^{-1}(x_t-\mu) \quad (139)$$

Taking derivative of μ, the first two terms are gone and the third term is already handled by example(8):

$$\frac{\partial L}{\partial \mu} = \sum_t \Sigma^{-1}(x_t-\mu) \quad (140)$$

Let $\dfrac{\partial L}{\partial \mu} = 0$, we solve for $\mu = \frac{1}{N}\sum_t x_t$. Then take derivative of Σ. It eaiser to be handled using our trace schema:

$$\mathrm{d}L \quad = \quad \mathrm{d}[-\frac{N}{2}\ln|\Sigma|] + \mathrm{d}[-\frac{1}{2}\sum_t (x_t-\mu)^{\mathrm{T}}\Sigma^{-1}(x_t-\mu)] \tag{141}$$

The first term is:

$$\begin{aligned}
& \mathrm{d}[-\frac{N}{2}\ln|\Sigma|] && (142)\\
= \quad & -\frac{N}{2}\mathrm{d}\ln|\Sigma| && (143)\\
(\text{section}(2.8)) \quad = \quad & -\frac{N}{2}\mathrm{Tr}\left[\Sigma^{-1}\mathrm{d}\Sigma\right] && (144)
\end{aligned}$$

The second term is:

$$\begin{aligned}
& \mathrm{d}[-\frac{1}{2}\sum_t (x_t-\mu)^{\mathrm{T}}\Sigma^{-1}(x_t-\mu)] && (145)\\
= \quad & -\frac{1}{2}\mathrm{d}\mathrm{Tr}\left[\sum_t (x_t-\mu)^{\mathrm{T}}\Sigma^{-1}(x_t-\mu)\right] && (146)\\
= \quad & -\frac{1}{2}\mathrm{d}\mathrm{Tr}\left[\sum_t (x_t-\mu)(x_t-\mu)^{\mathrm{T}}\Sigma^{-1}\right] && (147)\\
(\text{example}(5)) \quad = \quad & -\frac{1}{2}\mathrm{Tr}\left[[\sum_t (x_t-\mu)(x_t-\mu)^{\mathrm{T}}](-\Sigma^{-1}\mathrm{d}\Sigma\Sigma^{-1})\right] && (148)\\
= \quad & \frac{1}{2}\mathrm{Tr}\left[\Sigma^{-1}[\sum_t (x_t-\mu)(x_t-\mu)^{\mathrm{T}}]\Sigma^{-1}\mathrm{d}\Sigma\right] && (149)
\end{aligned}$$

Then we have:

$$\frac{\partial L}{\partial \Sigma} = -\frac{N}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}[\sum_t (x_t-\mu)(x_t-\mu)^{\mathrm{T}}]\Sigma^{-1} \tag{150}$$

Let $\dfrac{\partial L}{\partial \Sigma} = 0$, we solve for $\Sigma = \frac{1}{N}\sum_t (x_t-\mu)(x_t-\mu)^{\mathrm{T}}$.

3.4 Least Square Error Inference: a Comparison

This application is rather simple. The purpose is to compare the derivation with and without matrix calculus.

3.4.1 Problem Definition

Here we consider a simple version: [1]

[1] Sid Jaggi, Dec 2011, Final Exam Q1 of CUHK ERG2013

Given the linear system with noise $y = Ax + n$, where x is the input, y is the output and n is the noise term. A as the system parameter is a known $p \times q$ matrix. Now we have one observation of output, $\hat{y}$. We want to infer the "most possible" corresponding input $\hat{x}$ defined by the following formula: [2]

$$\hat{x} = \underset{x:\hat{y}=Ax+n}{\arg\min} \; ||n||^2 \tag{151}$$

3.4.2 Ordinary Formulation and Derivation

First, we write the error function explicitly:

$$\begin{aligned}
f(x) &= ||\hat{y} - Ax||^2 && (152)\\
&= \sum_{i=1}^{p} (\hat{y}_i - (Ax)_i)^2 && (153)\\
&= \sum_{i=1}^{p} (\hat{y}_i - \sum_{j=1}^{q} A_{ij}x_j)^2 && (154)
\end{aligned}$$

This is the quadratic form of all $x_k, k = 1, 2 \dots q$. We take derivative of x_k:

$$\begin{aligned}
\frac{\partial f}{\partial x_k} &= \sum_{i=1}^{p} [2(\hat{y}_i - \sum_{j=1}^{q} A_{ij}x_j)(-A_{ik})] && (155)\\
&= -2\sum_{i=1}^{p} [(\hat{y}_i - (Ax)_i)A_{ik}] && (156)\\
&= -2 < (\hat{y} - (Ax)), A_k > && (157)\\
&= -2A_k^{\mathrm{T}}(\hat{y} - Ax) && (158)
\end{aligned}$$

where A_k is the k-th column of A and $< ., . >$ denotes the inner product of two vectors. We let all $\frac{\partial f}{\partial x_k} = 0$ and put the q equations together to get the matrix form:

$$\begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_q} \end{bmatrix} = \begin{bmatrix} -2A_1^{\mathrm{T}}(\hat{y} - Ax) \\ -2A_2^{\mathrm{T}}(\hat{y} - Ax) \\ \vdots \\ -2A_q^{\mathrm{T}}(\hat{y} - Ax) \end{bmatrix} = -2A^{\mathrm{T}}(\hat{y} - Ax) = 0 \tag{159}$$

[2] Assuming Gaussian noise, the maximum likelihood inference results in the same expression of least squares. For simplicity, we ignore the argument of this conclusion. Readers can refer to the previous example to see the relation between least squares and Gaussian distribution.

We solve for:

$$x = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}\hat{y} \tag{160}$$

when $(A^{\mathrm{T}}A)$ is invertible.

3.4.3 Matrix Formulation and Derivation

The error function can be written as:

$$\begin{aligned} f(x) &= ||\hat{y} - Ax||^2 && (161) \\ &= (\hat{y} - Ax)^{\mathrm{T}}(\hat{y} - Ax) && (162) \end{aligned}$$

We apply the trace schema to obtain the differential:

$$\begin{aligned} \mathrm{d}f &= \mathrm{dTr}\left[(\hat{y} - Ax)^{\mathrm{T}}(\hat{y} - Ax)\right] && (163) \\ &= \mathrm{Tr}\left[\mathrm{d}(\hat{y} - Ax)^{\mathrm{T}}(\hat{y} - Ax) + (\hat{y} - Ax)^{\mathrm{T}}\mathrm{d}(\hat{y} - Ax)\right] && (164) \\ &= -\mathrm{Tr}\left[(\mathrm{d}x)^{\mathrm{T}}A^{\mathrm{T}}(\hat{y} - Ax) + (\hat{y} - Ax)^{\mathrm{T}}A\mathrm{d}x\right] && (165) \\ &= -\mathrm{Tr}\left[(\hat{y} - Ax)^{\mathrm{T}}A\mathrm{d}x + (\hat{y} - Ax)^{\mathrm{T}}A\mathrm{d}x\right] && (166) \\ &= -2\mathrm{Tr}\left[(\hat{y} - Ax)^{\mathrm{T}}A\mathrm{d}x\right] && (167) \end{aligned}$$

The derivative is:

$$\frac{\partial f}{\partial x} = -2A^{\mathrm{T}}(\hat{y} - Ax) \tag{168}$$

We let $\dfrac{\partial f}{\partial x} = 0$ and get the same result.

3.4.4 Remarks

Without matrix calculus, we can still achieve the goal. Every step is straightforward using ordinary calculus, except for the last one where we organize everything into the matrix form.

With matrix calculus, we utilize those well established matrix operators, like matrix multiplication. Thus the notation is sharply simplified. Comparing the two derivations, we find the one with matrix calculus is absent of annoying summations. This should be very advantegeous in more complex expressions.

This example aims to disclose the essence of matrix calculus. Whenever we need new rules, it's always workable in traditional calculus way.

4 Cheat Sheet

4.1 Definition

For scalar function f and matrix variable x, the derivative is:

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{12}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \frac{\partial f}{\partial x_{21}} & \frac{\partial f}{\partial x_{22}} & \cdots & \frac{\partial f}{\partial x_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \frac{\partial f}{\partial x_{m2}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix} \tag{169}$$

For column vector function f and column vector variable x, the derivative is:

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_m} & \frac{\partial f_2}{\partial x_m} & \cdots & \frac{\partial f_n}{\partial x_m} \end{bmatrix} \tag{170}$$

For $m \times n$ matrix A, the differential is:

$$\mathrm{d}A = \begin{bmatrix} \mathrm{d}A_{11} & \mathrm{d}A_{12} & \cdots & \mathrm{d}A_{1n} \\ \mathrm{d}A_{21} & \mathrm{d}A_{22} & \cdots & \mathrm{d}A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{d}A_{m1} & \mathrm{d}A_{m2} & \cdots & \mathrm{d}A_{mn} \end{bmatrix} \tag{171}$$

4.2 Schema for Scalar Function

1. $\mathrm{d}f = \mathrm{d}\mathrm{Tr}\left[f\right] = \mathrm{Tr}\left[\mathrm{d}f\right]$

2. Apply trace properties(see theorem(3)) and matrix differential properties(see theorem(7)) to get the following form:

$$\mathrm{d}f = \mathrm{Tr}\left[A^{\mathrm{T}}x\right] \tag{172}$$

3. Apply theorem(6) to get:

$$\frac{\partial f}{\partial x} = A \tag{173}$$

4.3 Schema for Vector Function

1. Apply (possibly) trace properties(see theorem(3)) and matrix differential properties(see theorem(7)) to get the following form:

$$\mathrm{d}f = A^{\mathrm{T}} x \tag{174}$$

2. Apply theorem(9) to get:

$$\frac{\partial f}{\partial x} = A \tag{175}$$

4.4 Properties

Trace properties and differential properties are collected in one list as follows:

1. $\mathrm{Tr}\left[A + B\right] = \mathrm{Tr}\left[A\right] + \mathrm{Tr}\left[B\right]$
2. $\mathrm{Tr}\left[cA\right] = c\mathrm{Tr}\left[A\right]$
3. $\mathrm{Tr}\left[AB\right] = \mathrm{Tr}\left[BA\right]$
4. $\mathrm{Tr}\left[A_1 A_2 \dots A_n\right] = \mathrm{Tr}\left[A_n A_1 \dots A_{n-1}\right]$
5. $\mathrm{Tr}\left[A^{\mathrm{T}} B\right] = \sum_i \sum_j A_{ij} B_{ij}$
6. $\mathrm{Tr}\left[A\right] = \mathrm{Tr}\left[A^{\mathrm{T}}\right]$
7. $\mathrm{d}(cA) = c\mathrm{d}A$
8. $\mathrm{d}(A + B) = \mathrm{d}A + \mathrm{d}B$
9. $\mathrm{d}(AB) = \mathrm{d}AB + A\mathrm{d}B$
10. $\mathrm{d}\mathrm{Tr}\left[A\right] = \mathrm{Tr}\left[\mathrm{d}A\right]$

4.5 Frequently Used Formula

For matrix A, column vector x:

- $\dfrac{\partial \mathrm{Tr}\left[A\right]}{\partial A} = I$
- $\dfrac{\partial x^{\mathrm{T}} A x}{\partial x} = Ax + A^{\mathrm{T}} x$
- $\dfrac{\partial x^{\mathrm{T}} x}{\partial x} = 2x$
- $\dfrac{\partial Ax}{\partial x} = A^{\mathrm{T}}$

- $\dfrac{\partial x^{\mathrm{T}}Ax}{\partial xx^{\mathrm{T}}} = \dfrac{\partial}{\partial x}(\dfrac{\partial x^{\mathrm{T}}Ax}{\partial x}) = A^{\mathrm{T}} + A$
- $\mathrm{d}\mathrm{Tr}\left[A\right] = \mathrm{Tr}\left[I\mathrm{d}A\right]$
- $\mathrm{Tr}\left[\mathrm{d}(x^{\mathrm{T}}Ax)\right] = \mathrm{Tr}\left[(x^{\mathrm{T}}A^{\mathrm{T}} + x^{\mathrm{T}}A)\mathrm{d}x\right]$
- $\mathrm{Tr}\left[\mathrm{d}(x^{\mathrm{T}}x)\right] = \mathrm{Tr}\left[2x^{\mathrm{T}}\mathrm{d}x\right]$

When $A = A^{\mathrm{T}}$ those formulae are simplified to:

- $\dfrac{\partial x^{\mathrm{T}}Ax}{\partial x} = 2Ax$
- $\dfrac{\partial x^{\mathrm{T}}Ax}{\partial xx^{\mathrm{T}}} = 2A$
- $\mathrm{Tr}\left[\mathrm{d}(x^{\mathrm{T}}Ax)\right] = \mathrm{Tr}\left[2(x^{\mathrm{T}}A)\mathrm{d}x\right]$

The inverse: (for non-singular A)

- $\mathrm{d}(X^{-1}) = -X^{-1}\mathrm{d}XX^{-1}$

Note we only have inverse formula for differential. Unless X is a scalar, we have difficulty organize the elements of $\dfrac{\partial(X^{-1})}{\partial X}$, which should be a 4-D tensor.

The determinant series: (for non-singular A)

- $\dfrac{\partial \det(A)}{\partial A} = C$ (C is cofactor matrix)
- $\dfrac{\partial \det(A)}{\partial A} = (\mathrm{adj}(A))^{\mathrm{T}}$ ($\mathrm{adj}(A)$ is adjugate matrix)
- $\dfrac{\partial \det(A)}{\partial A} = (\det(A)A^{-1})^{\mathrm{T}} = \det(A)(A^{-1})^{\mathrm{T}}$
- $\dfrac{\partial \ln\det(A)}{\partial A} = (A^{-1})^{\mathrm{T}}$
- $\mathrm{d}\det(A) = \mathrm{Tr}\left[\det(A)A^{-1}\mathrm{d}A\right]$
- $\mathrm{d}\ln\det(A) = \mathrm{Tr}\left[A^{-1}\mathrm{d}A\right]$

Note the trace operator can not be ignored in the differential formula, or the quantity at different sides of equality may be of different size. Although we derive this series from derivative to differential and from ordinary determinant to log determinant. However, the last one(differential of log determinant) is simplest for memorizing. Others could be derive quickly from it(using our bridging theorems).

4.6 Chain Rule

For all column vectors $x^{(i)}$:

$$\frac{\partial x^{(n)}}{\partial x^{(1)}} = \frac{\partial x^{(2)}}{\partial x^{(1)}} \frac{\partial x^{(3)}}{\partial x^{(2)}} \cdots \frac{\partial x^{(n)}}{\partial x^{(n-1)}} \tag{176}$$

Caution:

- The chain rule works only for our definition. Other definition may result in reverse order of multiplication.
- The chain rule here is listed only for completeness of the current topic. Use of chain rule is depricated. Most work can be done using the trace and differential schema.

Acknowledgements

Thanks prof. XU, Lei's tutorial on matrix calculus[12]. Besides, the author also benefits a lot from other online materials.

References

[1] Pili Hu. Matrix calculus. GitHub, https://github.com/hupili/tutorial/tree/master/matrix-calculus, 3 2012. HU, Pili's tutorial collection.

[2] Pili Hu. Tutorial collection. GitHub, https://github.com/hupili/tutorial, 3 2012. HU, Pili's tutorial collection.

[3] Matrix Calculus, Wikipedia, http://en.wikipedia.org/wiki/Matrix_calculus

[4] Matrix Norm, Wikipedia, http://en.wikipedia.org/wiki/Matrix_norm

[5] Matrix Trace, Wikipedia, http://en.wikipedia.org/wiki/Trace_%28linear_algebra%29

[6] Matrix Determinant, Wikipedia, http://en.wikipedia.org/wiki/Determinant

[7] Hessian Matrix, Wikipedia, http://en.wikipedia.org/wiki/Hessian_matrix

[8] Jacobian Matrix, Wikipedia, http://en.wikipedia.org/wiki/Jacobian_matrix_and_determinant

[9] Gaussian Distribution, Wikipedia, http://en.wikipedia.org/wiki/Normal_distribution

[10] Gaussian Integral, Wikipedia, http://en.wikipedia.org/wiki/Gaussian_integral

[11] Fubini's Theorem, Wikipedia, http://en.wikipedia.org/wiki/Fubini%27s_theorem

[12] XU, Lei, 2012, Machine Learning Theory(CUHK CSCI5030).

[13] Appendix from Introduction to Finite Element Methods book on University of Colorado at Boulder. http://www.colorado.edu/engineering/cas/courses.d/IFEM.d/IFEM.AppD.d/IFEM.AppD.pdf

Appendix

Message from Initiator(2012.04.03)

My three passes so far:

1. Learn the trace schema in the lecture.
2. Organize notes when doing the first homework.
3. In mid-term review, I revisited most properties.

Not until finishing the current version of this document do I grasp some details. It's very likely that everything is smooth when you read it. However, it may become a problem repeating some derivations. Those definitions I adopted are not all the initial version. Sometimes, I find it ugly to express certain result. Then I turn back modify the definition. Anyway, I don't see a unified dominant definition. If you check the history and talk page of Wikipedia[3], you'll find editing-wars happen every now and then. Even the single matrix calculus page is not coherent[3], let alone other pages. A precaution is, if you can not figure out the definition from the context, you'd better don't trust that formula. The safest thing is to learn his way of derivation and derive from your own definition. I also call for everyone's review of this document. It doesn't matter what background you have, for we try to build the system from scratch. I myself is an engineer rather than mathematician. When I first read others work, like [13], I bother to repeat the derivations and find some part is incorrect. That's why I reorganize my written notes, in order for me to reference in the future.

Shiney is a lazy girl who don't bother to derive anything but also wish to learn matrix calculus. Thus I compose all my notes for her. Besides, she wants to learn LaTeX in a fast food fashion. This document actually contains most things she needs in writing, from which the source and outcome are straightforward to learn.

Message Once in Cheat Sheet Section(2012.04.03)

This section is not yet available. I plan to sum up later. By the time I initiate this document, I had three passes of matrix calculus (appendix). The 4-th pass is left for me to refresh the mind in some future day.

Nevertheless, I think I already covered essential schema, property, formula, etc. Not all of them are in the form of theorems or propositions. A large part is covered using examples. Putting them all in one paper may be a good cheat sheet.

BTW, I also advocates version control and online collaboration tools for academic use. Such great platform should not be enjoyed by mere IT industry. I'll be glad if someone can contribute this section. Learning to use github costs no more than one hour. I'll of course pull any reasonable modifications from anyone. Typos, suggestions, technical fix, etc, will be acknowledged in the following section.

Amemdment Topics

20120424 The derivation of multivariate Fisher information matrix and Cramer-Rao bound uses quite a lot matrix calculus. It may be a good example of showing basic concepts. What's more, the derivation of general version(covariance $T(X)$ with $\mathbb{E}[T(X)] = \psi(\theta)$) benifits from chain rule. Fisher information do not seem of as wide interest as those applications shown above. Thus I temporarily keep it away from this tutorial.

Collaborator List

People who suggested content or pointed out mistakes:

- NULL.