A Practical Review of Matrix Calculus. (Version 1.2.)

Ali Taylan Cemgil
SNN, University of Nijmegen, The Netherlands
cemgil@mbfys.kun.nl
http://www.mbfys.kun.nl/∼cemgil

April 12, 2000

1 Introduction

Certain problems in statistics require the computation of derivatives of various quantities with respect to vectors or matrices. For example, in the maximum likelhod (ML) estimation of a multivariate Gaussian, the ML solution requires the computation of the derivative with respect the mean (vector) and the covariance matrix. If we want to compute standard error bars on the parameter estimates, we need to compute second derivatives, more precisely the inverse of negative Hessian of log likelihood. In the multivariate Gaussian example, the computation of error bars of the covariance matrix requires differentiation of a matrix quantity (the Jacobian) with respect to the covariance matrix. This expressions can be complicated to evaluate. Fortunately, with consistent definitions, it turns out that this derivatives have simple and convenient forms and are easily programmed in matrix computation packages such as MATLAB.

The results presented here are rather elementary. However, instead of providing a table of derivatives of common forms, we develop a practical notation for derivation of these quantities. In this respect this review is just about a slightly modified presentation of material from Magnus and Neudecker (1999). We use ideas from tensor notation surveyed in Arfken (1985).

2 Definition of Notation

We use small Greek letters to denote scalar variables, Latin small letter to denote vector variables and Latin capital letters to denote matrix variables. For example ξ is a scalar, x is a vector and X is a matrix. Function names obey the same convention and we will use ϕ, f and F to denote functions which evaluate to a scalar, vector and a matrix respectively. Some examples are

$$\phi(X) = \mathbf{Tr}\, X$$

$$f(x) = Ax$$

$$F(\xi) = \begin{pmatrix} \xi & 0 \\ 0 & \xi^2 \end{pmatrix}$$

The vector notation is very compact and easy to program and these features make it a very useful tool in derivations. For example Ax compactly encodes the expression $\sum_{j=1}^{n} A_{ij}x_j$. The fact that x is a column vector and A is a matrix of compatible size is implicit in this notation. Thus in analytic work one has to check the sizes and shapes of objects in an expression for consistency. Unfortunately, representation of objects involving more then two indices turns out to be less transparent. In this respect the index notation is more expressive however besides requiring "more ink" it can lead to ambiguities in transforming theexpression back to matrix notation. Suppose we wish to express $x^T A^T$ in index form. The result is again $\sum_{j=1}^{n} x_j A_{ij}$ and consequently we can not distinguish between the row or column vectors. Clearly, the reason is that the notation in its present form does not distinguish between row and column indices. We extend the index notation as follows: The expression

$$Ax$$

will be denoted as

$$A_i^j x_j$$

We get rid of the summation sign and adapt the convention introduced by Einstein, that repeated index means summation. Moreover, we organize the indices such that an index occurring in the subscript (superscript) corresponds to a row (column) index. The resulting notation is almost as compact as the matrix notation; we only keep track of indices explicitly. We can distinguish $x^T A^T$ from Ax since the former expression now corresponds to $A_i^j x^i$, an object with only one upper index, hence a row vector. We denote the size of A_i^j as $|i| \times |j|$, where $|i|$ denotes the number of states that the index i addresses.

In physics and tensor analysis (Arfken, 1985), the upper and lower indices are classified as *contravariant* and *covariant*, which determine the behavior of corresponding quantities (i.e. numbers indexed using these indices) under various transformations. Here, it is rather immaterial how one organizes these numbers in a data structure. In contrast, our motivation is to develop a practical notation to find simple expressions for various matrix derivatives. In that respect it turns out to be useful to classify the indices as row and column indices. [1] Consequently, the result of any operation can be expressed in form of a matrix regardless of the number of indicies involved. Beside possessing the advantage of easy implementation, one can use results from matrix theory directly on resulting objects. For example, the derivative of a matrix w.r.t. a matrix comes out to be itself a Jacobian, and questions relating to a vanishing Jacobian determinants remain meaningful (Magnus and Neudecker, 1999).

We demonstrate the power of the index notation by a few examples. The $\mathbf{vec}\, X$ of a matrix is defined to be a column vector obtained from X by stacking the columns of X on top of each other. For example if

$$X = \begin{pmatrix} \alpha & \beta \\ \gamma & \eta \end{pmatrix}$$

then

$$\mathbf{vec}\, X = \begin{pmatrix} \alpha \\ \gamma \\ \beta \\ \eta \end{pmatrix} \tag{1}$$

The *Kronecker product* of X with Y is defined to be the partitioned matrix

$$X \otimes Y = \begin{pmatrix} \alpha Y & \beta Y \\ \gamma Y & \eta Y \end{pmatrix}$$

In index notation $\mathbf{vec}\, X$ becomes X_{ji}. We note that this is just a column vector indexed with a double index. We apply here one more convention that the slowest index appears first. To understand what we mean by this

[1] Column vectors have row indices and row vectors have column indices.

convention consider the example in Eq. 1. This is a column vector indexed with one row index, say k. If k increases by one, in the original matrix X_i^j this corresponds first to an increment in i and then to j. In other words while k goes from 1 to 4, i, the fast index, traces the numbers $1, 2, 1, 2$ and j, the slower index, traces $1, 1, 2, 2$. By applying the same convention we have

$$Z = X \otimes Y \ = X_i^j Y_k^l = Z_{ik}^{jl}$$

The result can easily be seen as follows: Since Y is replicated in every cell of X, clearly the indices of Y are faster. Until i or j change we have to traverse over all k and l. Some additional examples are shown in Table 1.

Matrix Notation	Index Notation	MATLAB	Comment
X	X_i^j	`X`	$\|i\| \times \|j\|$ Matrix
X^T	X_j^i	`X'`	Matrix Transpose
X^{-1}	$(X^{-1})_j^i$	`inv(X)`	Matrix Inverse
$\mathbf{vec}\, X$	X_{ji}	`X(:)`	Slowest index first
XY	$X_i^j Y_j^l$	`X*Y`	Matrix Multiplication
$X \otimes Y$	$X_i^j Y_k^l$	`kron(X,Y)`	Kronecker Product
$\mathbf{Tr}\, X$	X_i^i	`trace(X)`	Trace (X must be square)

Table 1: Notation

Probabily the most important advantage of index notation is that it relaxes the structural constrains put onto an expression and enables us to move terms around once all components are expressed in index form. We demonstrate this by a proof of an important identity in matrix algebra which relates $\otimes$ and **vec**:

$$(A \otimes B)\, \mathbf{vec}\, X = \mathbf{vec}(BXA^T) \tag{2}$$

Here A is $|m| \times |n|$, B is $|p| \times |q|$ and X is consequently $|q| \times |n|$ so that the term BXA^T makes sense. The form of $(A \otimes B)\, \mathbf{vec}\, X$ is T_{mp}.

$$\begin{aligned}
((A \otimes B)\, \mathbf{vec}\, X)_{mp} &= (A \otimes B)_{mp}^{\alpha} (\mathbf{vec}\, X)_{\alpha} && (3)\\
&= (A \otimes B)_{mp}^{nq} (\mathbf{vec}\, X)_{nq} && (4)\\
&= \mathbf{vec}((A \otimes B)_{mp}^{nq} (\mathbf{vec}\, X)_{nq}) && (5)\\
&= \mathbf{vec}((A \otimes B)_{mp}^{nq} X_q^n) && (6)\\
&= \mathbf{vec}(A_m^n B_p^q X_q^n) && (7)\\
&= \mathbf{vec}(B_p^q X_q^n (A^T)_n^m) && (8)\\
&= \mathbf{vec}((BXA^T)_p^m) && (9)\\
&= (\mathbf{vec}(BXA^T))_{mp} && (10)
\end{aligned}$$

In Eq. 3, α is just a dummy index, since it will be summed over. This summation could be evaluated in arbitrary order and we choose to set it to $\alpha = nq$. The result will be an object of form T_{mp} so the additional **vec** operation is justified. Eq. 5 shows how the matrix structure is relaxed: the outside **vec** arranges the result as a vector T_{mp}, so we can arrange the terms arbitrarly in the argument of **vec** as long as the resulting object has the form T_p^m.

3 Matrix Calculus: Differentiation

3.1 The first Derivative : The Jacobian

Before the discussion of the general case, namely the derivative of a matrix function $F(X)$ we consider the derivative of a vector function $f(x)$. We define the derivative of $(f(x))_i$ w.r.t. x_α as

$$\mathsf{D}f(x) = \frac{\partial (f(x))_i}{\partial x^\alpha} = \frac{\partial f_i}{\partial x^\alpha} \tag{11}$$

Note that the row index α became a column index, i.e. we have

$$\mathsf{D}f(x) = \frac{\partial f(x)}{\partial x^T} \tag{12}$$

Let us consider an example:

$$\begin{aligned} f(x) &= Ax && (13) \\ &= \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} \\ \alpha_{2,1} & \alpha_{2,2} \\ \alpha_{3,1} & \alpha_{3,2} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} && (14) \\ &= \begin{pmatrix} \alpha_{1,1}\xi_1 + \alpha_{1,2}\xi_2 \\ \alpha_{2,1}\xi_1 + \alpha_{2,2}\xi_2 \\ \alpha_{3,1}\xi_1 + \alpha_{3,2}\xi_2 \end{pmatrix} && (15) \end{aligned}$$

We have three quantities to be differentiated w.r.t. two quantities. We organize these derivatives in a matrix of size 3×2. We get the intuitive result:

$$\begin{aligned}
\mathsf{D}f(x) &= \begin{pmatrix} \frac{\partial f_1}{\partial x^1} & \frac{\partial f_1}{\partial x^2} \\ \\ \frac{\partial f_2}{\partial x^1} & \frac{\partial f_2}{\partial x^2} \\ \\ \frac{\partial f_3}{\partial x^1} & \frac{\partial f_3}{\partial x^2} \end{pmatrix} && (16) \\
&= \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} \\ \alpha_{2,1} & \alpha_{2,2} \\ \alpha_{3,1} & \alpha_{3,2} \end{pmatrix} && (17) \\
&= A && (18)
\end{aligned}$$

Let $f(x) = x$, the identity function. Then we have

$$\mathsf{D}x = \frac{\partial x_i}{\partial x^\alpha} = \delta_i^\alpha = I$$

where δ_i^α is the Kronecker-delta symbol, which is defined as

$$\delta_i^\alpha = \begin{cases} 1 & i = \alpha \\ 0 & \text{otherwise} \end{cases}$$

This symbol turns out to be very useful in evaluation of expressions involving derivatives. To see this consider again the example $f(x) = Ax$:

$$\begin{aligned}
\mathsf{D}Ax &= A\mathsf{D}x && (19) \\
&= A_i^j \frac{\partial x_j}{\partial x^\alpha} && (20) \\
&= A_i^j \delta_j^\alpha && (21) \\
&= A_i^\alpha && (22)
\end{aligned}$$

The usefulness of δ can be better appreciated when we consider the general case.

3.1.1 General case

We define the derivative of $F(X)$ with respect to X as

$$\mathsf{D}F(X) \stackrel{\text{def}}{=} \frac{\partial \mathbf{vec}\, F(X)}{\partial (\mathbf{vec}\, X)^T} \tag{23}$$

Note that this definition is consistent with the case when F and X are vectors. The marices are merely vectorized. As a first example consider $F(X) = X$. Then by definition we have

$$\mathsf{D}F(X) = \frac{\partial X_{ji}}{\partial X^{\beta\alpha}} \tag{24}$$

since $\mathbf{vec}\, X = X_{ji}$ and $(\mathbf{vec}\, X)^T = (X_{\beta\alpha})^T = X^{\beta\alpha}$. This is an object of form $Z_{ji}^{\beta\alpha}$. Indeed as in the vector case

$$\begin{aligned} \frac{\partial X_{ji}}{\partial X^{\beta\alpha}} &= \delta_{ji}^{\beta\alpha} && (25) \\ &= \delta_j^\beta \delta_i^\alpha && (26) \\ &= I \otimes I && (27) \end{aligned}$$

Note that this result demonstrates the consistency of the definition. We get an identity matrix of size $(|i||j|) \times (|i||j|)$. If we had organized the derivatives on some other (arbitrary) order, (e.g. $Z_{i\alpha}^{\beta j}$) we would not get an identity matrix. In general, that result would not posess any structure. It would only display the derivatives but nothing more.

Example:

$$\begin{aligned} F(X) &= AXB = A_i^k X_k^l B_l^j && (28) \\ \mathsf{D}F(X) &= A(\mathsf{D}X)B = \frac{\partial F_{ji}}{\partial X^{\beta\alpha}} && (29) \\ &= A_i^k (\frac{\partial X_{lk}}{\partial X^{\beta\alpha}}) B_l^j && (30) \\ &= A_i^k \delta_l^\beta \delta_k^\alpha B_l^j && (31) \\ &= A_i^\alpha B_\beta^j && (32) \end{aligned}$$

In terms of tensor calculus this is the answer, i.e. the general term of the expression for $\mathsf{D}AXB$. However, we require by our definition that the resulting object has to be of form $Z_{ji}^{\beta\alpha}$. So we continue transforming further:

$$\begin{aligned} &= (B^T)_j^\beta A_i^\alpha && (33) \\ &= B^T \otimes A && (34) \end{aligned}$$

Corollary:

$$\begin{aligned} \mathsf{D}(AX) &= I \otimes A && (35) \\ \mathsf{D}(XA) &= A^T \otimes I && (36) \end{aligned}$$

Now we will consider the derivative of X^{-1} which appears to be a little more tricky. We will first solve by index notation, and then demonstrate how the result is obtained by using differentials:

$$\begin{aligned} I &= K^{-1}K && (37)\\ \delta_i^l &= (K^{-1})_i^j K_j^l && (38)\\ 0 &= \mathsf{D}((K^{-1})_i^j)K_j^l + (K^{-1})_i^j \mathsf{D}(K_j^l) && (39)\\ &= T_{ji}^{\beta\alpha} K_j^l + (K^{-1})_i^j \delta_l^\beta \delta_j^\alpha && (40) \end{aligned}$$

We multiply both sides with $(K^{-1})_l^r$,

$$\begin{aligned} 0 &= T_{ji}^{\beta\alpha}\delta_j^r + (K^{-1})_i^j \delta_l^\beta \delta_j^\alpha (K^{-1})_l^r && (41)\\ T_{ri}^{\beta\alpha} &= -(K^{-1})_i^\alpha (K^{-1})_\beta^r && (42)\\ &= -(K^{-1})^T \otimes K^{-1} && (43) \end{aligned}$$

A more compact notation is achieved by using differentials

$$\begin{aligned} 0 &= \mathsf{d}(K^{-1}K) && (44)\\ &= (\mathsf{d}K^{-1})K + K^{-1}(\mathsf{d}K) && (45)\\ \mathsf{d}K^{-1} &= -K^{-1}(\mathsf{d}K)K^{-1} && (46) \end{aligned}$$

At this stage we can introduce indicies and write

$$\mathsf{d}(K^{-1})_i^r = -(K^{-1})_i^j (\mathsf{d}K)_j^l (K^{-1})_l^r \qquad (47)$$

then we differentiate w.r.t X_α^β and obtain the result in the same lines as Eq. 43. The thing to keep in mind is that the object $\mathsf{d}K$ "behaves" like a matrix.

3.1.2 Special Cases: $\mathsf{D}\phi(X)$ and $\mathsf{D}F(\xi)$

The definition $\mathsf{D}F(X)$ in Eq. 23 is changed when we have the derivative of a scalar function w.r.t. a matrix ($\frac{\partial\phi(X)}{\partial X}$) or the derivative of a matrix function w.r.t. a scalar ($\frac{\partial F(\xi)}{\partial\xi}$). In these cases we define

$$\begin{aligned} \mathsf{D}\phi(X) &= \frac{\partial\phi}{\partial X_\alpha^\beta}\\ \mathsf{D}F(\xi) &= \frac{\partial F_i^j}{\partial\xi} \end{aligned}$$

Example

$$
\begin{aligned}
\phi(X) &= \mathbf{Tr}\, X^T X = (X^T)^i_j X^j_i = (X^j_i)^2 && (48)\\
\mathsf{D}\phi(X) &= 2X^j_i(\mathsf{D}X^j_i) && (49)\\
&= 2X^j_i \frac{\partial X_{ji}}{\partial X^{\beta\alpha}} && (50)\\
&= 2X^j_i \delta^\beta_j \delta^\alpha_i && (51)\\
&= 2X^\beta_\alpha && (52)\\
&= 2X && (53)
\end{aligned}
$$

A slightly general form arises in many applications such as the Kalman Filter:

$$
\begin{aligned}
\phi(X) &= \mathbf{Tr}\, BX^T CX && (54)\\
&= B^j_i (X^T)^k_j C^l_k X^i_l && (55)\\
\mathsf{D}\phi(X) &= \frac{\partial \phi(X)}{\partial X^\beta_\alpha} && (56)\\
&= B^j_i \frac{\partial X_{jk}}{\partial X^{\beta\alpha}} C^l_k X^i_l + B^j_i X^j_k C^l_k \frac{\partial X_{il}}{\partial X^{\beta\alpha}} && (57)\\
&= B^j_i \delta^\beta_j \delta^\alpha_k C^l_k X^i_l + B^j_i X^j_k C^l_k \delta^\beta_i \delta^\alpha_l && (58)\\
&= B^\beta_i C^l_\alpha X^i_l + B^j_\beta X^j_k C^\alpha_k && (59)\\
&= C^l_\alpha X^i_l B^\beta_i + (C^T)^k_\alpha X^j_k (B^T)^\beta_j && (60)\\
&= CXB + C^T X B^T && (61)
\end{aligned}
$$

3.2 The second Derivative : The Hessian

We define the Hessian of a matrix function as

$$\mathsf{H}F(X) = \mathsf{D}((\mathsf{D}F(X))^T)$$

To see the rationale behind this definition in contrast to appeerently more natural definition $\mathsf{D}(\mathsf{D}F)$, we consider the second derivative of $F(X) =$

$X^T A X$

$$
\begin{aligned}
F(X) &= \begin{pmatrix} \phi_{1,1} & \phi_{1,2} \\ \phi_{2,1} & \phi_{2,2} \end{pmatrix} = \begin{pmatrix} \xi & \gamma \\ \eta & \lambda \end{pmatrix} \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} \\ \alpha_{2,1} & \alpha_{2,2} \end{pmatrix} \begin{pmatrix} \xi & \eta \\ \gamma & \lambda \end{pmatrix} \qquad (62) \\
&= \begin{pmatrix} \alpha_{1,1}\xi^2 + (\alpha_{1,2}+\alpha_{2,1})\gamma\xi + \alpha_{2,2}\gamma^2 & \alpha_{1,1}\eta\xi + \alpha_{1,2}\lambda\xi + \alpha_{2,1}\eta\gamma + \alpha_{2,2}\lambda\gamma \\ \alpha_{1,1}\eta\xi + \alpha_{1,2}\eta\gamma + \alpha_{2,1}\lambda\xi + \alpha_{2,2}\lambda\gamma & \alpha_{1,1}\eta^2 + (\alpha_{1,2}+\alpha_{2,1})\lambda\eta + \alpha_{2,2}\lambda^2 \end{pmatrix} \\
\mathsf{D}F(X) &= \begin{pmatrix} 2\alpha_{1,1}\xi + (\alpha_{1,2}+\alpha_{2,1})\gamma & (\alpha_{1,2}+\alpha_{2,1})\xi + 2\alpha_{2,2}\gamma & 0 & 0 \\ \alpha_{1,1}\eta + \alpha_{2,1}\lambda & \alpha_{1,2}\eta + \alpha_{2,2}\lambda & \alpha_{1,1}\xi + \alpha_{1,2}\gamma & \alpha_{2,1}\xi + \alpha_{2,2}\gamma \\ \alpha_{1,1}\eta + \alpha_{1,2}\lambda & \alpha_{2,1}\eta + \alpha_{2,2}\lambda & \alpha_{1,1}\xi + \alpha_{2,1}\gamma & \alpha_{1,2}\xi + \alpha_{2,2}\gamma \\ 0 & 0 & 2\alpha_{1,1}\eta + (\alpha_{1,2}+\alpha_{2,1})\lambda & (\alpha_{1,2}+\alpha_{2,1})\eta + 2\alpha_{2,2}\lambda \end{pmatrix} \\
\mathsf{H}F(X) &= \begin{pmatrix} 2\alpha_{1,1} & \alpha_{1,2}+\alpha_{2,1} & 0 & 0 \\ \alpha_{1,2}+\alpha_{2,1} & 2\alpha_{2,2} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \alpha_{1,1} & \alpha_{2,1} \\ 0 & 0 & \alpha_{1,2} & \alpha_{2,2} \\ \alpha_{1,1} & \alpha_{2,1} & 0 & 0 \\ \alpha_{2,1} & \alpha_{2,2} & 0 & 0 \\ 0 & 0 & \alpha_{1,1} & \alpha_{1,2} \\ 0 & 0 & \alpha_{2,1} & \alpha_{2,2} \\ \alpha_{1,1} & \alpha_{1,2} & 0 & 0 \\ \alpha_{1,2} & \alpha_{2,2} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 2\alpha_{1,1} & \alpha_{1,2}+\alpha_{2,1} \\ 0 & 0 & \alpha_{1,2}+\alpha_{2,1} & 2\alpha_{2,2} \end{pmatrix} \qquad (63) \\
&= \begin{pmatrix} A + A^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{0} & A \\ A^T & \mathbf{0} \\ \mathbf{0} & A^T \\ A & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{0} & A + A^T \end{pmatrix} = \begin{pmatrix} \mathsf{H}\,\phi_{1,1} \\ \mathsf{H}\,\phi_{2,1} \\ \mathsf{H}\,\phi_{1,2} \\ \mathsf{H}\,\phi_{2,2} \end{pmatrix} \qquad (64)
\end{aligned}
$$

The reader can verify easly that the alternative definition $\mathsf{D}(\mathsf{D}F)$ would lead to a Hessian matrix which could not be expressed as the partitioned form in Eq. 64. Nevertheless, in practice Hessians are computed for scalar functions $\phi(X)$. For this case it is easy to see that $\mathsf{D}((\mathsf{D}F)^T) = \mathsf{D}(\mathsf{D}F)$. As an example we evaluate the Hessian of $\phi(X) = \mathbf{Tr}\, BX^T CX$ (Eq. 54). We identify the resulting object of size $Z^{\gamma\eta}_{\beta\alpha}$

$$
\begin{aligned}
\mathsf{D}(\mathsf{D}\phi(X)) &= \frac{\partial(\mathsf{D}\phi(X))_{\beta\alpha}}{\partial X^{\gamma\eta}} && (65)\\
&= B_i^\beta C_\alpha^l \frac{\partial X_{il}}{\partial X^{\gamma\eta}} + B_\beta^j \frac{\partial X_{jk}}{\partial X^{\gamma\eta}} C_k^\alpha && (66)\\
&= B_i^\beta C_\alpha^l \delta_i^\gamma \delta_l^\eta + B_\beta^j \delta_j^\gamma \delta_k^\eta C_k^\alpha && (67)\\
&= B_\gamma^\beta C_\alpha^\eta + B_\beta^\gamma C_\eta^\alpha && (68)\\
\mathsf{H}\phi(X) &= B^T \otimes C + B \otimes C^T && (69)
\end{aligned}
$$

4 Applications

4.1 ML Estimation of a multivariate Gaussian

The derived results can be used to estimate parameters and corresponding error bars of a multivariate Gaussian from data. The Gaussian distribution is given by

$$p(x) = |2\pi K|^{-\frac{1}{2}} \exp(-\frac{1}{2}(x-m)^T K^{-1}(x-m))$$

where $|K|$ denotes the determinant of K.

The log likelihood of a i.i.d. data set $\{x_i\}_{i=1}^N$ [2] is

$$
\begin{aligned}
\mathcal{L} &= \log \prod_{i=1}^N p(x_i) && (70)\\
&= -\frac{N}{2}\log|2\pi K| - \frac{1}{2}\sum_{i=1}^N (x_i - m)^T K^{-1}(x_i - m) && (71)
\end{aligned}
$$

To find the ML parameters we compute the derivatives $\frac{\partial\mathcal{L}}{\partial m_\alpha}$ and $\frac{\partial\mathcal{L}}{\partial K_\alpha^\beta}$ and set them to zero.

$$
\begin{aligned}
\frac{\partial\mathcal{L}}{\partial m} &= \frac{\partial}{\partial m}\left(-N s^T K^{-1} m + \frac{N}{2} m^T K^{-1} m\right) && (72)\\
&= N K^{-1}(m - s) && (73)
\end{aligned}
$$

where s is the sample mean $s = \frac{\sum_{i=1}^N x_i}{N}$. To find K, we write $\mathcal{L}$ as

$$\mathcal{L} = -\frac{N}{2}\,\mathbf{Tr}\log 2\pi K - \frac{N}{2}\,\mathbf{Tr}\,K^{-1}P$$

[2]Here, x_i denotes the i'th sample in the data set, not the i'th element of x. We denote this fact by explicitly showing the summation sign.

from which it follows that

$$\frac{\partial \mathcal{L}}{\partial K} = -\frac{N}{2}K^{-1} + \frac{N}{2}K^{-1}PK^{-1}$$

Here P is the sample covariance $P = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)(x_i - m)^T$. By setting the derivatives to zero we verify the standard ML solutions $m = s$ and $K = P$.

To find the error bars we need to compute second derivatives $\frac{\partial}{\partial m}(\frac{\partial \mathcal{L}}{\partial m})$, $\frac{\partial}{\partial K}(\frac{\partial \mathcal{L}}{\partial K})$ and $\frac{\partial}{\partial K}(\frac{\partial \mathcal{L}}{\partial m})$. These are found to be

$$\frac{\partial}{\partial m}(\frac{\partial \mathcal{L}}{\partial m}) = -NK^{-1} \tag{74}$$

$$\begin{aligned}
\frac{\partial}{\partial K}(\frac{\partial \mathcal{L}}{\partial K}) &= \frac{\partial}{\partial K}(-\frac{N}{2}K^{-1} + \frac{N}{2}K^{-1}PK^{-1}) && (75)\\
&= \frac{N}{2}K^{-1}(\mathsf{D}K)K^{-1} + N(-K^{-1}(\mathsf{D}K)K^{-1}PK^{-1} + K^{-1}PK^{-1}(\mathsf{D}K)K^{-1}) && (76)\\
&= -\frac{N}{2}K^{-1}(\mathsf{D}K)K^{-1} && (77)\\
&= -\frac{N}{2}K^{-1} \otimes K^{-1} && (78)\\
&= -(\frac{2}{N}K \otimes K)^{-1} && (79)
\end{aligned}$$

Eq. 77 follows because $K^{-1}P = I$ at ML solution.

$$\frac{\partial}{\partial m}(\frac{\partial \mathcal{L}}{\partial K}) = 0 \tag{80}$$

The Hessian is given by

$$\begin{aligned}
\mathsf{H}f(x) &= \mathsf{D}(\mathsf{D}f(x)) && (81)\\
&= \begin{pmatrix} -(\frac{1}{N}K)^{-1} & 0 \\ 0 & -(\frac{2}{N}K \otimes K)^{-1} \end{pmatrix} && (82)
\end{aligned}$$

The error bars for m and K are given by $\mathbf{diag}(\frac{1}{N}K)^{\frac{1}{2}}$ and $\mathbf{diag}(\frac{2}{N}K \otimes K)^{\frac{1}{2}}$ respectively.

References

George Arfken. *Mathematical Methods for Physicists*, chapter 3. Academic Press, 1985.

Jan R. Magnus and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Economics*. Wiley, 1999.