Statistics Netherlands
Division of Methodology and Quality
Department of Methodology
P.O. Box 4000
2270 JM Voorburg
The Netherlands

A Survey of Matrix Differentiation*

Kees Zeelenberg

*I wish to thank Anco Hundepool for converting the manuscript from PCWrite to LaTeX.
The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

Projectnumber: *M1-81-209*
BPAnumber: *12976-93-M1-1*
Date: *March 21, 2008*

Contents

Abstract **1**

1 Introduction **1**

2 Preliminaries **1**

2.1 The trace 1

2.2 The Kronecker product 2

2.3 The vec-operator 3

2.4 The commutation matrix 4

2.5 The duplication matrix 5

2.6 Diagonality 6

3 First-order differentiation **7**

3.1 Differentiability of vector functions 7

3.2 Differentiability of matrix functions 8

3.3 Chain rule 8

3.4 Properties of differentials 9

3.5 Examples 9

4 Second-order differentiation **12**

4.1 Twice-differentiability 12

4.2 The second differential 12

4.3 Matrix functions 13

4.4 Examples 14

5 Linearization of the regression estimator **17**

6 Maximum-likelihood estimation of the multivariate linear model **18**

6.1 Introduction 18

6.2 The model 18

6.3 First-order conditions 19

6.4 The Hessian matrix 19

6.5 The information matrix 20

7 Maximum-likelihood estimation of the factor-analysis model 21

7.1 Introduction 21

7.2 The model 21

7.3 First-order conditions 22

7.4 The Hessian matrix 23

7.5 The information matrix 24

A SURVEY OF MATRIX DIFFERENTIATION

An summary of first and second-order differentiation of matrix functions is given. As example, these techniques are applied to maximum-likelihood estimation of the multivariate linear model and the factor-analysis model.

Keywords: matrix differentiation, multivariate linear model, factor analysis, regression estimator
AEA-code: C30

Keywords:

1 Introduction

In this paper an overview of matrix differentiation techniques and some applications is given. In section 2 we collect some matrix results from Magnus (1988). Sections 3 and 4, drawn from Magnus and Neudecker (1988), treat first-order differentiation, respectively second-order differentiation. Section 5 gives an application to maximum-likelihood estimation of the multivariate linear model, section 6 to maximum-likelihood estimation of the factor-analysis model, and section 7 to linearization of the regression estimator.

2 Preliminaries

This section brings together several results from matrix analysis that are useful for matrix differentiation techniques; see Magnus (1988) for much more details and proofs. We use several operators such as *tr* (trace of a matrix) and *vec* (vector of a matrix). These operators have the highest priority, *e.g.* $\alpha\,\mathrm{tr}(A)$ means: $\alpha \times [\mathrm{tr}(A)]$. If the operator is followed by a space then it extends until the next space, closing bracket, comma, or period, *e.g.* $\mathrm{tr}\,AB$ means: $\mathrm{tr}(AB)$.

2.1 The trace

Let A and B be $n \times n$-matrices. The *trace* of a matrix is defined as the sum of its diagonal elements:

$$\mathrm{tr}\,A = \sum_{i=1}^{n} a_{ii}. \tag{2.1}$$

The trace has the following properties:

$$\mathrm{tr}(A + B) = \mathrm{tr}\,A + \mathrm{tr}\,B, \tag{2.2}$$

$$\operatorname{tr}(A') = \operatorname{tr} A, \tag{2.3}$$

$$\operatorname{tr} AB = \operatorname{tr} BA = \operatorname{tr} A'B' = \operatorname{tr} B'A', \tag{2.4}$$

$$\operatorname{tr} \alpha A = \alpha \operatorname{tr} A. \tag{2.5}$$

2.2 The Kronecker product

Let A be an $m \times n$-matrix and B a $p \times q$-matrix. The *Kronecker product* of A and B is the $mp \times nq$-matrix

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{pmatrix} \tag{2.6}$$

The Kronecker product has the same priority as the ordinary product. For example, $A \otimes B + C = (A \otimes B) + C$, $AB \otimes C = (AB) \otimes C$. Note that $a_{ij}b_{rs} = (A \otimes B)_{(i-1)p+r,(j-1)q+s}$ $(i = 1, 2, ..., m;\ j = 1, 2, ..., n;\ r = 1, 2, ..., p;\ s = 1, 2, ..., q)$.

Let C and D be matrices, x and y vectors, and α a scalar. The Kronecker product has the following properties (it is assumed that any product and sum exist):

$$(A \otimes B)' = A' \otimes B', \tag{2.7}$$

$$(A + B) \otimes (C + D) = A \otimes C + A \otimes D + B \otimes C + B \otimes D, \tag{2.8}$$

$$(A \otimes B)(C \otimes D) = AC \otimes BD, \tag{2.9}$$

$$\alpha A = \alpha \otimes A = A \otimes \alpha = A\alpha, \tag{2.10}$$

$$x \otimes y' = xy' = y' \otimes x. \tag{2.11}$$

If A and B are square of order m respectively n, then

$$\operatorname{tr}(A \otimes B) = (\operatorname{tr} A)(\operatorname{tr} B), \tag{2.12}$$

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}, \tag{2.13}$$

$$(A \otimes B)^{+} = A^{+} \otimes B^{+}, \tag{2.14}$$

with a + as superscript denoting the Moore-Penrose inverse,

$$r(A \otimes B) = r(A)r(B); \tag{2.15}$$

if A is an $m \times m$-matrix and B and $p \times p$-matrix, then

$$|A \otimes B| = |A|^p |B|^m; \tag{2.16}$$

if λ_i are the characteristic values $(i = 1, 2, ..., m)$ of A with characteristic vectors x_i and μ_j $(j = 1, 2, ..., p)$ the characteristic values of B, then the characteristic values of $A \otimes B$ are $\lambda_i \mu_j$ with characteristic values $x_i \otimes y_j$.

2.3 The vec-operator

Let A be an $m \times n$-matrix and a_i the i-th column of A. Then *vec A* is the *mn*-vector defined by

$$\operatorname{vec} A = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}. \tag{2.17}$$

Note that $a_{ij} = (\operatorname{vec} A)_{(j-1)m+i}$.

Let A, B, C, and D be matrices, and x and y vectors. The vec-operator has the following properties (it is assumed that any product exists):

$$\operatorname{vec} x' = \operatorname{vec} x = x, \tag{2.18}$$

$$\operatorname{vec} xy' = y \otimes x, \tag{2.19}$$

$$\operatorname{tr} AB = (\operatorname{vec} A')'(\operatorname{vec} B), \tag{2.20}$$

$$\operatorname{tr} ABCD = (\operatorname{vec} D')'(C' \otimes A)(\operatorname{vec} B) = (\operatorname{vec} D)'(A \otimes C')(\operatorname{vec} B'), \tag{2.21}$$

$$\operatorname{vec} ABC = (C' \otimes A)(\operatorname{vec} B), \tag{2.22}$$

$$ABx = (x' \otimes A)(\operatorname{vec} B) = (A \otimes x')(\operatorname{vec} B'). \tag{2.23}$$

If A is an $m \times n$-matrix and B an $n \times q$-matrix, then

$$\operatorname{vec} AB = (B' \otimes I_m)(\operatorname{vec} A) = (B' \otimes A)(\operatorname{vec} I_n) = (I_q \otimes A)(\operatorname{vec} B). \tag{2.24}$$

If A, B, and V are square matrices of the same order and V is symmetric, then

$$(\operatorname{vec} V)'(A \otimes B)(\operatorname{vec} V) = (\operatorname{vec} V)'(B \otimes A)(\operatorname{vec} V). \tag{2.25}$$

If x is an m-vector and y an n-vector, then from (2.19) and (2.24) we have

$$x \otimes y = \operatorname{vec}(yx') = (I_m \otimes y)x = (x \otimes I_n)y. \tag{2.26}$$

Let X be an $m \times n$-matrix, Y a $p \times q$-matrix, A an $n \times q$-matrix, and B an $m \times p$-matrix, such that $(\operatorname{vec} X)(\operatorname{vec} Y)' = A \otimes B$. Then

$$x_{ij} y_{rs} = a_{js} b_{ir}, \qquad i = 1, 2, ..., m;\ j = 1, 2, ..., n;\ r = 1, 2, ..., p;\ s = 1, 2, ..., q. \tag{2.27}$$

Similar formulae hold for matrices with the structure of $(\operatorname{vec} X)(\operatorname{vec} Y)'$. An example is the covariance matrix of the vec of a stochastic matrix: if $\operatorname{Var}(\operatorname{vec} X) = \operatorname{E}(\operatorname{vec} X - \operatorname{E} \operatorname{vec} X)(\operatorname{vec} X - \operatorname{E} \operatorname{vec} X)' = A \otimes B$, then

$$\operatorname{cov}(x_{ij}, x_{rs}) = a_{js} b_{ir}, \qquad i = 1, 2, ..., m;\ j = 1, 2, ..., n;\ r = 1, 2, ..., p;\ s = 1, 2, ..., q. \tag{2.28}$$

Another example is the matrix with second partial derivatives of a real-valued matrix function: if $\partial^2\phi/(\partial(\mathrm{vec}\,X)\partial(\mathrm{vec}\,X)') = A \otimes B$, then

$$\frac{\partial^2\phi}{\partial x_{ij}\partial x_{rs}} = a_{js}b_{ir},$$

$$i = 1, 2, ..., m;\ \ j = 1, 2, ..., n;\ \ r = 1, 2, ..., p;\ \ s = 1, 2, ..., q; \quad (2.29)$$

see section 4.4.

2.4 The commutation matrix

The *commutation* matrix is the permutation matrix that transforms the vec of a matrix into the vec of the transpose of that matrix:

$$K_{mn}(\mathrm{vec}\,A) = \mathrm{vec}\,A', \quad (2.30)$$

where A is an $m \times n$-matrix and K_{mn} is the $mn \times mn$-commutation matrix for matrices of order (m, n). Note that $\mathrm{vec}\,A'$ is the vector with the *rows* of A stacked; it is sometimes denoted as $\overline{\mathrm{vec}}(A)$.

The commutation matrix K_{mm} will be denoted by K_m. Since K_{mn} is a permutation matrix, it is orthogonal and thus

$$K_{mn}^{-1} = K_{mn}' = K_{nm}. \quad (2.31)$$

Also

$$K_{m1} = K_{1m} = I_m. \quad (2.32)$$

The commutation matrix derives its name from the fact that it reverses ('commutes') the order of Kronecker products:

$$K_{pm}(A \otimes B) = (B \otimes A)K_{qn}, \quad (2.33)$$

where B is a $p \times q$-matrix. The commutation matrix can be used to write the vec of a Kronecker product as the Kronecker product of the vec's:

$$\mathrm{vec}(A \otimes B) = (I_n \otimes K_{qm} \otimes I_p)[(\mathrm{vec}\,A) \otimes (\mathrm{vec}\,B)]. \quad (2.34)$$

An explicit expression for the commutation matrix is

$$K_{mn} = \sum_{i=1}^{m}\sum_{j=1}^{n}(H_{ij} \otimes H_{ij}'), \quad (2.35)$$

where H_{ij} is the $m \times n$-matrix with 1 as element (i, j) and 0 elsewhere.

Let X be an $m \times n$-matrix, Y a $p \times q$-matrix, A an $m \times q$-matrix, and B an $n \times p$-matrix, such that $(\mathrm{vec}\,X)(\mathrm{vec}\,Y)' = K_{nm}(A \otimes B)$. Then

$$x_{ij}y_{rs} = a_{is}b_{jr},$$

$$i = 1,2,...,m;\ j = 1,2,...,n;\ r = 1,2,...,p;\ s = 1,2,...,q. \quad (2.36)$$

Similar formulae hold for matrices with the structure of $(\text{vec}\, X)(\text{vec}\, Y)'$. An example is the covariance matrix of the vec of a stochastic matrix: if $\text{Var}(\text{vec}\, X) = \text{E}(\text{vec}\, X - \text{E}\,\text{vec}\, X)(\text{vec}\, X - \text{E}\,\text{vec}\, X)' = K_{nm}(A \otimes B)$, then

$$\text{cov}(x_{ij}, x_{rs}) = a_{is} b_{jr},$$
$$i = 1,2,...,m;\ j = 1,2,...,n;\ r = 1,2,...,p;\ s = 1,2,...,q. \quad (2.37)$$

Another example is the matrix with second partial derivatives of a real-valued matrix function: if $\partial^2\phi/(\partial(\text{vec}\, X)\partial(\text{vec}\, X)') = K_{nm}(A \otimes B)$, then

$$\frac{\partial^2\phi}{\partial x_{ij}\partial x_{rs}} = a_{is} b_{jr},$$
$$i = 1,2,...,m;\ j = 1,2,...,n;\ r = 1,2,...,p;\ s = 1,2,...,q; \quad (2.38)$$

see section 4.4.

2.5 The duplication matrix

Let A be a $n \times n$-matrix and let $\text{v}(A)$ denote the $\frac{1}{2}n(n+1)$-vector that is obtained from $\text{vec}\, A$ by deleting all supradiagonal elements of A. If A is symmetric then $\text{v}(A)$ contains only the distinct elements of A. The duplication matrix D_n is the $n^2 \times \frac{1}{2}(n+1)$-matrix that transforms, for symmetric A, $\text{v}(A)$ into $\text{vec}\, A$:

$$D_n\, \text{v}(A) = \text{vec}\, A. \quad (2.39)$$

The Moore-Penrose inverse of the duplication matrix is

$$D_n^+ = (D_n' D_n)^{-1} D_n'. \quad (2.40)$$

It is easily seen that for symmetric A:

$$\text{v}(A) = D_n^+\, \text{vec}\, A. \quad (2.41)$$

An explicit expression for the duplication matrix is

$$D_n = \sum_{i=j}^{n} \sum_{j=1}^{n} (\text{vec}\, T_{ij}) u_{ij}', \quad (2.42)$$

where $T_{ii} = E_{ii}$, $T_{ij} = E_{ij} + E_{ji}$ $(i \neq j)$, E_{ij} is the $n \times n$-matrix with 1 as element (i, j) and 0 elsewhere, and $u_{ij} = \text{v}(E_{ij})$ (note that $E_{ij} = e_i e_j'$). Also

$$D_n D_n^+ = \tfrac{1}{2}(I_{n^2} + K_n), \quad (2.43)$$

$$D_n^+ D_n = I_{\frac{1}{2}n(n+1)}, \quad (2.44)$$

and

$$[D_n'(A \otimes A) D_n]^{-1} = D_n^+ (A^{-1} \otimes A^{-1}) D_n^{+\prime}. \quad (2.45)$$

The $n^2 \times n^2$-matrix $\frac{1}{2}(I_{n^2} + K_n)$ will be denoted by N_n, and plays an important part in distribution theory, especially normal distribution theory. There holds

$$N_n \operatorname{vec} A = \operatorname{vec} \tfrac{1}{2}(A + A'), \tag{2.46}$$

$$N_n(A \otimes A)N_n = N_n(A \otimes A) = (A \otimes A)N_n, \tag{2.47}$$

and

$$N_n = N_n' = N_n^2, \tag{2.48}$$

so that N_n is orthogonal and idempotent.

2.6 Diagonality

Let A be a square $n \times n$-matrix and define $\mathrm{w}(A)$ as the vector containing just the diagonal elements of A:

$$\mathrm{w}(A) = (a_{11}, a_{22}, ..., a_{nn})'. \tag{2.49}$$

We define the $n \times n^2$-matrix G_n as the matrix that transforms for diagonal A, $\mathrm{w}(A)$ into $\operatorname{vec}(A)$:

$$G_n' w(A) = \operatorname{vec} A. \tag{2.50}$$

An explicit expression for G_n is

$$G_n = \sum_{i=1}^{n} e_i (\operatorname{vec} E_{ii})', \tag{2.51}$$

where e_i is the n-vector with 1 as element i and 0 elsewhere, and E_{ii} is the $n \times n$-matrix with 1 as element (i, i) and 0 elsewhere. It can be shown that

$$G_n K_n = G_n N_n = G_n,$$

$$G_n G_n' = I_n, \tag{2.52}$$

and

$$G_n^+ = G_n'. \tag{2.53}$$

The matrix G_n eliminates from $\operatorname{vec} A$ the off-diagonal elements, since for every square matrix A,

$$G_n(\operatorname{vec} A) = \mathrm{w}(A). \tag{2.54}$$

and

$$G_n' \,\mathrm{w}(A) = G_n' G_n(\operatorname{vec} A) = \operatorname{vec}(\operatorname{dg} A), \tag{2.55}$$

where $\operatorname{dg} A$ is the diagonal matrix containing the diagonal elements of A. The matrix G_n converts a Kronecker product into a Hadamard product:

$$G_n(A \otimes B)G_n' = A \odot B, \tag{2.56}$$

where A and B are matrices of the same size and the Hadamard product of two matrices is defined as their element-wise product, *i.e.* $(A \odot B)_{ij} = a_{ij} b_{ij}$.

Let X be an $m \times n$-matrix, y a p-vector, A an $n \times p$-matrix and B an $m \times p$-matrix, such that $(\operatorname{vec} X)y' = (A \otimes B)G_p'$. Then

$$x_{ij}y_r = a_{jr}b_{ir}, \; i = 1, 2, ..., m; \; j = 1, 2, ..., n; \; r = 1, 2, ..., p. \tag{2.57}$$

Similar formulae hold for other matrices with the same structure as $(\operatorname{vec} X)y'$. An example is the covariance matrix of a stochastic vector and the vec of a stochastic matrix: if $\operatorname{Cov}(\operatorname{vec} X, y) = \operatorname{E}(\operatorname{vec} X - \operatorname{E}\operatorname{vec} X)(y - \operatorname{E} y)' = (A \otimes B)G_p'$, then

$$\operatorname{cov}(x_{ij}, y_r) = a_{jr}b_{ir}, \; i = 1, 2, ..., m; \; j = 1, 2, ..., n; \; r = 1, 2, ..., p. \tag{2.58}$$

Another example is the matrix with second partial cross derivatives of a vector and a real-valued matrix function: if $\partial^2\phi/(\partial(\operatorname{vec} X)\partial(y)') = (A \otimes B)G_p'$, then

$$\frac{\partial^2 \phi}{\partial x_{ij} \partial y_r} = a_{jr}b_{ir}, \; i = 1, 2, ..., m; \; j = 1, 2, ..., n; \; r = 1, 2, ..., p; \tag{2.59}$$

see section 4.4.

3 First-order differentiation

3.1 Differentiability of vector functions

Let f be a function from an open set $S \subset \mathbb{R}^m$ to $\mathbb{R}^n$; let $x^0 \in S$ and $u \in \mathbb{R}^m$ such that $x^0 + u \in S$. The function f is *differentiable* at x^0 if there exists a real $n \times m$-matrix A_{x^0}, depending on x^0 but not on u, such that

$$f(x^0 + u) = f(x^0) + A_{x^0}u + o(u), \tag{3.1}$$

where $o(u)$ is a function such that $\lim_{|u| \to 0} |o(u)|/|u| = 0$. The matrix A_{x^0} is called the *(first) derivative* of f at x^0; it is denoted by $\mathsf{D} f(x^0)$ or by $\partial f/\partial x' \rfloor_{x=x^0}$, and is called the *Jacobian matrix* of f at x^0, and if $m = n$, its determinant is called the *Jacobian* of f. The Jacobian matrix $\mathsf{D} f$ is equal to the matrix of partial derivatives, *i.e.* $\mathsf{D} f(x)_{ij} = \partial f_i/\partial x_j$. The linear function $\mathrm{d} f_{x^0} : \mathbb{R}^m \to \mathbb{R}^n$ defined by $\mathrm{d} f_{x^0}(u) = A_{x^0} \times u$ is called the *(first) differential* of f at x^0; instead of u we often write $\mathrm{d} x$, so that: $\mathrm{d} f = A_{x^0} \times (\mathrm{d} x)$. Alternatively if A is a matrix such that $\mathrm{d} f = A \,\mathrm{d} x$ then A is the derivative of f at x^0 and contains the partial derivatives. This one-to-one relation between differentials and derivatives is very useful, since differentials are relatively easy to manipulate.

From (3.1) we see that the differential corresponds to the linear part of the function, which can also be written as

$$y - y^0 = A_{x^0}(x - x^0),$$

where $y^0 = f(x^0)$. Therefore the differential of a function is the linearization of the function: it is the equation of the hyperplane through the origin that is parallel to the hyperplane tangent to the graph of f at x^0; so the linearized function can be written as

$$f(x) \doteq f(x^0) + A_{x^0}(x - x^0). \tag{3.2}$$

3.2 Differentiability of matrix functions

A matrix function F from an open set $S \subset \mathbb{R}^{m\times n}$ to $\mathbb{R}^{p\times q}$ is differentiable if $\operatorname{vec} F$ is differentiable, *i.e.* if there exists a real $pq \times mn$-matrix A, depending on X^0, such that

$$\operatorname{vec} F(X^0 + U) = \operatorname{vec} F(X^0) + A_{X^0}(\operatorname{vec} U) + \operatorname{vec} o(U), \tag{3.3}$$

where U is a $p \times q$-matrix such that $X^0 + U \in S$, and $\lim_{|U|\to 0}|o(U)|/|U| = 0$ with the norm of a matrix X defined by $|X| = (\operatorname{tr} X'X)^{\frac{1}{2}}$. The differential of F at X^0 is the $m \times n$-matrix function $\mathsf{d}\, F_{X^0}$ defined by $\operatorname{vec} \mathsf{d}\, F_{X^0}(U) = A_{X^0}(\operatorname{vec} U)$. The $pq \times mn$-matrix $\mathsf{D}(\operatorname{vec} F)$ is called the Jacobian matrix of F at X^0 and is denoted by $\mathsf{D}\, F(X^0)$ or by $\partial(\operatorname{vec} F)/\partial(\operatorname{vec} X)'\rfloor_{X=X^0}$.

If either F or X is a scalar then $\mathsf{D}\, F$ is a vector. It is then useful to define some other matrices that also contain the partial derivatives. If F is a scalar function of a matrix (*i.e.* $p = q = 1$), then we define the $m \times n$-matrix $\partial F(X)/\partial X$ implicitly by

$$\mathsf{D}\, F(X) = \left(\operatorname{vec} \frac{\partial F(X)}{\partial X}\right)', \tag{3.4}$$

i.e. $(\partial F/\partial X)_{ij} = \partial F/\partial x_{ij}$. If X is a scalar (*i.e.* $m = n = 1$), then we define the $p \times q$-matrix $\partial F(X)/\partial X$ implicitly by

$$\mathsf{D}\, F(X) = \operatorname{vec} \frac{\partial F(X)}{\partial X}, \tag{3.5}$$

i.e. $(\partial F/\partial X)_{ij} = \partial F_{ij}/\partial X$. In all other cases the only useful definition of derivative is the Jacobian matrix $\partial \operatorname{vec} F/\partial(\operatorname{vec} X)'$, because only then there exists a general chain rule and the determinant of the derivative equals the Jacobian. Note that if X is a vector (*i.e.* $n = 1$) and F a scalar (*i.e.* $p = q = 1$), then $\mathsf{D}\, F$ is a row vector and $\partial F/\partial X = (\mathsf{D}\, F)'$ is a column vector.

An explicit expression for the Jacobian matrix is

$$\mathsf{D}\, F = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\operatorname{vec} \frac{\partial F}{\partial x_{ij}}\right)(\operatorname{vec} H_{ij})', \tag{3.6}$$

where H_{ij} is the $m \times n$-matrix with 1 as element (i, j) and 0 elsewhere. If $n = 1$, then (3.6) simplifies to

$$\mathsf{D}\, F = \sum_{i=1}^{m}\left(\operatorname{vec} \frac{\partial F}{\partial x_i}\right) e_i', \tag{3.7}$$

where e_i is the m-vector with 1 as element i and 0 elsewhere.

3.3 Chain rule

Let S be an open set in $\mathbb{R}^n$ and let $f : S \to \mathbb{R}^m$ be differentiable at a point x^0 in S. Let T be a subset of $\mathbb{R}^m$ such that $f(x) \in T$ for all $x \in S$ and let $g : T \to \mathbb{R}^p$ be differentiable at a point $y^0 = f(x^0) \in T$. Then the composite function $h = g \circ f : S \to \mathbb{R}^p$ defined by $h(x) = g[f(x)]$ is differentiable at x^0, and there holds $\mathsf{D}\, h(x^0) = \mathsf{D}\, g(y^0)\, \mathsf{D}\, f(x^0)$ and $\mathsf{d}\, h_{x^0}(u) = \mathsf{d}\, g_{y^0}[\mathsf{d}\, f_{x^0}(u)]$.

3.4 Properties of differentials

Let A be a matrix of constants, F and G matrix functions, and α a real scalar. Then, assuming that all differentials, products, inverses, etc. exist, we have

$$\mathrm{d}\,A = 0, \tag{3.8}$$

$$\mathrm{d}(\alpha F) = \alpha\,\mathrm{d}\,F, \tag{3.9}$$

$$\mathrm{d}(F + G) = \mathrm{d}\,F + \mathrm{d}\,G, \tag{3.10}$$

$$\mathrm{d}(FG) = (\mathrm{d}\,F)G + F(\mathrm{d}\,G), \tag{3.11}$$

$$\mathrm{d}(F \otimes G) = (\mathrm{d}\,F) \otimes G + F \otimes (\mathrm{d}\,G), \tag{3.12}$$

$$\mathrm{d}(F') = (\mathrm{d}\,F)', \tag{3.13}$$

$$\mathrm{d}(\operatorname{vec} F) = \operatorname{vec}(\mathrm{d}\,F), \tag{3.14}$$

$$\mathrm{d}(\operatorname{tr} F) = \operatorname{tr}(\mathrm{d}\,F), \tag{3.15}$$

$$\mathrm{d}\,F^{-1} = -F^{-1}(\mathrm{d}\,F)F^{-1}, \tag{3.16}$$

$$\mathrm{d}|F| = \operatorname{tr}(F^{\#}\,\mathrm{d}\,F), \tag{3.17}$$

where $F^{\#}$ is the adjoint matrix (*i.e.* the transpose of the matrix with cofactors) of F. In particular, at points where F has full rank:

$$\mathrm{d}|F| = |F|\operatorname{tr}(F^{-1}\,\mathrm{d}\,F). \tag{3.18}$$

The use of differentials makes it unnecessary to remember many matrix derivatives, since they follow easily from the above properties.

Formula (3.16) is easily proved by taking the differential of $FF^{-1} = I$, and rearranging.

As an example of the chain rule we will prove (3.17). Define the function $g : \mathbb{R}^{m\times n} \to \mathbb{R}$ by $g(Y) = |Y|$. Note that g is a function without restrictions and that all variables y_{ij} are independent. Then $|F|$ is the composite of g and F. Expanding $|Y|$ along the i-th row we get $|Y| = \sum_j y_{ij}|Y_{ij}|$, where $|Y_{ij}|$ is the cofactor of y_{ij}. Since Y_{ij} is independent of y_{ij}, there holds $\partial|Y|/\partial y_{ij} = Y_{ij}$ and thus $\mathrm{d}|Y| = \sum_i \sum_j Y_{ij}\,\mathrm{d}\,y_{ij} = \operatorname{tr}(Y^{\#}\,\mathrm{d}\,Y)$. By the chain rule we then have $\mathrm{d}\,F = \operatorname{tr}(F^{\#}\,\mathrm{d}\,F)$. Note that (3.17) and (3.18) hold independently of any restrictions, such as symmetry, on F.

3.5 Examples

Example 3.1. Some special cases for real functions of one variable are

$$\mathrm{d}\,x^n = nx^{n-1}\,\mathrm{d}\,x \tag{3.19}$$

$$\mathrm{d}\,e^x = e^x\,\mathrm{d}\,x, \tag{3.20}$$

$$\mathrm{d}\log x = \frac{\mathrm{d}\,x}{x} \qquad (x > 0). \tag{3.21}$$

Example 3.2. For the real function $f(x, y) = x^2 + 2xy - y^2$ we have

$$\begin{aligned} \mathrm{d} f &= \mathrm{d}(x^2) + 2\,\mathrm{d}(xy) - \mathrm{d}(y^2) = 2x\,\mathrm{d}\,x + 2(y\,\mathrm{d}\,x + x\,\mathrm{d}\,y) - 2y\,\mathrm{d}\,y \\ &= 2(x + y)\,\mathrm{d}\,x + 2(x - y)\,\mathrm{d}\,y. \end{aligned}$$

Example 3.3. For the real function $f(x, y) = x^\alpha y^\beta \quad (x > 0, y > 0)$ we have

$$\mathrm{d}\log f = \alpha\,\mathrm{d}\log x + \beta\,\mathrm{d}\log y.$$

Example 3.4. (Implicit differentiation) Consider the equation

$$f(x, y) = 0 \tag{3.22}$$

Suppose (3.22) holds on an open set in $\mathbb{R}^2$ in the sense that there exists a function $g : \mathbb{R} \to \mathbb{R}$ implicitly defined by $f[x, g(x)] = 0$. Define the function $h : \mathbb{R} \to \mathbb{R}$ by $x \mapsto f[x, g(x)]$ and the function $\phi : \mathbb{R}^2 \to \mathbb{R}$ by $x \mapsto [x, g(x)]$. Then h is the composite of f and ϕ, so that by the chain rule we have

$$h'(x) = \mathsf{D}\,g(x) = \mathsf{D}\,\phi \times \mathsf{D}\,f = \begin{pmatrix} 1 \\ g'(x) \end{pmatrix} \begin{pmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{pmatrix} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} g'(x).$$

On the other hand, from (3.22) we have $h'(x) = 0$, so that

$$\frac{\partial y}{\partial x} = g'(x) = -\frac{\partial f/\partial x}{\partial f/\partial y}. \tag{3.23}$$

This result also follows by the chain rule for differentials, since it implies

$$\frac{\partial f}{\partial x}\,\mathrm{d}\,x + \frac{\partial f}{\partial y}\frac{\partial y}{\partial x}\,\mathrm{d}\,x = 0,$$

from which (3.23) follows after rearranging and dividing through by $\mathrm{d}\,x$.

Example 3.5. $\mathrm{d}(Ax) = A(\mathrm{d}\,x)$ and so

$$\mathsf{D}(Ax) = \frac{\partial Ax}{\partial x'} = A. \tag{3.24}$$

Example 3.6. $\mathrm{d}(x'Ax) = x'A(\mathrm{d}\,x) + (\mathrm{d}\,x)'Ax = x'(A + A')(\mathrm{d}\,x)$ and so

$$\mathsf{D}(x'Ax) = \frac{\partial x'Ax}{\partial x'} = x'(A + A') \tag{3.25}$$

and

$$\frac{\partial x'Ax}{\partial x} = (A + A')x. \tag{3.26}$$

Example 3.7. $\mathrm{d}(y'Az) = y'(\mathrm{d}\,A)z = (z' \otimes y')(\mathrm{vec}\,\mathrm{d}\,A)$, so that

$$\mathsf{D}(y'Az) = \frac{\partial y'Az}{\partial(\mathrm{vec}\,A)'} = z' \otimes y', \tag{3.27}$$

and, with (2.19),

$$\frac{\partial y'Az}{\partial A} = yz'. \tag{3.28}$$

Example 3.8. $\mathrm{d}(AXB) = A(\mathrm{d}\,X)B$, so that $\mathrm{d}(\mathrm{vec}\,AXB) = (B' \otimes A)(\mathrm{vec}\,\mathrm{d}\,X)$ and therefore

$$\mathsf{D}(AXB) = \frac{\partial\,\mathrm{vec}\,AXB}{\partial(\mathrm{vec}\,X)'} = B' \otimes A. \tag{3.29}$$

Example 3.9. An application of (3.16) is

$$\mathsf{D}\operatorname{vec} X^{-1} = \frac{\partial \operatorname{vec} X^{-1}}{\partial(\operatorname{vec} X)'} = -X'^{-1} \otimes X^{-1}, \tag{3.30}$$

and, using (2.29), we get

$$\frac{\partial x^{ij}}{\partial x_{rs}} = -x^{ir} x^{sj}.$$

Example 3.10.

$$\mathsf{D}\operatorname{vec}(X \otimes Y) = \frac{\partial \operatorname{vec}(X \otimes Y)}{\partial(\operatorname{vec} Z)'} = ?????? \tag{3.31}$$

Example 3.11. An application of (3.18) is: $\mathrm{d}\log|A| = |A|^{-1}\,\mathrm{d}|A| = \operatorname{tr}(A^{-1}\,\mathrm{d}A) = (\operatorname{vec} A'^{-1})'(\mathrm{d}\operatorname{vec} A)$, and so if A is a function of a scalar α, there holds

$$\mathrm{d}\log|A| = (\operatorname{vec} A'^{-1})' \frac{\partial \operatorname{vec} A}{\partial \alpha}\,\mathrm{d}\,\alpha = \operatorname{tr} A'^{-1} \frac{\partial A}{\partial \alpha}\;\mathrm{d}\,\alpha, \tag{3.32}$$

and therefore

$$\frac{\partial \log|A|}{\partial \alpha} = \operatorname{tr} A'^{-1} \frac{\partial A}{\partial \alpha}. \tag{3.33}$$

Example 3.12. Let X be an $m \times n$-matrix; then $\mathrm{d}(X'X) = (\mathrm{d}\,X')X + X'(\mathrm{d}\,X)$, so that

$$\begin{aligned}
\mathrm{d}\operatorname{vec}(X'X) &= (X' \otimes I_n)(\operatorname{vec}\mathrm{d}\,X') + (I_n \otimes X')(\operatorname{vec}\mathrm{d}\,X) \\
&= (X' \otimes I_n)K_{mn}(\operatorname{vec}\mathrm{d}\,X) + (I_n \otimes X')(\operatorname{vec}\mathrm{d}\,X) \\
&= K_n(I_n \otimes X')(\operatorname{vec}\mathrm{d}\,X) + (I_n \otimes X')(\operatorname{vec}\mathrm{d}\,X) \\
&= 2N_n(I_n \otimes X')(\operatorname{vec}\mathrm{d}\,X);
\end{aligned}$$

therefore

$$\mathsf{D}\operatorname{vec} X'X = \frac{\partial \operatorname{vec} X'X}{\partial(\operatorname{vec} X)'} = 2N_n(I_n \otimes X').$$

Similarly,

$$\mathsf{D}\operatorname{vec} XX' = \frac{\partial \operatorname{vec} XX'}{\partial(\operatorname{vec} X)'} = 2N_m(X \otimes I_m). \tag{3.34}$$

Example 3.13. Let X be an $m \times n$-matrix and A and $m \times m$-matrix. Then $\mathrm{d}\operatorname{tr} X'AX = \operatorname{tr}(\mathrm{d}\,X)'AX + \operatorname{tr} X'A(\mathrm{d}\,X) = 2\operatorname{tr} X'A(\mathrm{d}\,X) = 2(\operatorname{vec} A'X)'(\operatorname{vec}\mathrm{d}\,X)$, so that

$$\mathsf{D}\operatorname{tr} X'AX = \frac{\partial \operatorname{tr} X'AX}{\partial(\operatorname{vec} X)'} = 2(\operatorname{vec} A'X)',$$

and

$$\frac{\partial \operatorname{tr} X'AX}{\partial X} = 2A'X.$$

Example 3.14. Let X be an $n \times n$-matrix. Then $\mathrm{d}\operatorname{tr} X^2 = \operatorname{tr}[(\mathrm{d}\,X)X + X(\mathrm{d}\,X)] = 2\operatorname{tr} X(\mathrm{d}\,X) = 2(\operatorname{vec} X')'(\operatorname{vec}\mathrm{d}\,X)$, so that

$$\mathsf{D}\operatorname{tr} X^2 = \frac{\partial \operatorname{tr} X^2}{\partial(\operatorname{vec} X)'} = 2(\operatorname{vec} X)',$$

and

$$\frac{\partial \operatorname{tr} X^2}{\partial X} = 2X.$$

Example 3.15. Let X be an $n{\times}n$-matrix. Then $\mathrm{d}\log|X| = \operatorname{tr} X^{-1}(\mathrm{d}\,X) = (\operatorname{vec} X'^{-1})'(\operatorname{vec}\mathrm{d}\,X)$, so that

$$\mathsf{D}\log|X| = \frac{\partial \log|X|}{\partial(\operatorname{vec} X)'} = (\operatorname{vec} X'^{-1})',$$

and

$$\frac{\partial \log|X|}{\partial X} = X'^{-1}.$$

4 Second-order differentiation

4.1 Twice-differentiability

Let f be a function from an open set $S \subset \mathbb{R}^m$ to $\mathbb{R}^n$; let $x \in S$, $u \in \mathbb{R}^m$, and let f be differentiable at x. The function f is *twice differentiable* at x if $\mathsf{D} f$ is differentiable at x, *i.e.* if there exists a real $mn \times m$-matrix B, depending on x but not on u, such that

$$\operatorname{vec} \mathsf{D} f(x+u) = \operatorname{vec} \mathsf{D} f(x) + B(x)u + o(u), \tag{4.1}$$

where $o(u)$ is a function such that $\lim_{|u|\to 0}|o(u)|/|u| = 0$. The matrix B is the derivative of $\operatorname{vec} \mathsf{D} f$ at x, *i.e.* $B(x) = \mathsf{D}[\mathsf{D} f] = \partial[\operatorname{vec} \mathsf{D} f(x)]/\partial x'$; note that for $\mathsf{D} f$ to be differentiable, it must exist in a neighborhood of x, *i.e.* f must be differentiable in a neighborhood of x.

It will be easier to work with the $mn \times m$-matrix $\mathsf{H} f(x) = \mathsf{D}[\mathsf{D} f]' = K_{mn}B(x) = \partial \operatorname{vec}[\mathsf{D} f(x)]'/\partial x'$; this matrix is called the *Hessian matrix* of f at x, and if $n = 1$, its determinant is called the *Hessian* of f. The Hessian matrix is equal to the Hessian matrices of the component functions of f stacked below each other:

$$\mathsf{H} f(x) = \begin{pmatrix} \mathsf{H} f_1(x) \\ \mathsf{H} f_2(x) \\ \vdots \\ \mathsf{H} f_n(x) \end{pmatrix}. \tag{4.2}$$

It can be shown that the component Hessian matrices of f at x are symmetric, *i.e.* $\mathsf{H} f_i(x) = (\mathsf{H} f_i(x))'$ if f is twice differentiable at x (Dieudonné, 1960, section 8.12). Note that $\mathsf{H} f$ also exists if the partial derivatives $\partial f/\partial x_i$ are differentiable and $\mathsf{D} f$ is not differentiable; then the $\mathsf{H} f_i$ are not necessarily symmetric (in this case a sufficient condition for $\mathsf{H} f_i$ to be symmetric is that each partial derivative is continuous).

4.2 The second differential

The second differential is the differential of the first differential:

$$\mathrm{d}^2 f = \mathrm{d}(\mathrm{d} f), \tag{4.3}$$

where we consider $\mathrm{d} f$ as a function of x only, holding $\mathrm{d} x$ constant. The second differential exists if and only if f is twice differentiable, since

$$\begin{aligned}
\mathrm{d}^2 f = \mathrm{d}(\mathrm{d} f) &= \mathrm{d}[\mathsf{D} f(x)\,\mathrm{d} x] = [(\mathrm{d} x)' \otimes I_n] \operatorname{vec}[\mathrm{d}\,\mathsf{D} f(x)] \\
&= [(\mathrm{d} x)' \otimes I_n]\frac{\partial \operatorname{vec} \mathsf{D} f(x)}{\partial x'}(\mathrm{d} x) = [(\mathrm{d} x)' \otimes I_n]B(x)(\mathrm{d} x) \\
&= [(\mathrm{d} x)' \otimes I_n]K_{mn}\,\mathsf{H} f(x)(\mathrm{d} x) = [I_n \otimes (dx)']\,\mathsf{H} f(x)(\mathrm{d} x) \\
&= \begin{pmatrix} (\mathrm{d} x)'\,\mathsf{H} f_1(x)(\mathrm{d} x) \\ (\mathrm{d} x)'\,\mathsf{H} f_2(x)(\mathrm{d} x) \\ \vdots \\ (\mathrm{d} x)'\,\mathsf{H} f_n(x)(\mathrm{d} x) \end{pmatrix}
\end{aligned}$$

Thus $\mathrm{d}^2 f$ is a n-vector of quadratic forms in $\mathsf{H}\, f_i(x)$.

Alternatively, if f is twice differentiable and $B(x)$ is an $mn \times m$-matrix such that for all $\mathrm{d}\, x$

$$d^2 f = [I_n \otimes (dx)']B(x)(dx), \tag{4.4}$$

then

$$\mathsf{H}\, f(x) = \tfrac{1}{2}\{B(x) + [B'(x)]_v\}, \tag{4.5}$$

where

$$B(x) = \begin{pmatrix} B_1(x) \\ B_2(x) \\ \vdots \\ B_n(x) \end{pmatrix}, \quad [B'(x)]_v = \begin{pmatrix} B_1'(x) \\ B_2'(x) \\ \vdots \\ B_n'(x) \end{pmatrix}, \tag{4.6}$$

and each B_i is a $m \times m$-matrix. For example, if $n = 1$ then

$$d^2 f = (dx)' B(x)(dx), \tag{4.7}$$

for all $\mathrm{d}\, x$, if and only if

$$\mathsf{H}\, f(x) = \tfrac{1}{2}[B(x) + [B'(x)], \tag{4.8}$$

4.3 Matrix functions

A matrix function F from an open set $S \subset \mathbb{R}^{m\times n}$ to $\mathbb{R}^{p\times q}$ is twice differentiable if $\operatorname{vec} F$ is twice differentiable. The Hessian matrix $\mathsf{H}\, F$ of F is the $mnpq \times mn$-matrix

$$\mathsf{H}\, F(X) = \begin{pmatrix} \mathsf{H}\, F_{11}(\operatorname{vec} X) \\ \vdots \\ \mathsf{H}\, F_{p1}(\operatorname{vec} X) \\ \vdots \\ \vdots \\ \mathsf{H}\, F_{1q}(\operatorname{vec} X) \\ \vdots \\ \mathsf{H}\, F_{pq}(\operatorname{vec} X) \end{pmatrix},$$

where each $\mathsf{H}_{k\ell}$ is a $mn \times mn$-matrix. The second differential of F is defined as $\mathrm{d}^2\, F(X; \mathrm{d}\, X) = \mathrm{d}[\mathrm{d}\, F(X; \mathrm{d}\, X)]$, *i.e.* $\operatorname{vec} \mathrm{d}^2\, F(X; \mathrm{d}\, X) = \mathrm{d}^2[\operatorname{vec} F(\operatorname{vec} X; \operatorname{vec} \mathrm{d}\, X)]$. If F is twice differentiable then

$$\operatorname{vec} \mathrm{d}^2\, F = [I_{pq} \otimes (\operatorname{vec} \mathrm{d}\, X)']B(X)(\operatorname{vec} \mathrm{d}\, X)] \tag{4.9}$$

for every $\mathrm{d}\, X \in \mathbb{R}^{p\times q}$, if and only if

$$\mathsf{H}\, F(X) = \tfrac{1}{2}\{B(X) + [B'(X)]_v\}. \tag{4.10}$$

An explicit expression for the Hessian matrix is

$$\mathsf{H}\, F = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{\ell=1}^{n}\left(\operatorname{vec} \frac{\partial^2 F}{\partial x_{k\ell}\partial x_{ij}}\right) \otimes (\operatorname{vec} H_{ij}) \otimes (\operatorname{vec} H_{k\ell})' \tag{4.11}$$

where H_{ij} is the $m \times n$-matrix with 1 as element (i, j) and 0 elsewhere. If $n = 1$, then (4.11) simplifies to

$$\mathsf{H} F = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\operatorname{vec} \frac{\partial^2 F}{\partial x_j \partial x_i} \right) \otimes e_i e_j', \tag{4.12}$$

where e_i is the m-vector with 1 as element i and 0 elsewhere.

4.4 Examples

Example 4.1. (Example 3.2 continued). The second differential of $f(x, y) = x^2 + 2xy - 2y^2$ is

$$\mathrm{d}(\mathrm{d} f) = \mathrm{d}[2(x + y)\,\mathrm{d}\,x + 2(x - y)\,\mathrm{d}\,y] = 2(\mathrm{d}\,x)^2 + 4(\mathrm{d}\,x)(\mathrm{d}\,y) - 2(\mathrm{d}\,y)^2$$

so that

$$B(x) = \begin{pmatrix} 2 & 4 \\ 0 & -2 \end{pmatrix}.$$

Thus the Hessian matrix is

$$\mathsf{H} f = \tfrac{1}{2}[B(x) + B'(x)] = \begin{pmatrix} 2 & 2 \\ 2 & -2 \end{pmatrix}.$$

Example 4.2. Often the second differential of a real-valued matrix function has the form $\operatorname{tr} B(\mathrm{d}\,X)'C(\mathrm{d}\,X)$ or $\operatorname{tr} B(\mathrm{d}\,X)C(\mathrm{d}\,X)$. Then the following result is useful.

Let $\phi : S \subset \mathbb{R}^{m \times n} \to \mathbb{R}$ be a real-valued matrix function.

a. Suppose $d^2\phi = \operatorname{tr} B(dX)'C(dX)$ with B an $n \times n$-matrix and C an $m \times m$-matrix. Then $d^2\phi = (\mathrm{d} \operatorname{vec} X)'(B' \otimes C)(\mathrm{d} \operatorname{vec} X)$, so that

$$\mathsf{H}\,\phi(X) = \tfrac{1}{2}(B' \otimes C + B \otimes C'), \tag{4.13}$$

and

$$\frac{\partial^2 \phi(x)}{\partial x_{ij} \partial x_{rs}} = \tfrac{1}{2}(b_{sj} c_{ir} + b_{js} c_{ri}),$$
$$i, r = 1, 2, ..., m; \;\; j, s = 1, 2, ..., n. \tag{4.14}$$

b. Suppose $\mathrm{d}^2\,\phi = \operatorname{tr} B(\mathrm{d}\,X)C(\mathrm{d}\,X)$ with B and C $n \times m$-matrices. Then $d^2\phi = (\mathrm{d} \operatorname{vec} X')'(B' \otimes C)(\mathrm{d} \operatorname{vec} X) = (\mathrm{d} \operatorname{vec} X)' K_{nm} (B' \otimes C)(\mathrm{d} \operatorname{vec} X)$, so that

$$\mathsf{H}\,\phi(X) = \tfrac{1}{2} K_{nm}(B' \otimes C + C' \otimes B), \tag{4.15}$$

and

$$\frac{\partial^2 \phi(x)}{\partial x_{ij} \partial x_{rs}} = \tfrac{1}{2}(b_{si} c_{jr} + b_{jr} c_{si}),$$
$$i, r = 1, 2, ..., m; \;\; j, s = 1, 2, ..., n. \tag{4.16}$$

Equations (4.14) and (4.16) can be derived from respectively (4.13) and (4.15) using (2.29) and (2.38).

Some special cases are:

a. (Example 3.13 continued) For $\phi(X) = \operatorname{tr} X'AX$ with A an $m \times m$-matrix, we get $\mathrm{d}^2\,\phi(X) = 2\operatorname{tr}(\mathrm{d}\,X)'A(\mathrm{d}\,X)$, so that

$$\mathsf{H}(\operatorname{tr} X'AX) = I_n \otimes (A + A'), \tag{4.17}$$

and

$$\frac{\partial^2 \operatorname{tr} X'AX}{\partial x_{ij}\partial x_{rs}} = \delta_{js}(a_{ir} + a_{ri}); \qquad i, r = 1, 2, ..., m;\ \ j, s = 1, 2, ..., n. \tag{4.18}$$

where δ_{js} is the Kronecker delta ($\delta_{jj} = 1$, and $\delta_{js} = 0$ for $j \neq s$).

b. (Example 3.14 continued) For $\phi(X) = \operatorname{tr} X^2$ we get $\mathrm{d}\,\phi(X) = \operatorname{tr}[(\mathrm{d}\,X)X + X(\mathrm{d}\,X)]$ and $\mathrm{d}^2\,\phi(X) = 2\operatorname{tr}(\mathrm{d}\,X)^2$, so that

$$\mathsf{H}(\operatorname{tr} X^2) = K_n(I_n \otimes I_n + I_n \otimes I_n) = 2K_n, \tag{4.19}$$

and

$$\frac{\partial^2 \operatorname{tr} X^2}{\partial x_{ij}\partial x_{rs}} = 2\delta_{is}\delta_{jr}, \qquad i, r = 1, 2, ..., m;\ \ j, s = 1, 2, ..., n. \tag{4.20}$$

c. (Example 3.15 continued) For $\phi(X) = \log|X|$, we get $\mathrm{d}\,\phi(X) = \operatorname{tr} X^{-1}(\mathrm{d}\,X)$ and $\mathrm{d}^2\,\phi(X) = -\operatorname{tr} X^{-1}(\mathrm{d}\,X)X^{-1}(\mathrm{d}\,X)$ and therefore

$$\mathsf{H}\log|X| = -K_n(X'^{-1} \otimes X^{-1}), \tag{4.21}$$

and

$$\frac{\partial^2 \log|X|}{\partial x_{ij}\partial x_{rs}} = -\, x^{si}x^{jr}, \qquad i, r = 1, 2, ..., m;\ \ j, s = 1, 2, ..., n. \tag{4.22}$$

Note that if X is a symmetric positive definite matrix, then $\mathrm{d}^2 \log|X| = -(\operatorname{vec}\mathrm{d}\,X)'(X^{-1} \otimes X^{-1})(\operatorname{vec}\mathrm{d}\,X) < 0$, so that $\log|X|$ is a strictly concave function on the space of positive definite matrices.

Example 4.3. (Example 3.12 continued) Consider the matrix function $F : \mathbb{R}^n \to \mathbb{R}^{n\times n}$, defined by $F(x) = \frac{1}{2}xx'$, where x is a n-vector. There holds $\mathrm{d}\,F(x) = \frac{1}{2}[x(\mathrm{d}\,x)' + (\mathrm{d}\,x)x']$ so that, cf. (3.34),

$$\mathsf{D}\,F(x) = 2N_n(x \otimes I_n)$$

and $\mathrm{d}^2\,F(x) = (\mathrm{d}\,x)(\mathrm{d}\,x)'$, so that $\mathrm{d}^2 \operatorname{vec} F(x) = \operatorname{vec}(\mathrm{d}\,x)(\mathrm{d}\,x)' = (I_n \otimes \mathrm{d}\,x)\,\mathrm{d}\,x$. Now, $\mathrm{d}\,x = [I_n \otimes (\mathrm{d}\,x)'](\operatorname{vec} I_n)$, so that

$$I_n \otimes \mathrm{d}\,x = I_n \otimes [I_n \otimes (\mathrm{d}\,x)'](\operatorname{vec} I_n)$$

$$= [I_n \otimes I_n \otimes (\mathrm{d}\,x)'][I_n \otimes (\operatorname{vec} I_n)]. \tag{4.23}$$

Therefore

$$\mathrm{d}^2 \operatorname{vec} F(x) = [I_{n^2} \otimes (\mathrm{d}\,x)'](I_n \otimes \operatorname{vec} I_n)\,\mathrm{d}\,x,$$

and

$$\mathsf{H}\,F(x) = \tfrac{1}{2}\{I_n \otimes \operatorname{vec} I_n + [(I_n \otimes \operatorname{vec} I_n)']_v\}. \tag{4.24}$$

There holds

$$I_n \otimes \operatorname{vec} I_n = \begin{pmatrix} \operatorname{vec} I_n & 0 & \dots & 0 \\ 0 & \operatorname{vec} I_n & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \operatorname{vec} I_n \end{pmatrix} = \begin{pmatrix} E_{11} \\ E_{12} \\ \vdots \\ E_{1n} \\ E_{21} \\ \vdots \\ \vdots \\ E_{nn} \end{pmatrix}, \tag{4.25}$$

so that

$$[(I_n \otimes \operatorname{vec} I_n)']_v = \begin{pmatrix} E'_{11} \\ E'_{12} \\ \vdots \\ E'_{1n} \\ E'_{21} \\ \vdots \\ \vdots \\ E'_{nn} \end{pmatrix} = \begin{pmatrix} E_{11} \\ E_{21} \\ \vdots \\ E_{n1} \\ E_{12} \\ \vdots \\ \vdots \\ E_{nn} \end{pmatrix} = (K_n \otimes I_n)(I_n \otimes \operatorname{vec} I_n). \tag{4.26}$$

Therefore

$$\begin{aligned} \mathsf{H}\,F(x) &= [\tfrac{1}{2}(I_{n^2} + K_n) \otimes I_n](I_n \otimes \operatorname{vec} I_n) \\ &= (N_n \otimes I_n)(I_n \otimes \operatorname{vec} I_n). \end{aligned} \tag{4.27}$$

We can also use the explicit expression (4.12), which in this case is more straightforward. There holds

$$\frac{\partial^2 F(x)}{\partial x_j \partial x_i} = \tfrac{1}{2}(e_i e'_j + e_j e'_i), \tag{4.28}$$

and thus

$$\begin{aligned} \mathsf{H}\,F(x) &= \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}[\operatorname{vec}(e_i e'_j + e_j e'_i)] \otimes e_i e'_j \\ &= \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(e_j \otimes e_i \otimes e'_j \otimes e_i + e_i \otimes e_j \otimes e'_j \otimes e_i) \\ &= \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}[(e_j \otimes e'_j \otimes e_i \otimes e_i + (K_n \otimes I_n)(e_j \otimes e'_j \otimes e_i \otimes e_i)] \\ &= [\tfrac{1}{2}(I_{n^2} + K_n) \otimes I_n](I_n \otimes \operatorname{vec} I_n) \\ &= (N_n \otimes I_n)(I_n \otimes \operatorname{vec} I_n). \end{aligned} \tag{4.29}$$

5 Linearization of the regression estimator

Design-based sampling variances of non-linear statistics are often calculated by means of a linear approximation obtained by a Taylor expansion; examples are the variances of the general regression coefficient estimator and the regression estimator. The linearizations usually need some complicated differentiations. In this section, taken from Zeelenberg (1997), it is shown how matrix calculus can simplify these derivations, to the extent that even the Taylor expansion of the regression coefficient estimator can be derived in one line, which should be compared with the nearly one page that Särndal et al (1992, pp. 205-6) need. To be honest, the use of matrix calculus requires some more machinery to be set up, which is not needed for traditional methods. However this set-up can be regarded as an investment: once it has been learned, it can be used fruitfully in many other applications. See also Binder (1996) for applications similar to those of this section.

The π-estimator (Horvitz-Thompson estimator) of the finite population regression coefficient (cf. Särndal et al, 1992, section 5.10) is

$$\hat{B} = \hat{T}^{-1}\hat{t}, \tag{5.1}$$

where

$$\hat{T} = \sum_{k \in s} \frac{x_k x_k'}{\pi_k},$$
$$\hat{t} = \sum_{k \in s} \frac{x_k y_k}{\pi_k},$$

y_k is the variable of interest for individual k, x_k is the vector with the auxiliary variables for individual k, π_k is the inclusion probability for individual k, and s denotes the sample. Taking the total differential of (5.1), and evaluating at the point where $\hat{T} = T$, $\hat{t} = t$, we get

$$\mathrm{d}\,\hat{B} = -T^{-1}(\mathrm{d}\,\hat{T})T^{-1}t + T^{-1}(\mathrm{d}\,\hat{t}). \tag{5.2}$$

Because of the connection between differentials and linear approximation, as given in equation (3.2), it immediately follows that (5.2) corresponds to the linearization of the regression coefficient estimator:

$$\hat{B} \doteq B - T^{-1}(\hat{T} - T)T^{-1}t + T^{-1}(\hat{t} - t) = B + T^{-1}(\hat{t} - \hat{T}B),$$

where $B = T^{-1}t$.

The regression estimator of a population total is (cf. Särndal et al, 1992, section 6.6)

$$\hat{t}_{yr} = \hat{t}_{y\pi} + (t_x - \hat{t}_{x\pi})'\hat{B}, \tag{5.3}$$

where $\hat{t}_{y\pi}$ is the π-estimator of the variable of interest, t_x is the vector with the population totals of the auxiliary variables, $\hat{t}_{x\pi}$ is the vector with the π-estimators of the auxiliary variables, and $\hat{B}$ is the estimator of the regression coefficient of the auxiliary variables on the variable of interest. Taking the total differential of (5.3), and evaluating at the point where $\hat{t}_{y\pi} = t_y$, $\hat{t}_{x\pi} = t_x$, and $\hat{B} = B$, we get the linear approximation of the regression estimator

$$d\,\hat{t}_{yr} = d\,\hat{t}_{y\pi} - (d\,\hat{t}_{x\pi})'B,$$

so that

$$\hat{t}_{yr} \doteq t_y + \hat{t}_{y\pi} - t_y + (t_x - \hat{t}_{x\pi})'B = \hat{t}_{y\pi} + (t_x - \hat{t}_{x\pi})'B.$$

Note that for the linearization of the regression estimator we do not need that of the regression coefficient estimator B.

6 Maximum-likelihood estimation of the multivariate linear model

6.1 Introduction

This section is a generalization of section 15.8 of Magnus and Neudecker (1988) to the case where each equation may have different explanatory variables.

6.2 The model

Consider the model

$$y_{ij} = x'_{ij}\beta_i + \epsilon_{ij},\ i = 1, 2, ..., m,\ j = 1, 2, ..., n, \tag{6.1}$$

where x_{ij} is a k_i-vector and ϵ_{ij} is a random variable. For a given j we can stack the equations (6.1) as

$$y_j = X_j\beta + \epsilon_j,\ j = 1, 2, ..., n, \tag{6.2}$$

where $y_j = (y_{1j}, y_{2j}, ..., y_{mj})'$, $\beta = (\beta'_1, \beta'_2, ..., \beta'_m)'$, $\epsilon_j = (\epsilon_{1j}, \epsilon_{2j}, ..., \epsilon_{mj})'$, and

$$X_j = \begin{pmatrix} x'_1 & 0 & \dots & 0 \\ 0 & x'_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & x'_m \end{pmatrix}. \tag{6.3}$$

It is assumed that

$$\epsilon_j \sim \mathcal{N}_m(0, \Omega), \tag{6.4}$$

where Ω is a positive definite $m \times m$-matrix, and that ϵ_j and ϵ_s $(s \neq t)$ are independent. Thus

$$y_j \sim \mathcal{N}_m(X\beta, \Omega), \tag{6.5}$$

and y_j and y_s $(s \neq j)$ are independent. The log-likelihood of the model (6.2) is therefore

$$\mathcal{L} = -\tfrac{1}{2}\log 2\pi - \tfrac{1}{2}n\log|\Omega| - \tfrac{1}{2}\sum_{j=1}^{n}\epsilon'_j\Omega^{-1}\epsilon_j. \tag{6.6}$$

6.3 First-order conditions

Thus

$$\mathrm{d}\,\mathcal{L} = \tfrac{1}{2}n\,\mathrm{tr}\,\Omega^{-1}(S-\Omega)\Omega^{-1}(\mathrm{d}\,\Omega) - \sum_{j=1}^{n}\epsilon_j'\Omega^{-1}(\mathrm{d}\,\epsilon_j) =$$

$$= \tfrac{1}{2}n\,\mathrm{tr}\,\Omega^{-1}(S-\Omega)\Omega^{-1}(\mathrm{d}\,\Omega) + \sum_{j=1}^{n}\epsilon_j'\Omega^{-1}X_j(\mathrm{d}\beta), \tag{6.7}$$

where $S = (n^{-1})\sum_{j=1}^{n}\epsilon_j\epsilon_j'$ is the covariance matrix of the sample; note that $\mathrm{E}(S) = \Omega$. Vectorizing we get

$$\mathrm{d}\,\mathcal{L} = \tfrac{1}{2}n[\mathrm{vec}\,\Omega^{-1}(S-\Omega)\Omega^{-1}]'(\mathrm{vec}\,\mathrm{d}\,\Omega) + \sum_{j=1}^{n}\epsilon_j'\Omega^{-1}X_j(\mathrm{d}\beta) \tag{6.8}$$

$$= \tfrac{1}{2}[\mathrm{vec}\,\Omega^{-1}(S-\Omega)\Omega^{-1}]'D_m(\mathrm{d}\,\omega) + \sum_{j=1}^{n}\epsilon_j'\Omega^{-1}X_j(\mathrm{d}\beta), \tag{6.9}$$

where $\omega = \mathrm{v}(\Omega)$ contains only the distinct elements of Ω. Thus

$$\frac{\partial\mathcal{L}}{\partial\beta'} = \sum_{j=1}^{n}\epsilon_j'\Omega^{-1}X_j = \sum_{j=1}^{n}(y_j - X_j\beta)'\Omega^{-1}X_j, \tag{6.10a}$$

and

$$\frac{\partial\mathcal{L}}{\partial\omega'} = \tfrac{1}{2}n[\mathrm{vec}\,\Omega^{-1}(S-\Omega)\Omega^{-1}]'D_m. \tag{6.10b}$$

Setting the derivatives equal to zero and rearranging, we obtain as estimator of β:

$$\hat{\beta} = \left(\sum_{j=1}^{n}X_j'\hat{\Omega}^{-1}X_j\right)^{-1}\left(\sum_{j=1}^{n}X_j'\hat{\Omega}^{-1}y_j\right). \tag{6.11}$$

For Ω we get

$$\tfrac{1}{2}nD_m'(\Omega^{-1}\otimes\Omega^{-1})\,\mathrm{vec}(S-\Omega) = 0;$$

thus

$$\tfrac{1}{2}nD_m'(\Omega^{-1}\otimes\Omega^{-1})D_m\,\mathrm{v}(S-\Omega) = 0,$$

and, since $D_m'(\Omega^{-1}\otimes\Omega^{-1})D_m$ is non-singular (see 2.41), we get $\mathrm{v}(S-\Omega) = 0$, and so $\mathrm{vec}(S-\Omega) = 0$, which implies

$$\hat{\Omega} = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - X_j\hat{\beta}\right)\left(y_j - X_j\hat{\beta}\right)'. \tag{6.12}$$

6.4 The Hessian matrix

Taking the differential of (6.7) we get the second differential of the log-likelihood:

$$\begin{aligned}\mathrm{d}^2\,\mathcal{L} = &-\tfrac{1}{2}n\,\mathrm{tr}\,\Omega^{-1}(S-\Omega)\Omega^{-1}(\mathrm{d}\,\Omega)\Omega^{-1}(\mathrm{d}\,\Omega)\\ &-\tfrac{1}{2}n\,\mathrm{tr}\,\Omega^{-1}(\mathrm{d}\,\Omega)\Omega^{-1}(S-\Omega)\Omega^{-1}(\mathrm{d}\,\Omega)\end{aligned}$$

$$
\begin{aligned}
&- \tfrac{1}{2} n \sum_{j=1}^{n} \operatorname{tr} \Omega^{-1} \epsilon_j (\mathrm{d}\beta)' X_j' \Omega^{-1} (\mathrm{d}\,\Omega) \\
&- \tfrac{1}{2} n \sum_{j=1}^{n} \operatorname{tr} \Omega^{-1} X_j (\mathrm{d}\beta) \epsilon_j' \Omega^{-1} (\mathrm{d}\,\Omega) - \tfrac{1}{2} n \operatorname{tr} \Omega^{-1} (\mathrm{d}\,\Omega) \Omega^{-1} (\mathrm{d}\,\Omega) \\
&- (\mathrm{d}\beta)' \left(\sum_{j=1}^{n} X_j' \Omega^{-1} X_j \right) (\mathrm{d}\beta) - \sum_{j=1}^{n} \epsilon_j' \Omega^{-1} (\mathrm{d}\,\Omega) \Omega^{-1} X_j (\mathrm{d}\beta) \\
= &- \tfrac{1}{2} n (\operatorname{vec} \mathrm{d}\,\Omega)' [\Omega^{-1} \otimes \Omega^{-1} (2\Omega - S) \Omega^{-1}] (\operatorname{vec} \mathrm{d}\,\Omega) \\
&- (\mathrm{d}\beta)' \left(\sum_{j=1}^{n} X_j' \Omega^{-1} X_j \right) (\mathrm{d}\beta) \\
&- 2n (\operatorname{vec} \mathrm{d}\,\Omega)' (\Omega^{-1} \otimes \Omega^{-1}) \left(\sum_{j=1}^{n} \epsilon_j \otimes X_j \right) (\mathrm{d}\beta),
\end{aligned} \tag{6.13}
$$

where the last equality sign rests among others on (2.4). Thus the Hessian matrix of the log-likelihood is

$$
\mathsf{H}\,\mathcal{L}(\beta, \omega) = - \begin{pmatrix} \sum_{j=1}^{n} X_j' \Omega^{-1} X_j & n \left(\sum_{j=1}^{n} \epsilon_j' \otimes X_j' \right) (\Omega^{-1} \otimes \Omega^{-1}) D_m \\ n D_m' (\Omega^{-1} \otimes \Omega^{-1}) \left(\sum_{j=1}^{n} \epsilon_j \otimes X_j \right) & \tfrac{1}{2} n D_m' [\Omega^{-1} \otimes \Omega^{-1} (2\Omega - S) \Omega^{-1}] D_m \end{pmatrix}. \tag{6.14}
$$

6.5 The information matrix

Taking expectations and multiplying by -1 we get the information matrix

$$
\mathcal{I}(\beta, \omega) = \begin{pmatrix} \sum_{j=1}^{n} X_j' \Omega^{-1} X_j & 0 \\ 0 & \tfrac{1}{2} n D_m' [\Omega^{-1} \otimes \Omega^{-1}] D_m \end{pmatrix}, \tag{6.15}
$$

which can also be obtained by taking the expectation of the outer product of the first derivatives. It follows that the asymptotic covariance matrix of $\hat{\Omega}$ is given by

$$
\begin{aligned}
\mathrm{V}_{\mathrm{as}}[\sqrt{n} \operatorname{vec}(\hat{\Omega} - \Omega)] &= D_m \{ \mathrm{V}_{\mathrm{as}}[\sqrt{n}(\hat{\omega} - \omega)] \} D_m' \\
&= 2 D_m [D_m' (\Omega^{-1} \otimes \Omega^{-1}) D_m']^{-1} D_m' \\
&= 2 D_m D_m^{+} (\Omega \otimes \Omega) D_m^{+\prime} D_m' = 2 N_m (\Omega \otimes \Omega) N_m \\
&= 2 N_m (\Omega \otimes \Omega).
\end{aligned} \tag{6.16}
$$

Using (2.28) and (2.37) we get from (6.16)

$$
\operatorname{cov}_{\mathrm{as}}[\sqrt{n}(\hat{\omega}_{ij} - \omega_{ij}), \sqrt{n}(\hat{\omega}_{rs} - \omega_{rs})] = \omega_{ir}\omega_{js} + \omega_{is}\omega_{jr}.
$$

In particular

$$
\operatorname{var}_{\mathrm{as}}[\sqrt{n}(\hat{\omega}_{ij} - \omega_{ij})] = \omega_{ii}\omega_{jj} + \omega_{ij}^2,
$$

and

$$
\operatorname{var}_{\mathrm{as}}[\sqrt{n}(\hat{\omega}_{ii} - \omega_{ii})] = 2\omega_{ii}^2.
$$

Using the information matrix one easily shows that the method of scoring amounts to iterated generalized least squares according to (6.12) and (6.13). Since any information matrix is positive definite, this algorithm always leads to a maximum.

7 Maximum-likelihood estimation of the factor-analysis model

7.1 Introduction

In this section we give an application to the factor-analysis model and derive the Hessian matrix and the information matrix. Many books on multivariate analysis derive the first-order conditions, see *e.g.* Anderson (1958, chapter 14), Bartholomew (1987, chapter 3), Lawley and Maxwell (1963, chapter 4), and Morrison (1967, chapter 9). Lawley and Maxwell (1963, chapter 5) and Jöreskog (1972) also derive the information matrix, but they do not use only matrix methods.

Neudecker and Satorra (1991)...........................

Subsection 2 gives the model, subsection 3 derives the first-order conditions for maximum-likelihood estimation, subsection 4 the Hessian matrix, and subsection 5 the information matrix.

7.2 The model

Consider the model

$$x_{ji} = \mu_i + \sum_{k=1}^{q} \lambda_{ik} y_k + \epsilon_{ji}, \; i = 1, 2, ..., p, \; j = 1, 2, ..., n, \tag{7.1}$$

where μ_i and λ_{ik} are coefficients and y_k and ϵ_{ji} are random variables. In matrix notation we can write (7.1) as

$$x_j = \mu + \Lambda y + \epsilon_j, \; j = 1, 2, ..., n, \tag{7.2}$$

where $x_j = (x_{j1}, x_{j2}, ..., x_{jp})'$, $\mu = (\mu_1, \mu_2, ..., \mu_p)'$, $y = (y_1, y_2, ..., y_q)'$, $\epsilon_j = (\epsilon_{j1}, \epsilon_{j2}, ..., \epsilon_{jp})'$, and $\Lambda = (\lambda_{ik})$. It is assumed that

$$y \sim \mathcal{N}_q(0, I_q) \tag{7.3}$$

and

$$\epsilon_j \sim \mathcal{N}_p(0, \Psi), \tag{7.4}$$

The conditional distribution of x_j given y is therefore

$$x_j \mid y \sim N_p(\mu + \Lambda y, \Psi) \tag{7.5}$$

and the unconditional distribution of x_j is

$$x_j \sim N_p(\mu, \Lambda\Lambda' + \Psi). \tag{7.6}$$

For Λ and Ψ to be identifiable from a sample, we must impose restrictions on Λ and Ψ. Usually one assumes that Ψ is diagonal and imposes in addition some other restrictions on Λ and Ψ. For example, in confirmatory factor analysis one has information on Λ, such as that some λ_{ik} are zero; in exploratory factor analysis one usually assumes that

$\Lambda'\Psi^{-1}\Lambda$ is diagonal. We proceed to derive the maximum-likelihood estimates of μ, Λ, and Ψ under the assumption that Ψ is diagonal.

Suppose we have n observations on the vector x. The log-likelihood of the sample is then

$$\mathcal{L} = -\tfrac{1}{2}np\log 2\pi - \tfrac{1}{2}n\log|\Omega| - \tfrac{1}{2}n\operatorname{tr}\Omega^{-1}S, \tag{7.7}$$

where $\Omega = \Lambda\Lambda' + \Psi$, and $S = n^{-1}\sum_{j=1}^{n}(x_j-\mu)(x_j-\mu)'$ is the covariance matrix of the sample.

7.3 First-order conditions

The differential of the log-likelihood is

$$\begin{aligned}
\mathrm{d}\,\mathcal{L} &= -\tfrac{1}{2}n\,\mathrm{d}(\log|\Omega|) - \tfrac{1}{2}n\operatorname{tr}\Omega^{-1}\,\mathrm{d}\,S - \tfrac{1}{2}n\operatorname{tr}(\mathrm{d}\,\Omega^{-1})S \\
&= -\tfrac{1}{2}n\operatorname{tr}\Omega^{-1}(\mathrm{d}\,\Omega)\Omega^{-1}(\Omega - S) + n\operatorname{tr}\Omega^{-1}u(\mathrm{d}\,\mu)',
\end{aligned} \tag{7.8}$$

where $u = n^{-1}\sum_j(x_j-\mu)$; note that $\mathrm{E}\,u = 0$. Using $\mathrm{d}\,\Omega = \Lambda(\mathrm{d}\,\Lambda)' + (\mathrm{d}\,\Lambda)\Lambda' + \mathrm{d}\,\Psi$, we get

$$\begin{aligned}
\mathrm{d}\,\mathcal{L} &= -\,n\operatorname{tr}\Omega^{-1}(\mathrm{d}\,\Lambda)\Lambda'\Omega^{-1}(\Omega - S) - \tfrac{1}{2}n\operatorname{tr}\Omega^{-1}(\mathrm{d}\,\Psi)\Omega^{-1}(\Omega - S) \\
&\quad + n\operatorname{tr}\Omega^{-1}u(\mathrm{d}\,\mu)' \\
&= -\,n[\operatorname{vec}\Omega^{-1}(\Omega - S)\Omega^{-1}\Lambda]'(\mathrm{d}\,\lambda) - \tfrac{1}{2}n[\operatorname{vec}\Omega^{-1}(\Omega - S)\Omega^{-1}]G_p'\psi \\
&\quad + n(\operatorname{vec}\Omega^{-1}u)'(\mathrm{d}\,\mu),
\end{aligned} \tag{7.9}$$

where $\lambda = \operatorname{vec}\Lambda$, and $\psi = \mathrm{w}(\Psi) = G_p'\operatorname{vec}\Psi$ is the vector with the diagonal elements of Ψ. Thus, the first derivatives of L are

$$\frac{\partial\mathcal{L}}{\partial\mu'} = n\operatorname{vec}u'\Omega^{-1} = nu'\Omega^{-1}, \tag{7.10a}$$

$$\frac{\partial\mathcal{L}}{\partial\lambda'} = -n[\operatorname{vec}\Omega^{-1}(\Omega - S)\Omega^{-1}\Lambda]', \tag{7.10b}$$

$$\frac{\partial\mathcal{L}}{\partial\psi'} = -\tfrac{1}{2}n[\operatorname{vec}\Omega^{-1}(\Omega - S)\Omega^{-1}]'G_p'. \tag{7.10c}$$

Setting the first derivatives equal to zero we get from (7.10a) $u = 0$ and so

$$\hat{\mu} = \frac{1}{n}\sum_{j=1}^{n}x_j = \bar{x}, \tag{7.11}$$

$$\hat{S} = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})(x_j - \bar{x})'; \tag{7.12}$$

from (7.10b) we get

$$(\hat{\Omega} - \hat{S})\hat{\Omega}^{-1}\hat{\Lambda} = 0; \tag{7.13}$$

from (7.10c) we get

$$\operatorname{dg}[(\hat{\Omega} - \hat{S})\hat{\Omega}^{-1}] = 0. \tag{7.14}$$

7.4 The Hessian matrix

From (7.8) we obtain as the second differential of the log-likelihood equation:

$$\begin{aligned}
\mathrm{d}^2\mathcal{L} &= -\tfrac{1}{2}n\,\mathrm{tr}\,\Omega^{-1}(\mathrm{d}\,\Omega)\Omega^{-1}(\mathrm{d}\,\Omega-\mathrm{d}\,S)+n\,\mathrm{tr}\,\Omega^{-1}(\mathrm{d}\,\Omega)\Omega^{-1}(\mathrm{d}\,\Omega)\Omega^{-1}(\Omega-S)\\
&\quad+n\,\mathrm{tr}\,\Omega^{-1}(\mathrm{d}\,u)(\mathrm{d}\,\mu)'-n\,\mathrm{tr}\,\Omega^{-1}(\mathrm{d}\,\Omega)\Omega^{-1}u(\mathrm{d}\,\mu)'\\
&= -\tfrac{1}{2}n\,\mathrm{tr}\,\Omega^{-1}\Phi\Omega^{-1}(\mathrm{d}\,\Omega)\Omega^{-1}(\mathrm{d}\,\Omega)\\
&\quad-n(\mathrm{d}\,\mu)'\Omega^{-1}(\mathrm{d}\,\mu)-2n\,\mathrm{tr}\,\Omega^{-1}(\mathrm{d}\,\Omega)\Omega^{-1}u(\mathrm{d}\,\mu)',
\end{aligned} \tag{7.15}$$

where $\Phi = 2S - \Omega$. Using $\mathrm{d}\,\Omega = \Lambda(\mathrm{d}\,\Lambda)' + (\mathrm{d}\,\Lambda)\Lambda' + \mathrm{d}\,\Psi$, we get

$$\begin{aligned}
\mathrm{d}^2\mathcal{L} &= -\tfrac{1}{2}n\,\mathrm{tr}\,\Omega^{-1}\Phi\Omega^{-1}\Lambda(\mathrm{d}\,\Lambda)'\Omega^{-1}\Lambda(\mathrm{d}\,\Lambda)'\\
&\quad-\tfrac{1}{2}n\,\mathrm{tr}\,\Omega^{-1}\Phi\Omega^{-1}\Lambda(\mathrm{d}\,\Lambda)'\Omega^{-1}(\mathrm{d}\,\Lambda)\Lambda'\\
&\quad-\tfrac{1}{2}n\,\mathrm{tr}\,\Omega^{-1}\Phi\Omega^{-1}(\mathrm{d}\,\Lambda)\Lambda'\Omega^{-1}(\mathrm{d}\,\Lambda)\Lambda'-\tfrac{1}{2}n\,\mathrm{tr}\,\Omega^{-1}\Phi\Omega^{-1}(\mathrm{d}\,\Lambda)\Lambda'\Omega^{-1}\Lambda(\mathrm{d}\,\Lambda)'\\
&\quad-n\,\mathrm{tr}\,\Omega^{-1}\Phi\Omega^{-1}\Lambda(\mathrm{d}\,\Lambda)'\Omega^{-1}(\mathrm{d}\,\Psi)-n\,\mathrm{tr}\,\Omega^{-1}\Phi\Omega^{-1}(\mathrm{d}\,\Lambda)\Lambda'\Omega^{-1}(\mathrm{d}\,\Psi)\\
&\quad-\tfrac{1}{2}n\,\mathrm{tr}\,\Omega^{-1}\Phi\Omega^{-1}(\mathrm{d}\,\Psi)\Omega^{-1}(\mathrm{d}\,\Psi)-n(\mathrm{d}\,\mu)'\Omega^{-1}(\mathrm{d}\,\mu)\\
&\quad-2n\,\mathrm{tr}\,\Omega^{-1}\Lambda(\mathrm{d}\,\Lambda)'\Omega^{-1}u(\mathrm{d}\,\mu)'-2n\,\mathrm{tr}\,\Omega^{-1}(\mathrm{d}\,\Lambda)\Lambda'\Omega^{-1}u(\mathrm{d}\,\mu)'\\
&\quad-2n\,\mathrm{tr}\,\Omega^{-1}(\mathrm{d}\,\Psi)\Omega^{-1}u(\mathrm{d}\,\mu)'\\
&= -\tfrac{1}{2}n(\mathrm{vec}\,\mathrm{d}\,\Lambda')'(\Omega^{-1}\Phi\Omega^{-1}\Lambda\otimes\Lambda'\Omega^{-1})(\mathrm{vec}\,\mathrm{d}\,\Lambda)\\
&\quad-\tfrac{1}{2}n(\mathrm{vec}\,\mathrm{d}\,\Lambda)'(\Lambda'\Omega^{-1}\Phi\Omega^{-1}\Lambda\otimes\Omega^{-1})(\mathrm{vec}\,\mathrm{d}\,\Lambda)\\
&\quad-\tfrac{1}{2}n(\mathrm{vec}\,\mathrm{d}\,\Lambda')'(\Omega^{-1}\Lambda\otimes\Lambda'\Omega^{-1}\Phi\Omega^{-1})(\mathrm{vec}\,\mathrm{d}\,\Lambda)\\
&\quad-\tfrac{1}{2}n(\mathrm{vec}\,\mathrm{d}\,\Lambda)'(\Lambda'\Omega^{-1}\Lambda\otimes\Omega^{-1}\Phi\Omega^{-1})(\mathrm{vec}\,\mathrm{d}\,\Lambda)\\
&\quad-n(\mathrm{vec}\,\mathrm{d}\,\Psi)'(\Omega^{-1}\Phi\Omega^{-1}\Lambda\otimes\Omega^{-1})(\mathrm{vec}\,\mathrm{d}\,\Lambda)\\
&\quad-n(\mathrm{vec}\,\mathrm{d}\,\Psi)'(\Omega^{-1}\Lambda\otimes\Omega^{-1}\Phi\Omega^{-1})(\mathrm{vec}\,\mathrm{d}\,\Lambda)\\
&\quad-\tfrac{1}{2}n(\mathrm{vec}\,\mathrm{d}\,\Psi)'(\Omega^{-1}\otimes\Omega^{-1}\Phi\Omega^{-1})(\mathrm{vec}\,\mathrm{d}\,\Psi)\\
&\quad-n(\mathrm{d}\,\mu)'\Omega^{-1}(\mathrm{d}\,\mu)-2n(\mathrm{vec}\,\mathrm{d}\,\mu)(u'\Omega^{-1}\otimes\Omega^{-1}\Lambda)(\mathrm{vec}\,\mathrm{d}\,\Lambda')\\
&\quad-2n(\mathrm{vec}\,\mathrm{d}\,\mu)'(u'\Omega^{-1}\Lambda\otimes\Omega^{-1})(\mathrm{vec}\,\mathrm{d}\,\Lambda)\\
&\quad-2n(\mathrm{vec}\,\mathrm{d}\,\mu)'(u'\Omega^{-1}\otimes\Omega^{-1})(\mathrm{vec}\,\mathrm{d}\,\Psi).
\end{aligned} \tag{7.16}$$

After some algebra it appears that the Hessian matrix of the log-likelihood has the form

$$\mathsf{H}\,\mathcal{L}(\mu,\lambda,\psi)=\begin{pmatrix}\mathsf{H}_{\mu\mu} & \mathsf{H}_{\mu\lambda} & \mathsf{H}_{\mu\psi}\\ \mathsf{H}_{\lambda\mu} & \mathsf{H}_{\lambda\lambda} & \mathsf{H}_{\lambda\psi}\\ \mathsf{H}_{\psi\mu} & \mathsf{H}_{\psi\lambda} & \mathsf{H}_{\psi\psi}\end{pmatrix}, \tag{7.17}$$

with

$$\mathsf{H}_{\mu\mu} = -\,n\Omega^{-1}, \tag{7.18}$$

$$\begin{aligned}
\mathsf{H}_{\mu\lambda} = \mathsf{H}'_{\lambda\mu} &= -n(u'\Omega^{-1}\otimes\Omega^{-1})K_{pq}-n(u'\Omega^{-1}\Lambda\otimes\Omega^{-1})\\
&= -\,n(u'\Omega^{-1}\otimes\Omega^{-1})[(I_p\otimes\Lambda)K_{pq}+(\Lambda\otimes I_p)]\\
&= -\,n(u'\Omega^{-1}\otimes\Omega^{-1})(K_{pp}+I_{p^2})(\Lambda\otimes I_p)\\
&= -\,2n(u'\Omega^{-1}\otimes\Omega^{-1})N_p(\Lambda\otimes I_p)
\end{aligned} \tag{7.19}$$

$$
\begin{aligned}
\mathsf{H}_{\mu\psi} &= \mathsf{H}'_{\psi\mu} = -n(u'\Omega^{-1} \otimes \Omega^{-1})G'_p, && (7.20)\\
\mathsf{H}_{\lambda\lambda} &= -\tfrac{1}{2}nK_{qp}(\Omega^{-1}\Phi\Omega^{-1}\Lambda \otimes \Lambda'\Omega^{-1}) - \tfrac{1}{2}n(\Lambda'\Omega^{-1}\Phi\Omega^{-1}\Lambda \otimes \Omega^{-1}) \\
&\quad - \tfrac{1}{2}n(\Lambda'\Omega^{-1}\Lambda \otimes \Omega^{-1}\Phi\Omega^{-1}) - \tfrac{1}{2}nK_{qp}(\Omega^{-1}\Lambda \otimes \Lambda'\Omega^{-1}\Phi\Omega^{-1}) \\
&= -n(\Lambda'\Omega^{-1} \otimes \Omega^{-1})N_p[(\Omega \otimes \Phi) + (\Phi \otimes \Omega)](\Omega^{-1}\Lambda \otimes \Omega^{-1}) \\
&= -2n(\Lambda' \otimes I_p)N_p(\Omega^{-1} \otimes \Omega^{-1}\Phi\Omega^{-1})N_p(\Lambda \otimes \Omega^{-1}), && (7.21)\\
\mathsf{H}_{\lambda\psi} &= \mathsf{H}'_{\psi\lambda} = -\tfrac{1}{2}n(\Lambda'\Omega^{-1} \otimes \Omega^{-1}\Phi\Omega^{-1})G'_p \\
&\quad - \tfrac{1}{2}n(\Lambda'\Omega^{-1}\Phi\Omega^{-1} \otimes \Omega^{-1})G'_p \\
&= -n(\Lambda' \otimes I_p)N_p(\Omega^{-1} \otimes \Omega^{-1}\Phi\Omega^{-1})G'_p, && (7.22)\\
\mathsf{H}_{\psi\psi} &= -\tfrac{1}{2}nG_p(\Omega^{-1} \otimes \Omega^{-1}\Phi\Omega^{-1})G'_p = -\tfrac{1}{2}n(\Omega^{-1} \odot \Omega^{-1}\Phi\Omega^{-1}). && (7.23)
\end{aligned}
$$

Note that if $u = 0$ (which holds at the maximum-likelihood estimate, see (7.10a)), then $\mathsf{H}_{\mu\lambda} = 0$ and $\mathsf{H}_{\mu\psi} = 0$. It follows from (7.17)-(7.23) that the Hessian matrix is

$$
\mathsf{H}\,\mathcal{L}(\mu, \lambda, \psi) = -n\times
$$

$$
\begin{pmatrix}
\Omega^{-1} & 2(u'\Omega^{-1}\otimes\Omega^{-1})N_p(\Lambda\otimes I_p) & (u'\Omega^{-1}\otimes\Omega^{-1})G'_p \\
2(\Lambda'\otimes I_p)N_p(\Omega^{-1}u\otimes\Omega^{-1}) & 2(\Lambda'\otimes I_p)N_p(\Omega^{-1}\otimes\Omega^{-1}\Phi\Omega^{-1})N_p(\Lambda\otimes I_p) & (\Lambda'\otimes I_p)N_p(\Omega^{-1}\otimes\Omega^{-1}\Phi\Omega^{-1})G'_p \\
G_p(\Omega^{-1}u\otimes\Omega^{-1}) & G_p(\Omega^{-1}\otimes\Omega^{-1}\Phi\Omega^{-1})N_p(\Lambda\otimes I_p) & \tfrac{1}{2}(\Omega^{-1}\odot\Omega^{-1}\Phi\Omega^{-1})
\end{pmatrix} \quad (7.24)
$$

Expressions for the individual elements of H, such as $\partial^2\mathcal{L}/(\partial\lambda_{ir}\partial\lambda_{js})$, can be obtained from (7.18)-(7.23) with the help of equations (2.29), (2.38), and (2.59)

7.5 The information matrix

Taking expectations of the Hessian matrix (7.24) and multiplying by -1, we get as information matrix

$$
\mathcal{I}(\mu, \lambda, \psi) = n\times
$$

$$
\begin{pmatrix}
\Omega^{-1} & 0 & 0 \\
0 & 2(\Lambda' \otimes I_p)N_p(\Omega^{-1} \otimes \Omega^{-1})N_p(\Lambda \otimes I_p) & (\Lambda' \otimes I_p)N_p(\Omega^{-1} \otimes \Omega^{-1})G'_p \\
0 & G_p(\Omega^{-1} \otimes \Omega^{-1})N_p(\Lambda \otimes I_p) & \tfrac{1}{2}(\Omega^{-1} \odot \Omega^{-1})
\end{pmatrix} \quad (7.25)
$$

From (7.25) we get

$$
\mathcal{I}_{\lambda\lambda} = n(\Lambda'\Omega^{-1}\Lambda) \otimes \Omega^{-1} + nK_{qp}(\Omega^{-1}\Lambda \otimes \Lambda'\Omega^{-1}); \quad (7.26)
$$

and therefore, using (2.27) and (2.36),

$$
\mathcal{I}(\lambda_{ij}, \lambda_{rs}) = n(\omega^{ir}\gamma_{js} + \theta_{is}\theta_{rj}), \quad (7.27)
$$

where $\Gamma = \Lambda'\Omega^{-1}\Lambda$ and $\Theta = \Omega^{-1}\Lambda$. Similarly, we get

$$
\mathcal{I}(\psi_k, \lambda_{ij}) = n\omega^{ik}\theta_{kj}, \quad (7.28)
$$

and

$$
\mathcal{I}(\psi_k, \psi_\ell) = \tfrac{1}{2}n(\omega^{k\ell})^2. \quad (7.29)
$$

Equations (7.27)-(7.29) coincide with the formulae derived without matrix differentiation techniques by Jöreskog (1972). If there are no other restrictions on the parameters and if the model is identified, then we can derive the asymptotic variances of (μ, λ, ψ) by inverting the information matrix.

References

Anderson, T.W., 1958, An introduction to multivariate statistical analysis, second edition 1984. Wiley, New York.

Bartholomew, D.J., 1987, Latent variable models and factor analysis. Griffin, London.

Binder, D.A. 1996, Linearization methods for single phase and two-phase samples: a cookbook approach. Survey Methodology 22, 17-22.

Dieudonné, J., 1960, Foundations of modern analysis, 2nd edition (1969). Academic Press, New York.

Jöreskog, K.G., 1971, Simultaneous factor analysis in several populations. Psychometrika 36(4), pp. 409-426.

Lawley, D.N. and A.E. Maxwell, 1963, Factor analysis as a statistical method, second edition, 1971. Buttersworth, London.

Magnus, J.R., 1988, Linear structures. Griffin, London.

Magnus, J.R. and H. Neudecker, 1988, Matrix differential calculus. Wiley, New York.

Morrison, D.F., 1967, Multivariate statistical methods, second edition 1976. McGraw-Hill, New York.

Neudecker, H. and A. Satorra, 1991, Linear structural relations: gradient and Hessian of the fitting function. Statistics and Probability Letters 11, 57-61.

Särndal, C.-E., Swensson, B., and Wretman, J., 1992, Model Assisted Survey Sampling. New York: Springer.

Zeelenberg, K., 1997, A simple derivation of the linearization of the regression estimator, Survey Methodology 23, 77-8.