Bilevel factor analysis models

by

Jacobus Johannes Pietersen

Submitted in partial fulfilment of the requirements

for the degree of

Doctor of Philosophy in the subject Applied Statistics

in the Faculty of Science

University of Pretoria

Pretoria

January 2000

CONTENTS

Page

SUMMARY

TITLE: Bilevel factor analysis models
CANDIDATE: J.J. PIETERSEN
PROMOTER: PROFESSOR S.H.C. DU TOIT
DEPARTMENT: STATISTICS
DEGREE: PH.D. (APPLIED STATISTICS)

The theory of ordinary factor analysis and its application by means of software packages do not make provision for data sampled from populations with hierarchical structures. Since data are often obtained from such populations - educational data for example - the lack of procedures to analyse data of this kind needs to be addressed.

A review of the ordinary factor analysis model and maximum likelihood estimation of the parameters in exploratory and confirmatory models is provided, together with practical applications. Subsequently, the concept of hierarchically structured populations and their models, called multilevel models, are introduced. A general framework for the estimation of the unknown parameters in these models is presented. It contains two estimation procedures. The first is the marginal maximum likelihood method in which an iterative expected maximisation approach is used to obtain the maximum likelihood estimates. The second is the Fisher scoring method which also provides estimated standard errors for the maximum likelihood parameter estimates. For both methods, the theory is presented for unconstrained as well as for constrained estimation. A two-stage procedure - combining the mentioned procedures - is proposed for parameter estimation in practice.

Multilevel factor analysis models are introduced next, and subsequently a particular two-level factor analysis model is presented. The general estimation theory that was presented earlier is applied to this model - in exploratory and confirmatory analysis. First, the marginal maximum likelihood method is used to obtain the equations for determining the parameter estimates. It is then shown how an iterative expected maximisation algorithm is used to obtain these estimates in unconstrained and constrained

optimisation. This method is applied to real life data using a FORTRAN program. Secondly, equations are derived by means of the Fisher scoring method to obtain the maximum likelihood estimates of the parameters in the two-level factor analysis model for exploratory and confirmatory analysis. A FORTRAN program was written to apply this method in practice. Real life data are used to illustrate the method.

Finally, flowing from this research, some areas for possible further research are proposed.

OPSOMMING

ONDERWERP:	Tweepeil faktorontledingsmodelle
KANDIDAAT:	J.J. PIETERSEN
PROMOTOR:	PROFESSOR S.H.C. DU TOIT
DEPARTEMENT:	STATISTIEK
GRAAD:	PH.D. (TOEGEPASTE STATISTIEK)

Die teorie van gewone faktorontleding en die toepassing deur middel van statistiese sagtewarepakette maak nie voorsiening vir die ontleding van data afkomstig uit populasies met hiërargiese strukture nie. Aangesien data in baie gevalle uit sulke populasies afkomstig is - onderwys is 'n goeie voorbeeld - behoort die afwesigheid van beramingsprosedures in die ontleding van modelle vir sulke data aangespreek te word.

'n Oorsig oor die gewone faktorontledingsmodel en die beskouing van maksimum aanneemlikheidsberaming van die parameters in ondersoekende en bevestigende modelle word gegee. Die teorie word dan toegepas op werklike data. Daarna word die konsep van hiërargies gestruktureerde populasies en modelle vir data uit sulke populasies, naamlik meerpeilmodelle, beskryf. 'n Algemene raamwerk vir die beraming van die onbekende parameters in hierdie modelle word gegee. Dit bevat twee beramingsprosedures. Die eerste is die marginale maksimum aanneemlikheidsmetode waarin 'n iteratiewe verwagte maksimeringsbenadering gebruik word om die maksimum aanneemlikheidsberamers te verkry. Die tweede is die bekende "Fisher scoring" metode wat ook beraamde standaardfoute vir die maksimum aanneemlikheidsberamers van die parameters gee. Vir beide metodes word die teorie bespreek ten opsigte van gevalle waar geen beperkings op die parameters geplaas word nie en ook waar daar wel beperkings op hulle geplaas word. 'n Twee-stadium prosedure - 'n kombinasie van die genoemde prosedures - word voorgestel vir die beraming van parameters in die praktyk.

Meerpeilfaktorontledingsmodelle word volgende bespreek, en vervolgens word 'n spesifieke tweepeilfaktorontledingsmodel beskryf. Die algemene beramingsteorie wat vroeër bespreek is, word toegepas op hierdie model - in ondersoekende en bevestigende ontled-

ing. Eerstens word die marginale maksimum aanneemlikheidsmetode gebruik om die vergelykings te bepaal vir die berekening van die parameterberamings. Dan word aangetoon hoe 'n iteratiewe verwagte maksimeringsalgoritme gebruik word om die beramings, met en sonder beperkings, te bereken. Hierdie metode word toegepas op werklike data deur gebruik te maak van 'n FORTRAN-program wat spesiaal vir die doel geskryf is. Tweedens word vergelykings afgelei deur gebruik te maak van die "Fisher scoring" metode om maksimum aanneemlikheidsberamers te bereken vir die parameters in die tweepeilfaktorontledingsmodel in ondersoekende en bevestigende ontleding. Weereens word die gevalle beskou waar beperkings en geen beperkings op die parameters geplaas word. 'n FORTRAN-program is geskryf om hierdie metode in die praktyk toe te pas. Werklike data word gebruik om die metode te illustreer.

Ten slotte, voortspruitend uit hierdie navorsing, word 'n aantal onderwerpe vir moontlike verdere navorsing voorgestel.

Notation

The following notation will be adopted in this thesis:

π	:	constant, $\pi = 3.14159...$
e	:	Euler's constant, $e = 2.71828...$
$\exp(x)$	:	$e^x, -\infty < x < \infty$
$\ell n\ x$	:	natural logarithm of the real number x, $x \geq 0$
δ_{ij}	:	Kronecker's delta (1 if $i = j$ and 0 if $i \neq j$)
$\mathbf{A} : (p \times q)$	:	matrix of order $p \times q$
$\mathbf{a} : (p \times 1)$	:	column vector of order $p \times 1$
a	:	scalar
$\mathbf{A}'$	:	transpose of $\mathbf{A}$
$\mathbf{a}'$	:	transpose of $\mathbf{a}$ (a row vector)
a_{ij} or $[A]_{ij}$	:	the element in the i-th row and j-th column of $\mathbf{A}$
a_i or $[\mathbf{a}]_{i1}$	:	the i-th element of $\mathbf{a}$
$\mathbf{A}^{-1}$	:	inverse of $\mathbf{A}$
a^{ij}	:	$[\mathbf{A}^{-1}]_{ij}$
$\|\mathbf{A}\|$	:	determinant of $\mathbf{A}$
tr[$\mathbf{A}$]	:	trace of $\mathbf{A}$
$\mathbf{D}_a$	:	diagonal matrix with diagonal elements $a_{11}, a_{22}, \cdots$
Diag[$\mathbf{A}$]	:	diagonal matrix formed from the diagonal elements of $\mathbf{A}$
diag[$\mathbf{A}$]	:	columnvector formed from the diagonal elements of $\mathbf{A}$
vec[$\mathbf{A}$]	:	$(pq \times 1)$ vector formed from the q columns of the $p \times q$ matrix $\mathbf{A}$
vecs[$\mathbf{A}$]	:	$(p(p+1)/2 \times 1)$ vector formed from the nonduplicated elements of the $(p \times p)$ symmetric matrix $\mathbf{A}$
vecs*[$\mathbf{A}$]	:	$(p(p-1)/2 \times 1)$ vector formed from the nonduplicated off-diagonal elements of the $(p \times p)$ symmetric matrix $\mathbf{A}$
$\mathbf{0}$	:	null matrix, $[\mathbf{0}]_{ij} = 0$
$\mathbf{0}$ or $\mathbf{0}_p$	:	$(p \times 1)$ null vector, $[\mathbf{0}]_{i1} = 0$

$\mathbf{j}$ or $\mathbf{j}_p$: $(p \times 1)$ vector with unit elements, $[\mathbf{j}]_{i1} = 1$

$\mathbf{I}$ or $\mathbf{I}_p$: $(p \times p)$ identity matrix

$\mathbf{J}_{ij}$: matrix with all elements equal to zero with the exception of the element in the i-th row and j-th column which is equal to unity

$\mathbf{J}_{i1}$: column vector with all elements equal to zero with the exception of the i-th element which is equal to unity

$\mathbf{A} \otimes \mathbf{B}$: The right direct product or "Kronecker product " of matrices $\mathbf{A}$ and $\mathbf{B}$ defined by:

$$\begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1q}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2q}\mathbf{B} \\ \vdots & \vdots & \vdots & \vdots \\ a_{p1}\mathbf{B} & a_{p2}\mathbf{B} & \cdots & a_{pq}\mathbf{B} \end{pmatrix}$$

$\frac{\partial \mathbf{A}}{\partial x}$: matrix with typical element $\frac{\partial a_{ij}}{\partial x}$

$\frac{\partial \mathbf{a}}{\partial x}$: column vector with typical element $\frac{\partial [\mathbf{a}]_{i1}}{\partial x}$

$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$: column vector with typical element $\frac{\partial f(\mathbf{x})}{\partial [\mathbf{x}]_{i1}}$

$\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'}$: symmetric matrix with typical element $\frac{\partial^2 f(\mathbf{x})}{\partial [\mathbf{x}]_{i1} \partial [\mathbf{x}]_{j1}}$

$E(\mathbf{y}) : (p \times 1)$: expected value of the random vector $\mathbf{y}$ with typical element $E(y_i)$

$\mathrm{Cov}(\mathbf{y}, \mathbf{y}') : (p \times p)$: covariance matrix of the random vector $\mathbf{y}$ with typical element $E[y_i - E(y_i)]\ [y_j - E(y_j)]$

CHAPTER 1

INTRODUCTION

In many fields of research two important aspects of the process are the collection of information and its analysis. Data analysis has received much attention over the years; a variety of statistical techniques have been developed and are used daily to analyse many kinds of data. The task has been made relatively easy by modern technology. Many statistical software packages are now available for analysing data, even on personal computers; and the packages are regularly updated to take care of new developments in analytic techniques.

Most computer packages are written to analyse data obtained from individuals of a particular population, or to compare data obtained from individuals of more than one population. The techniques are not designed to deal with information about populations in which there are built-in hierarchies - i.e. where the individuals are grouped into clusters or groups, for a two-level hierarchy, and also where these clusters may be grouped into even larger homogeneous groups, for a three-level hierarchy.

Incorrect ways of treating hierarchical data that were used in the past, and are probably still used by some researchers, are to aggregate or disaggregate the observed variables. By "aggregation" is meant that variables at the individual level are aggregated to the higher level and that the analysis is then done at that higher level. By "disaggregation" is meant that the higher order variables are disaggregated to the individual level where the analysis is then done. The aggregation of variables leads to the loss of all the within-group information, the disaggregation of variables leads to individuals within a group having the same value on the group variables, and this violates the assumption of independence of the observations. These two methods of analysing hierarchical data are therefore unsatisfactory (Bryk and Raudenbush, 1992).

Because of the necessity to take the structure of the data into account in the analysis

process, a general framework for nested data was introduced by Lindley and Smith (1972). At that time only very simple problems could be subjected to data analysis; a general estimation procedure was not then available, and the analysis of hierarchical data requires fairly sophisticated estimation procedures. It was the development of the EM algorithm by Dempster, Laird and Rubin in 1977 (Bryk and Raudenbush, 1992) that provided a method of estimation appropriate to the analysis of this kind of data.

Computer software for analysing hierarchical data was developed only in the late 1980s and is not yet widely available. Four such computer programs for fitting models are GENMOD, HLM, ML2 and VARCL (Bryk and Raudenbush, 1992). Since they do not make provision for analysing latent variable models applied to hierarchical data, the aim of this thesis is to present estimation procedures in the analysis of a specific latent variable model; namely a factor analysis model for two-level hierarchical data. In addition to setting out the theory of the estimation procedures, a computer program written in FORTRAN is used to apply the derived theory to some real life data in order to show the feasibility of the procedures. A summary of the contents of each chapter in this thesis will now be provided.

In Chapter 2 a review of factor analysis as a statistical technique is presented. The classical application of factor analysis to multivariate data obtained from a single population in which no hierarchy is present, is considered. The mathematical model used in factor analysis is defined, and a procedure to estimate the parameters in the model is then described. Distinction is made between an exploratory and a confirmatory model. The whole issue of factor rotation is discussed; and some practical applications of factor analysis to real life data are offered in the final section of the chapter.

The concept of multilevel models for univariate as well as for multivariate data obtained from hierarchical populations is introduced in Chapter 3. General linear models are defined for multilevel data using Goldstein and McDonald (1988) as main reference. The general models, and two methods for obtaining estimators of the unknown parameters in the case of a general multivariate two-level model, are described. The first of these is

the marginal maximum likelihood method; and the EM algorithm is used here to obtain the parameter estimates iteratively. The second is the Fisher scoring method to obtain maximum likelihood estimates of the parameters, also in an iterative way. Both these methods are then adapted to provide the parameter estimators if constraints are to be imposed on a subset of the parameters.

Having considered general models in the previous chapter, latent variable multilevel models are introduced in Chapter 4. A specific latent variable model - namely, a two-level factor analysis model - is described, and it is shown how this model fits into the general framework described in Chapter 3. The method of marginal maximum likelihood in estimating the parameters in this factor analysis model is then discussed in detail. Expressions are obtained in this procedure that are shown to be ideal for use in an EM algorithm to obtain the estimates. Typically, non-linear constraints are imposed on some of the parameters in an exploratory factor model. Since these constraints are also imposed on the parameters in the two-level model, it is indicated how these non-linear constraints may be approximated by a set of linear constraints in order to simplify some derivations. The linear constraints are then imposed on the parameters, and it is shown how the EM algorithm may be adjusted to obtain the parameter estimates subject to these constraints. An EM algorithm is proposed for exploratory as well as for confirmatory analysis. The final section of the chapter is devoted to practical applications where this method is used to obtain parameter estimates in the modelling of real life data.

In Chapter 5 a further method of estimating the parameters in the two-level factor analysis model is described. This is the Fisher scoring method which has the advantage of providing, in addition to the parameter estimates, an estimate of the covariance matrix of these estimators. First, the likelihood function for maximisation under normality assumptions is obtained and is changed to a function for minimisation, called the discrepancy function. Subsequently the gradient vector and expected Hessian matrix of this discrepancy function are obtained. These quantities are then used in the Fisher scoring method to obtain the parameter estimates and their estimated covariance ma-

trix, in both exploratory and confirmatory models. It is also indicated how this method may be used when non-linear constraints are imposed on some of the parameters, as is the case in exploratory analysis. Some remarks are made on testing the goodness of fit of this model and on hypotheses that may also be tested. Finally, real life data are used to show the application of this method in practice.

In the final chapter, namely Chapter 6, a few topics are mentioned which have not been investigated in this thesis and which may lead to interesting further research.

CHAPTER 2

A REVIEW OF FACTOR ANALYSIS

2.1 Introduction

In this chapter a review of factor analysis will be given as it has been, and still is, employed by many researchers and other users of statistics. The technique, as described here, is well known and very popular in some fields of research - specifically in psychology - and is currently a very useful tool for analysing sample correlation and covariance matrices.

As a statistical method, factor analysis dates back to the second half of the nineteenth century (Mulaik, 1972). It has since been developed as a very powerful technique in data analysis. Many books have been written on the topic, of which the one by Mulaik (1972), which also provides a brief history of the development of factor analysis as a linear model, qualifies as an important reference.

There are two distinct stages in factor analysis, namely exploratory and confirmatory factor analysis. Of these two, exploratory factor analysis was first developed and use was mainly made of 'approximate' and easy-to-calculate methods in the era before modern computers became available. However, in the mid to late 1960s Bock and Bargmann (1966) and Jöreskog (1969) introduced a more confirmatory approach to factor analysis in the sense that various parameters in the model could be specified *a priori*, followed by a goodness of fit test of the model. This means that more meaningful constraints are imposed on the parameters whereas in exploratory factor analysis arbitrary constraints, mainly incorporated for computational convenience only, are used. It is these advantages of confirmatory factor analysis, namely substantively motivated constraints and the statistical test of the model, that are responsible for the current tendency of researchers to gradually move away from exploratory, and more towards confirmatory

factor analysis.

In the next section the mathematical model used in exploratory factor analysis, and the parameters involved in this model, will be introduced. Section 2.3 will then deal with a problem inherent to the factor analysis model, namely identification. In Section 2.4 the estimation of the parameters in the exploratory factor analysis model will be dealt with while the issue of factor rotation will be discussed in Section 2.5. The last two sections will be used to discuss parameter estimation in the confirmatory factor analysis model and to present some practical applications.

2.2 The factor analysis model

A number of distinct rationales can be given to express the basic assumptions of the common factor model in equivalent ways. Two of these - used by McDonald (1985) in his definition of the model - are, firstly, that there are a number of unobserved variables that explain the observed covariances (or correlations), i.e. if these unobserved variables are partialled out, the covariances (or correlations) between the observed variables are zero, and secondly, that each observed variable can be expressed as its regression on a number of unobserved variables plus a residual about that regression, with uncorrelated residuals.

In order to write the mathematical form for this model, let $\mathbf{y}$ be the $p \times 1$ vector containing the p observed variables $y_1, y_2, \ldots, y_p$. Assume that $E(\mathbf{y}) = \mathbf{0}$ and $\text{Cov}(\mathbf{y}, \mathbf{y}') = \boldsymbol{\Sigma}$ where $\boldsymbol{\Sigma}$ is a $p \times p$ covariance matrix. Let $\mathbf{x}$ be an $m \times 1$ vector, the co-ordinates of which are the m unobserved variables $x_1, x_2, \ldots, x_m$, also called the common factors, and assume that $E(\mathbf{x}) = \mathbf{0}$ and $\text{Cov}(\mathbf{x}, \mathbf{x}') = \boldsymbol{\Phi}$. The first definition of the factor analysis model in the previous paragraph can now be expressed mathematically by the expression

$$\boldsymbol{\Sigma}_{y.x} = \boldsymbol{\Sigma} - \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}'$$

(see e.g. Mulaik (1972) for a derivation of this equation) where $\mathbf{\Sigma}_{y.x}$ is the covariance matrix of $\mathbf{y}$ with $\mathbf{x}$ partialled out, and $\mathbf{\Lambda}$ is the $p \times m$ matrix of regression coefficients of $\mathbf{y}$ on $\mathbf{x}$. The partial covariance matrix $\mathbf{\Sigma}_{y.x}$, however, is diagonal by definition. Therefore, let it be denoted by $\mathbf{D}_\Psi$, and consequently the testable common factor model is

$$\mathbf{\Sigma} = \mathbf{\Lambda \Phi \Lambda}' + \mathbf{D}_\Psi. \tag{2.1}$$

The second definition of the model can be stated as

$$\begin{aligned} y_1 &= \lambda_{11}x_1 + \lambda_{12}x_2 + \ldots + \lambda_{1m}x_m + e_1 \\ y_2 &= \lambda_{21}x_1 + \lambda_{22}x_2 + \ldots + \lambda_{2m}x_m + e_2 \\ &\vdots \\ y_p &= \lambda_{p1}x_1 + \lambda_{p2}x_2 + \ldots + \lambda_{pm}x_m + e_p \end{aligned}$$

where λ_{ij} is the regression coefficient of y_i on x_j - also called the common factor loading of variable y_i on factor x_j - and e_i is the residual of y_i about its regression on the common factors, also called the unique factor. These residuals are assumed to be uncorrelated.

In matrix notation it follows that

$$\mathbf{y} = \mathbf{\Lambda x} + \mathbf{e} \tag{2.2}$$

which is the common factor model itself (McDonald, 1985).

The parameters in the factor analysis model that need to be estimated from sampled observations, are the pm factor loadings (regression coefficients) in $\mathbf{\Lambda}$, the $m(m+1)/2$ non-duplicated elements of $\mathbf{\Phi}$ and the p diagonal elements of $\mathbf{D}_\Psi$, i.e. the unique variances. Before the estimation of these parameters by means of the method of maximum

likelihood is presented, the problem of the uniqueness of the parameters will be given attention in the next section.

2.3 Identification in the factor analysis model

A model is said to be identified when the parameters in the model are uniquely determined. In models that are not identified, the estimates of the parameters are arbitrary and have meaningless interpretation (Long, 1983). It is, however, possible to remove this unidentifiability by imposing restrictions on the parameters in the model. In the factor analysis model, the parameters are not uniquely determined and consequently the model is not identified: Let $\mathbf{\Lambda}^\star = \mathbf{\Lambda}\mathbf{T}^{-1}$ and $\mathbf{\Phi}^\star = \mathbf{T}\mathbf{\Phi}\mathbf{T}'$. It follows that

$$\begin{aligned}\mathbf{\Lambda}^\star\mathbf{\Phi}^\star\mathbf{\Lambda}^{\star\prime} + \mathbf{D}_\Psi &= (\mathbf{\Lambda}\mathbf{T}^{-1})(\mathbf{T}\mathbf{\Phi}\mathbf{T}')(\mathbf{\Lambda}\mathbf{T}^{-1})' + \mathbf{D}_\Psi \\ &= \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' + \mathbf{D}_\Psi\end{aligned}$$

which indicates unidentification, since $\mathbf{\Lambda}^\star$ and $\mathbf{\Phi}^\star$ are clearly different from $\mathbf{\Lambda}$ and $\mathbf{\Phi}$ (for all $\mathbf{T} \neq \mathbf{I}$), but the model is unaltered if any of these two sets of parameters is used. It is therefore necessary to impose at least m^2 restrictions on the elements of $\mathbf{\Lambda}$ and $\mathbf{\Phi}$ - since $\mathbf{T}$ is of the order $m \times m$ and has m^2 elements - in order to define them uniquely (Lawley and Maxwell, 1971). In exploratory factor analysis these m^2 restrictions are often imposed in the following way: The factor covariance matrix $\mathbf{\Phi}$ is restricted to be an $m \times m$ identity matrix, imposing $m(m+1)/2$ restrictions on $\mathbf{\Phi}$, while the remaining $m(m-1)/2$ restrictions are imposed on $\mathbf{\Lambda}$ in restricting $\mathbf{\Lambda}'\mathbf{D}_\Psi^{-1}\mathbf{\Lambda}$ to be a diagonal matrix. These restrictions are chosen for convenience only and does generally not result in interpretable solutions. However, this does not matter in exploratory factor analysis since no hypothesis concerning the factors are involved in such an analysis (Jöreskog, 1969); subsequent rotation of the factors is usually performed in an attempt to obtain interpretable solutions.

In situations where some knowledge of the problem under investigation is available - possibly through previous research - the researcher has certain hypotheses regarding the population parameters that he may wish to test. A factor analysis done under these circumstances is called a confirmatory analysis and requires that the values of certain elements in $\mathbf{\Lambda}$, $\mathbf{\Phi}$ and $\mathbf{D}_\Psi$ are specified in advance. For example, if previous results show that variable i has no relationship with factor j, one should specify $\lambda_{ij} = 0$. Also if factors r and s are expected to be uncorrelated, one should specify $\phi_{rs} = 0$. The parameters that are specified in advance are called the *fixed* parameters while the remaining parameters that need to be estimated are called the *free* parameters.

Let n_Λ and n_Φ represent the number of specified parameters in $\mathbf{\Lambda}$ and $\mathbf{\Phi}$ respectively. Then a necessary condition for uniqueness is that

$$n_\Lambda + n_\Phi \geq m^2.$$

This condition, however, is not sufficient, since it is not only the *number* of specified parameters, but also their *position* that is important in defining the free parameters uniquely (see e.g. Everitt (1984)).

One way to determine whether or not a specific model is identified, is to look in detail at the equations relating the elements of the sample covariance matrix to the model parameters. The equations need not be solved; one should only try to assess whether the parameters have unique solutions (Everitt, 1984).

Jöreskog (1979) gives a number of sufficient conditions for $\mathbf{\Lambda}$ to be uniquely determined. These conditions are valid for oblique solutions with fixed zero elements - the most interesting case in practice - and are the following:

i. $\mathbf{\Phi}$ is a symmetric positive definite matrix with unit diagonal elements.

ii. $\mathbf{\Lambda}$ has at least $m - 1$ fixed zeroes in each column.

iii. $\mathbf{\Lambda}_s$ has rank $m - 1$, where $\mathbf{\Lambda}_s$, $s = 1, 2, \ldots, m$ is the submatrix of $\mathbf{\Lambda}$ that contains the rows of $\mathbf{\Lambda}$ which have fixed zero elements in the s-th column.

The fixed unities on the diagonal of $\mathbf{\Phi}$ could be relaxed if one *nonzero* value is fixed in each column of $\mathbf{\Lambda}$, since both these restrictions merely fix the unit of measurement of the factors.

2.4 Parameter estimation in exploratory factor analysis

Several methods for estimating the parameters in a factor analysis model have been developed and can be found in books such as Harman (1976) and Mulaik (1972). A few iterative methods of factor analysis - one of which is principal factor analysis - are discussed in an expository paper by McDonald (1970) while an account of three methods yielding maximum likelihood estimators of the parameters may be found in Browne (1969).

The method of maximum likelihood is frequently employed to obtain estimators of the population parameters. In many cases, these estimators have desirable asymptotic distributional properties. Furthermore, one may perform with them tests of significance on the goodness of fit of factor analysis models (Browne, 1968). The historical development of maximum likelihood estimation in factor analysis may be found in Mulaik (1972) and also in Jackson (1991) and consists of the following major developments.

The theoretical development of maximum likelihood estimation in factor analysis as it is known today is mainly due to the work of Lawley (1940) when he made a major breakthrough in deriving the equations for the maximum likelihood estimators of the parameters. His computational recommendations, however, were not practical for problems with many variables since electronic computers were not yet available to do the complex calculations required by his method. It was later found that even with modern computing facilities his algorithm did not converge effectively in all applications. In the 1950s, however, the first computers became available and methods for factor analytic problems were investigated by a number of workers in this field. Algorithms for obtaining maximum likelihood estimates were provide by Howe (1955) and Rao (1955);

Howe also showed that Lawley's maximum likelihood estimators could be derived from a model for which no distributional assumptions are made.

In the late 1960s the major computational problem in maximum likelihood factor analysis could be seen as something of the past; Jöreskog (1966) used a Fletcher-Powell algorithm in testing a simple structure hypothesis and subsequently used an improved version of the algorithm to minimise the maximum likelihood discrepancy function in factor analysis. The algorithm was found to converge to the desired solution and did so faster than any other algorithm tried before. It should also be mentioned that the Jennrich-Robinson algorithm was developed independently of Jöreskog (1966) and was found to be superior.

The maximum likelihood equations for the parameters in the factor analysis model may be obtained by means of the well known method of maximising the likelihood function of a sample with respect to the unknown population parameters. It has been general practice to maximise the natural logarithm of the likelihood function. The latter may be based only on the information in the sample covariance matrix $\mathbf{S}$ that follows a Wishart distribution under the assumption of multivariate normality of the data vector, or alternatively, the likelihood function may be based on the joint likelihood of the individual observations drawn from the population. Both these approaches have been used to derive the equations for the maximum likelihood estimators; Morrison (1990) and Lawley and Maxwell (1971) use the Wishart distribution of $\mathbf{S}$ while the second approach is used by Anderson (1984) and Mulaik (1972).

A brief indication will now be given as to how the maximum likelihood estimators in exploratory factor analysis may be obtained by maximising the likelihood function based on the joint likelihood of individual observations. Note that the restrictions $\mathbf{\Phi} = \mathbf{I}$ and $\mathbf{\Lambda}'\mathbf{D}_{\Psi}^{-1}\mathbf{\Lambda}$=diagonal are used here. Assume now that a sample of N p-variate vectors of observations has been drawn from a p-variate normal population with $\boldsymbol{\mu} = \mathbf{0}$ and $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{D}_{\Psi}$, and denote the sample values by $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$. The likelihood for this sample is given by

$$L = (2\pi)^{-Np/2}|\boldsymbol{\Sigma}|^{-N/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{N}\mathbf{x}_i'\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right)$$

while the natural logarithm of L is

$$\ell nL = -\frac{Np}{2}\ell n(2\pi) - \frac{N}{2}\ell n|\boldsymbol{\Sigma}| - \frac{1}{2}\text{tr}[\mathbf{A}\boldsymbol{\Sigma}^{-1}] \tag{2.3}$$

where $\mathbf{A} = \sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'$.

The estimators of $\boldsymbol{\Lambda}$ and $\mathbf{D}_\Psi$ are obtained by setting the partial derivatives of ℓnL with respect to $\boldsymbol{\Lambda}$ and $\mathbf{D}_\Psi$ equal to zero. The details of the derivations will be omitted here since it can be found in many references, e.g. Anderson (1984).

The partial derivatives are given by

$$\frac{\partial \ell nL}{\partial \psi_i} = -\frac{N}{2}\left([\boldsymbol{\Sigma}^{-1}]_{ii} - [\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}]_{ii}\right)$$

and

$$\frac{\partial \ell nL}{\partial \lambda_{ij}} = -N\left([\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ij} - [\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]_{ij}\right).$$

Equating these to zero, the following expressions are obtained for $\hat{\boldsymbol{\Lambda}}$ and $\hat{\mathbf{D}}_\Psi$:

$$\text{diag}(\hat{\boldsymbol{\Sigma}}^{-1}) = \text{diag}(\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{S}\hat{\boldsymbol{\Sigma}}^{-1})$$

and

$$\mathbf{S}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Lambda}} = \hat{\boldsymbol{\Lambda}}.$$

The above two expressions may be simplified further to yield

$$\text{diag}(\hat{\boldsymbol{\Sigma}}) = \text{diag}(\mathbf{S}) \tag{2.4}$$

and

$$\mathbf{S}\hat{\mathbf{D}}_{\Psi}^{-1}\hat{\boldsymbol{\Lambda}} = \hat{\boldsymbol{\Lambda}}(\mathbf{I} + \hat{\boldsymbol{\Gamma}}) \tag{2.5}$$

where $\hat{\boldsymbol{\Gamma}} = \hat{\boldsymbol{\Lambda}}'\hat{\mathbf{D}}_{\Psi}^{-1}\hat{\boldsymbol{\Lambda}}$ and is therefore a diagonal matrix.

Equation (2.5) may be manipulated and written in the form

$$\left(\hat{\mathbf{D}}_{\Psi}^{-\frac{1}{2}}(\mathbf{S} - \hat{\mathbf{D}}_{\Psi})\hat{\mathbf{D}}_{\Psi}^{-\frac{1}{2}}\right)\left(\hat{\mathbf{D}}_{\Psi}^{-\frac{1}{2}}\hat{\boldsymbol{\Lambda}}\right) = \left(\hat{\mathbf{D}}_{\Psi}^{-\frac{1}{2}}\hat{\boldsymbol{\Lambda}}\right)\hat{\boldsymbol{\Gamma}}$$

which shows that the columns of $\hat{\mathbf{D}}_{\Psi}^{-\frac{1}{2}}\hat{\boldsymbol{\Lambda}}$ are the characteristic vectors of $\hat{\mathbf{D}}_{\Psi}^{-\frac{1}{2}}(\mathbf{S} - \hat{\mathbf{D}}_{\Psi})\hat{\mathbf{D}}_{\Psi}^{-\frac{1}{2}}$ and $\hat{\boldsymbol{\Gamma}}$ contains the corresponding roots.

Since the estimators $\hat{\boldsymbol{\Lambda}}$ and $\hat{\mathbf{D}}_{\Psi}$ cannot be obtained in closed form, use must be made of iterative procedures to obtain the estimates. One such method - the Fletcher and Powell method - is described by Lawley and Maxwell (1971). Other methods that are also used include the Newton-Raphson and Gauss-Newton methods (Bentler, 1986).

In the process of estimating the parameters in exploratory factor analysis, it is argued implicitly that the number of factor variables, m, is fixed and known. This, however,

is not always the case in practice. Consequently there is a need to test if m factors are sufficient for equation (2.1) to hold. The answer to this problem may be found by using the likelihood ratio principle to derive an appropriate test for the null hypothesis that $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{D}_{\Psi}$ where the order of $\mathbf{\Lambda}$ is $p \times m$.

The test statistic is given by (see e.g. Morrison (1990))

$$\lambda_m = K\ell n\frac{|\hat{\mathbf{\Sigma}}|}{|\mathbf{S}|}$$

where

$$K = N - \frac{1}{6}(2p + 11) - \frac{2}{3}m.$$

The test statistic λ_m obtained in this way is asymptotically distributed as a chi-square variate with ν degrees of freedom where

$$\nu = \frac{1}{2}[(p - m)^2 - (p + m)].$$

Often a different function is optimised to obtain the maximum likelihood estimators and to simultaneously obtain the likelihood ratio test statistic described above. This function is

$$F = \ell n|\mathbf{\Sigma}| + \text{tr}(\mathbf{S}\mathbf{\Sigma}^{-1}) - \ell n|\mathbf{S}| - p \tag{2.6}$$

whose minimum will yield the same maximum likelihood estimators of the parameters than the maximum of ℓnL and also, $N-1$ times the minimum value of F is the likelihood ratio test statistic of goodness of fit (Jöreskog, 1969).

An estimate of the number of factors may now be obtained by using a sequence of likelihood ratio tests, each time extracting a larger number of factors, until the null hypothesis is not rejected. The value of m, for which the null hypothesis is not rejected for the first time, is taken as an estimate for the number of factors. It should be mentioned that such a series of tests are not independent and that the true significance level may be very different from the level used with each test (Morrison, 1990).

2.5 Factor rotation

Once an initial solution of $\mathbf{\Lambda}$ has been obtained in exploratory factor analysis, it may be transformed into a different solution without changing its ability to represent the observed covariances. This can easily be seen to be true since for any $m \times m$ orthogonal matrix $\mathbf{T}$ it follows that

$$\mathbf{\Lambda T}(\mathbf{\Lambda T})' = \mathbf{\Lambda TT'\Lambda'} = \mathbf{\Lambda\Lambda'} \tag{2.7}$$

since $\mathbf{TT'} = \mathbf{I}$. An unlimited number of solutions can therefore be obtained for the factor matrix once an initial solution has been obtained. Also, because of (2.7), any one of this unlimited number of solutions equally well reproduces the covariances among the observed variables. It is also evident that the test statistic for testing the hypothesis of sufficient number of factors will be identical for all such solutions. The only major difference between these solutions is the complexity, and therefore the interpretability, of the factors. It is this issue, namely the interpretability of the factors, that factor rotation addresses.

To assist in understanding the rotation of factors, the rationale behind rotation will be demonstrated for the case of two factors. An initial factor matrix for this special case contains a pair of coefficients (one for each factor) for each variable. These coefficients may act as co-ordinates to represent the variables in a two-dimensional space that is

called the common-factor space (Harman, 1976). This space consists of a pair of axes which are at right angle with each other since the initial factors are uncorrelated. The points in this space, which represent the variables, may be transformed to another solution by rotating the axes through some angle while retaining the orthogonality of the axes. The co-ordinates of the points with respect to these new axes now represent the coefficients of some other solution. It is shown by Harman (1976) that such a transformation from one set of axes to another may be put in the form

$$\mathbf{B} = \mathbf{AT}$$

where $\mathbf{B}$ and $\mathbf{A}$ are of the order $p \times 2$ and are respectively the coefficients of the new and the initial solution. The matrix $\mathbf{T}$ is of the order 2×2 and is called the transformation matrix because it transforms one solution into another. Harman continues and shows that this reasoning may be extended to cases where there are more than two factors.

At first, the rotation of factors was done in a subjective manner. A review of such subjective, graphical transformations to obtain a rotated solution from some initial solution can be found in Harman (1976) and Mulaik (1972). It was only after the development of large electronic computers that objective, analytic methods could be used, because these methods require extensive computations.

Using an objective factor rotation method to obtain interpretable factors, one needs some kind of criterion, or set of criteria, to evaluate the rotated factors. Of these criteria, the most important are the five criteria for simple structure proposed by Thurstone in 1947. They are given in Morrison (1990), Harman (1976) and Mulaik (1972). The last of these references also includes a technical explanation of what is meant by simple structure.

Several factor analysts came up with criteria for use in orthogonal analytic rotation. Among the first of these criteria was the one arrived at by Carroll in 1953. It involves the minimisation of fourth degree terms obtained from cross-products of squared factor

loadings. The criterion he proposed is that f be a minimum, where

$$f = \sum_{k<\ell=1}^{m} \sum_{i=1}^{p} b_{ik}^2 b_{i\ell}^2.$$

A year later Ferguson (see Harman(1976)) suggested that the sum of the fourth powers of all factor loadings should be a maximum, i.e. q should be a maximum where

$$q = \sum_{i=1}^{p} \sum_{j=1}^{m} b_{ij}^4. \tag{2.8}$$

In 1954 Neuhaus and Wrigley (Harman, 1976) proposed a criterion in which the variance of the squared loadings of a variable is maximised. This involves the maximisation of fourth powers of factor loadings, and consequently - following C. Burt's suggestion (Harman, 1976) - the method is termed the "quartimax" method. The criterion proposed by Niehaus and Wrigley, however, is equivalent to Ferguson's criterion in (2.8) under orthogonal transformation.

In 1958 Kaiser introduced a criterion that he called the "varimax". It has been used with great success, and is still probably the most widely employed criterion in orthogonal analytic rotation. Harman (1976) and Malaik (1972) provide much detail on the development of these criteria. The varimax criterion will be discussed briefly here.

Since the quartimax criterion is concerned with the simplification of the *rows* of a factor matrix and would frequently give a general factor, Kaiser suggested a criterion to simplify the *columns* of a factor matrix. As a measure of a certain factor's simplicity, the variance of the squared loadings of the observed variables on that factor was taken. This variance is maximised so that there will only be a small number of large loadings on the factor, which will increase its interpretability. The criterion therefore proposed by Kaiser is the maximum of the sum, over all factors, of these variances. This criterion can be written as

$$v^{\star} = \sum_{j=1}^{m} \left(\frac{1}{p} \sum_{i=1}^{p} (b_{ij}^2)^2 - \frac{1}{p^2} (\sum_{i=1}^{p} b_{ij}^2)^2 \right) \tag{2.9}$$

where $\mathbf{B} = (b_{ij})$ is the rotated factor matrix.

The use of the criterion given by (2.9) was only moderately successful if measured against intuitive-graphical methods. It was then that (2.9) was modified, at D.R. Saunders' suggestion (Harman, 1976), by weighting the observed variables equally in the rotation. To accomplish this, the rows of $\mathbf{B}$ are normalised to unit-length vectors by dividing the loadings of each observed variable (which form the rows of $\mathbf{B}$) by the square root of that variable's communality. These weighted loadings are then used to replace the unweighted loadings in (2.9), which leads to the function for maximisation

$$v = \sum_{j=1}^{m} \left(\frac{1}{p} \sum_{i=1}^{p} \left(\frac{b_{ij}^2}{h_i^2} \right)^2 - \frac{1}{p^2} \left(\sum_{i=1}^{p} \frac{b_{ij}^2}{h_i^2} \right)^2 \right). \tag{2.10}$$

When the maximum of this function has been obtained, the rotated loadings are re-weighted so that the rows of $\mathbf{B}$ assume their original lengths. This is accomplished by multiplying each variable's loadings by the square root of that variable's communality.

The implementation of (2.10) instead of (2.9) resulted in considerably improved orthogonal simple structure solutions when compared with intuitive or graphic solutions.

Before we move on to oblique rotation, it is necessary to give meaning to the concepts "primary" and "reference", since they form different bases for two types of oblique solutions. In the case of m factors, the m oblique primary axes are those that pass through the centroids of clusters of variables, while the m oblique reference axes are each normal to the m co-ordinate hyperplanes of $m - 1$ dimensions (Harman, 1976). An oblique solution is now said to be a primary factor solution if the coefficients in the solution are the co-ordinates of the variables with respect to the primary axes. An oblique reference axes solution is defined similarly in terms of the reference axes.

Different criteria were proposed by several people for obtaining an objective oblique solution. Carroll (1953) was among the first to propose the same criterion for orthogonal and for oblique rotation. This method was named "quartimin" since, as was mentioned previously, it involves the minimisation of fourth degree terms. Harman (1976) and Mulaik (1972) provide some detail on a number of other objective oblique rotation methods and criteria that were proposed and applied in practice by different authors.

These criteria were all used to obtain a simple structure within the reference structure matrix, i.e. the matrix of correlations between the variables and the reference axes. The resulting solution is then used to obtain the primary factor pattern through a transition formula that simply requires multiplication by a diagonal matrix (see e.g. Harman (1976)).

Jennrich and Sampson (1966) made an important breakthrough by deriving an analytical procedure to rotate an initial factor matrix directly into a primary factor solution. They applied this procedure to the quartimin criterion after which it got the name "direct quartimin". A consequence of this new development of directly obtaining the primary factor solution is that the earlier procedure of first obtaining the reference structure solution is now only of historical interest and is not recommended in practice.

2.6 Parameter estimation in confirmatory factor analysis

After identification of the model in confirmatory factor analysis has been established, one may proceed to estimate the free parameters in the model and subsequently test the model for fit. In this case the function that is minimised, F (see expression (2.6)), is considered as a function of only the free parameters in $\mathbf{\Lambda}$, $\mathbf{\Phi}$ and $\mathbf{D}_\Psi$. The first-order derivatives of F with respect to the free parameters are given by the expressions (see for example Lawley and Maxwell (1971) and Jöreskog (1969))

$$\frac{\partial F}{\partial \mathbf{\Lambda}} = 2\mathbf{\Sigma}^{-1}(\mathbf{\Sigma} - \mathbf{S})\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Phi},$$

$$\frac{\partial F}{\partial \mathbf{\Phi}} = c\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}(\mathbf{\Sigma} - \mathbf{S})\mathbf{\Sigma}^{-1}\mathbf{\Lambda}$$

and

$$\frac{\partial F}{\partial \mathbf{D}_\Psi} = \text{diag}\left[\mathbf{\Sigma}^{-1}(\mathbf{\Sigma} - \mathbf{S})\mathbf{\Sigma}^{-1}\right]$$

where c=1 for diagonal elements of $\mathbf{\Phi}$ and c=2 for non-diagonal elements. Note that the elements in the derivative matrices on the left that correspond to the fixed parameters in $\mathbf{\Lambda}$, $\mathbf{\Phi}$ and $\mathbf{D}_\Psi$ are taken to be zero.

An iterative procedure is described in Jöreskog (1969) where the first-order derivative information is used as well as expectations of the second-order derivatives. The procedure starts with a few steepest descent iterations, and only then the second-order derivative information is calculated. Subsequent iterations are then performed using the Fletcher and Powell method.

It is also pointed out by Jöreskog (1969) that after the minimum value of F has been obtained, say $\hat{F}$, a test for the goodness of fit of $\hat{\mathbf{\Sigma}}$ to the sample covariance matrix $\mathbf{S}$ is given by the likelihood ratio test statistic $(N-1)\hat{F}$. This statistic is asymptotically distributed as χ^2 with degrees of freedom equal to

$$\nu = p(p+1)/2 - pm - m(m+1)/2 - p + n_\psi + \sum_{i=1}^{m} \max(r_i, m)$$

where n_ψ is the number of fixed parameters in $\mathbf{D}_\Psi$ and r_i is the number of independent restrictions on the i-th factor, including the restrictions on the ϕ_{ii} and ϕ_{ij} in the factor covariance matrix $\mathbf{\Phi}$.

Practical problems with the use of this test in assessing the goodness of fit of a model are pointed out by Bentler and Bonett (1980). The main reason for these problems is the role of the sample size in calculating the test statistic. Consequently it is suggested by Jöreskog and Sörbom (1981) that this test should only be used to give an indication of fit, rather than as a formal test of a hypothesis. It is recommended that differences

in the test statistics for two models, the one nested within the other, are used. Such a difference is also distributed as χ^2, with degrees of freedom equal to the difference in degrees of freedom of the two models.

Yet another suggestion made by Jöreskog and Sörbom (1981) is to normalize the residuals $(s_{ij} - \hat{\sigma}_{ij})$ by the square root of their asymptotic variances, estimated by $(\hat{\sigma}_{ii}\hat{\sigma}_{jj} + \hat{\sigma}_{ij}^2)/N$. These normalized residuals should be approximately normally distributed if the model fits the data adequately and could therefore be used to assess the goodness of fit (see also Everitt (1984)).

A further measure of fit - one that is used in RAMONA (Browne and Mels, 1990) - is the Root Mean Square Error of Approximation (RMSEA). This measure gives an indication of the fit of the model to the *population* covariance matrix; one would obtain this by the minimum value of the population discrepancy function, say F_0, which is the minimum value of F with $\mathbf{S}$ replaced by the population covariance matrix $\mathbf{\Sigma}_0$. Since F_0 cannot be calculated, an estimate should be employed: Such an estimate is given by $\hat{F} - d/N$ where $\hat{F}$ is the minimum of the sample discrepancy function and d is the degrees of freedom. The RMSEA is now defined as $\sqrt{F_0/d}$ (Steiger and Lind, 1980) which adjusts F_0 for the number of parameters, and does not involve the sample size. The estimated RMSEA value, RMSEA_e, may be calculated as

$$\text{RMSEA}_e = \sqrt{\frac{\hat{F} - d/N}{d}}$$

and a confidence interval on the RMSEA may be obtained from a confidence interval on the noncentrality parameter of a χ^2 distribution (Steiger and Lind, 1980).

In order to obtain accurate estimates of the standard errors of the parameter estimates, Jöreskog (1969) suggests that the second-order derivative matrix should be calculated at the minimum of F and inverted. If this matrix is denoted by $\mathbf{C}$ with diagonal elements denoted by c_{ii}, $i = 1, 2, \ldots, q$ - where q is the number of free parameters - then an approximate 95% confidence interval for the i-th free parameter, say θ_i, is given by

$$\hat{\theta}_i - 2\sqrt{(2/(N-1))c_{ii}} < \theta_i < \hat{\theta}_i + 2\sqrt{(2/(N-1))c_{ii}}$$

where the understanding is that this formula should only be used when the model is identified.

2.7 Practical applications

The data that will be used to illustrate the use of factor analysis in practice were provided by the Education division of the Human Sciences Research Council. The data were captured in 1994 for the purpose of doing item analyses on some tests on cognitive skills. The respondents are students still in school and the total sample size is N=5635. For the first application, six highly correlated tests are selected and only one factor is assumed to account for the covariances between them. In the second application, six additional highly correlated tests are selected, not strongly correlated with the first six, and in this example two factors are assumed to reproduce their covariance matrix sufficiently. In both applications the analysis is done using the covariance matrix instead of the correlation matrix. This is done so that the results obtained here can be compared to further analyses on the same data in Chapters 4 and 5, where the covariance matrix is analysed.

Example 2.7.1: One factor

Let the six variables for this example be denoted by $y_1, y_2, \ldots, y_6$ and let these components make up the six-variate vector $\mathbf{y}$. Assuming that only one factor is sufficient to explain the covariances among the six variables, the mathematical model to be fitted to the data obtained from the 5635 students is that the relation

$$\mathbf{y} = \mathbf{\Lambda} x + \mathbf{e} \tag{2.11}$$

holds for each student where x is a scalar representing the one factor variable, $\mathbf{\Lambda}$ is a 6×1 column vector containing the six factor loading parameters and $\mathbf{e}$ is the 6×1 column vector of errors. It is assumed that the random vector $\mathbf{y}$ follows a six-variate normal distribution with covariance matrix

$$\text{Cov}(\mathbf{y}, \mathbf{y}') = \mathbf{\Lambda}\phi\mathbf{\Lambda}' + \mathbf{D}_\Psi \tag{2.12}$$

since it is assumed that x and $\mathbf{e}$ are independent variates and that $x \sim N(0, \phi)$ and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{D}_\Psi)$ where $\mathbf{D}_\Psi$ is the 6×6 diagonal matrix containing the six error variances $\Psi_1, \Psi_2, \ldots, \Psi_6$ on the diagonal.

Since there is only one factor, one identification condition is required to obtain a unique solution. This is done by fixing the scale of x through fixing ϕ=1.

The observed data obtained from the sample of 5635 students are now used to obtain the maximum likelihood estimates $\hat{\mathbf{\Lambda}}$ and $\hat{\mathbf{D}}_\Psi$ of the unknown population parameters $\mathbf{\Lambda}$ and $\mathbf{D}_\Psi$. The program RAMONA (Browne and Mels, 1990) was used to obtain these estimates from the sample covariance matrix which is given in Table 2.1.

TABLE 2.1

Sample Covariance Matrix - Six Variables

	y_1	y_2	y_3	y_4	y_5	y_6
y_1	1,342					
y_2	0,840	1,121				
y_3	0,939	0,811	1,186			
y_4	0,834	0,716	0,848	1,234		
y_5	0,678	0,581	0,673	0,672	0,823	
y_6	0,713	0,616	0,696	0,674	0,616	0,906

The results obtained from this analysis are provided in Table 2.2 and Table 2.3.

TABLE 2.2
Matrix (Column Vector) of Factor Loadings and Standard Errors

	Factor1
y_1	0,964 (0,013)
y_2	0,833 (0,012)
y_3	0,949 (0,012)
y_4	0,886 (0,013)
y_5	0,731 (0,010)
y_6	0,759 (0,011)

The above matrix of factor loadings shows that all loadings are highly significant, due to their small standard errors.

TABLE 2.3
Residual Covariance Matrix - Six Variables

	y_1	y_2	y_3	y_4	y_5	y_6
y_1	0,000					
y_2	0,037	0,000				
y_3	0,024	0,020	0,000			
y_4	−,020	−,022	0,007	0,000		
y_5	−,027	−,028	−,021	0,024	0,000	
y_6	−,019	−,017	−,024	0,002	0,061	0,000

In the table above, the residual covariances are provided. The estimates $\hat{\Psi}_1, \hat{\Psi}_2, \ldots, \hat{\Psi}_6$, which are the error variance estimates of the observed variables, and their estimated standard errors are given below in Table 2.4.

TABLE 2.4

Variance Estimates and their estimated Standard Errors

$\hat{\Psi}_i$	Std.Error
0,412	0,010
0,427	0,009
0,285	0,007
0,449	0,010
0,288	0,006
0,330	0,007

In addition to the parameter estimates and their estimated standard errors, RAMONA provides different measures for assessing the goodness of fit of the model. One is the RMSEA; a point estimate of 0,097 is obtained, indicating that the fit is fairly reasonable - Browne and Mels (1990) noted that "A value of about 0,08 or less indicates a reasonable fit of the model in relation to the degrees of freedom." RAMONA also provides a 90% confidence interval for the RMSEA, being given as $(0,090\ ;\ 0,105)$.

Another methodology for assessing the goodness of fit is the accept/reject strategy using the test statistic, $(N-1)\hat{F}$, which has a limiting chi-square distribution with 9 degrees of freedom. For this data, the value of the test statistic is 489,6, implying rejection of the null hypothesis that the model fits. Keeping in mind, however, the influence of the large sample on this testing procedure, and looking at the small residual covariances in Table 2.3, lead to the conclusion that the fit of the model is not at all that bad and one should not blindly reject the model.

Example 2.7.2: Two factors

In this example, six additional highly correlated variables, only moderately correlated with the six variables of the previous example, are considered. Together, twelve variables are therefore considered and two factors are assumed to be sufficient due to high correlations between the six variables in each of the two groups of variables and low correlations between the two groups of variables.

Let the 12×1 vector $\mathbf{y}$ contain these twelve variables and assume the model

$$\mathbf{y} = \mathbf{\Lambda x} + \mathbf{e}.$$

Now, $\mathbf{\Lambda}$ is a 12×2 matrix of factor loadings, $\mathbf{x}$ is a 2×1 vector containing the two factor variables and $\mathbf{e}$ is a 12×1 vector of error terms. If $\mathbf{x} \sim N(\mathbf{0}, \mathbf{\Phi})$ and, independently, $\mathbf{e} \sim N(\mathbf{0}, \mathbf{D}_\Psi)$, then the structured covariance matrix of $\mathbf{y}$ follows as

$$\text{Cov}(\mathbf{y}, \mathbf{y}') = \mathbf{\Lambda \Phi \Lambda}' + \mathbf{D}_\Psi.$$

Since there are two factors ($m = 2$) in this application, the model is unidentified and m^2=4 identification conditions need to be imposed on the parameters. In order to fix the scale of the factors, their variances are fixed at one. The remaining two conditions are chosen so that there are two reference variables to represent the two factors. y_1 is chosen to represent the first factor and y_7 to represent the second factor. Therefore, λ_{12} and λ_{71} are additionally fixed at zero to obtain a unique solution.

The sample covariance matrix of $\mathbf{y}$, using observations made on $\mathbf{y}$ for the same 5635 students of the previous example, is provided in Table 2.5. RAMONA was again used to obtain the maximum likelihood estimates of the free parameters in the model.

TABLE 2.5

Sample Covariance Matrix - Twelve Variables

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
y_1	1,342									
y_2	0,840	1,121								
y_3	0,939	0,811	1,186							
y_4	0,834	0,716	0,848	1,234						
y_5	0,678	0,581	0,673	0,672	0,823					
y_6	0,713	0,616	0,696	0,674	0,616	0,906				
y_7	−0,827	−0,608	−0,741	−0,756	−0,680	−0,601	9,638			
y_8	−1,006	−0,814	−1,038	−1,030	−0,823	−0,804	5,860	21,011		
y_9	−1,676	−1,318	−1,578	−1,637	−1,524	−1,260	8,268	11,068	26,296	
y_{10}	−1,386	−1,091	−1,285	−1,393	−1,295	−1,063	6,450	9,925	15,000	21,948
y_{11}	−1,224	−0,998	−1,184	−1,258	−1,167	−0,930	6,491	9,959	14,705	14,952
y_{12}	−1,309	−1,021	−1,253	−1,183	−0,998	−0,916	5,269	7,925	12,200	9,178

	y_{11}	y_{12}
y_{11}	22,176	
y_{12}	9,370	20,928

The maximum likelihood parameter estimates obtained from analysing the above covariance matrix are provided in Tables 2.6 and 2.7 and 2.8. The estimates $\hat{\Psi}_1, \hat{\Psi}_2, \ldots, \hat{\Psi}_{12}$ and estimates of their standard errors are provided in Table 2.9.

TABLE 2.6

Matrix of Factor Loadings and Standard Errors

	Factor1	Factor2
y_1	0,965 (0,013)	0*
y_2	0,849 (0,014)	0,035 (0,014)
y_3	0,957 (0,013)	0,016 (0,013)
y_4	0,870 (0,014)	−,035 (0,014)
y_5	0,702 (0,011)	−,066 (0,011)
y_6	0,763 (0,012)	0,010 (0,012)
y_7	0*	1,871 (0,040)
y_8	0,152 (0,082)	2,800 (0,070)
y_9	0,060 (0,099)	4,128 (0,075)
y_{10}	0,227 (0,092)	3,825 (0,069)
y_{11}	0,392 (0,093)	3,875 (0,071)
y_{12}	−,153 (0,080)	2,629 (0,069)

A '⋆' indicates a fixed parameter value.

TABLE 2.7

Factor Covariance Matrix

	Factor1	Factor2
Factor1	1,000*	
Factor2	−,435 (0,022)	1,000*

A '⋆' indicates a fixed parameter value.

TABLE 2.8

Residual Covariance Matrix - Twelve Variables

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}
y_1	0,000											
y_2	0,036	0,000										
y_3	0,023	0,019	0,000									
y_4	−,020	−,021	0,008	0,000								
y_5	−,027	−,027	−,021	0,023	0,000							
y_6	−,019	−,017	−,024	0,003	0,063	0,000						
y_7	−,043	0,017	0,007	0,016	0,015	0,000	0,000					
y_8	0,022	−,005	−,063	−,009	0,106	−,020	0,746	0,000				
y_9	−,003	0,011	0,015	0,013	−,035	0,019	0,594	−,155	0,000			
y_{10}	−,001	−,002	0,028	−,015	−,040	−,007	−,522	−,293	−,298	0,000		
y_{11}	0,023	−,029	−,007	−,005	−,014	0,018	−,441	−,220	−,512	1,072	0,000	
y_{12}	−,059	−,015	−,056	0,038	0,090	0,044	0,227	0,575	1,152	−,838	−,567	0,000

TABLE 2.9

Variance Estimates and their Estimated Standard Errors

$\hat{\Psi}_i$	Std.Error
0,410	0,010
0,425	0,009
0,283	0,007
0,449	0,010
0,285	0,006
0,331	0,007
6,138	0,126
13,517	0,278
9,468	0,242
8,018	0,206
8,323	0,213
13,645	0,277

The point estimate of the RMSEA for this example is 0,060, indicating that the fit of the model is quite good - Browne and Mels (1990) report that a value of about 0,05 or less indicates a close fit. In terms of the chi-squared test statistic, a value of 903,25 with 43 degrees of freedom is obtained by RAMONA. This value is highly significant, but since the large sample is greatly responsible for this, and since the residuals indicate a satisfactory fit, the model should not be rejected.

2.8 Summary

This chapter gives an overview of factor analysis as a data analytic technique. The model that has been developed and used over many years, is provided together with the assumptions that go with the model. The problem of identification in the factor analysis model is discussed, and the (different) identification conditions that are imposed in exploratory and confirmatory factor analysis are considered.

The maximum likelihood method of parameter estimation is described as it applies to the exploratory factor analysis model; a test statistic is also given for testing whether the number of factors being extracted is sufficient. The rationale behind factor rotation is discussed, and some indication is given on the development of orthogonal and oblique rotation criteria over the years.

Maximum likelihood estimation of the free parameters in a confirmatory factor analysis model is considered. Also, a few methods for assessing the goodness of fit of the model are discussed.

Finally, real life data are analysed by RAMONA - extracting one and two factors respectively from six and twelve variables. In the second example (two factors), a confirmatory factor analysis model was used.

CHAPTER 3

INTRODUCTION TO MULTILEVEL MODELS

3.1 Introduction

This chapter deals with a brief introduction to the concept of hierarchically structured populations and the linear mathematical models that are fitted to data obtained from them. According to Goldstein and McDonald (1988), it has been shown that it may be misleading to ignore the hierarchical structure in data where such a structure is present.

In real life, data are often obtained in systems in which there are built-in hierarchies. An example is education, where pupils are grouped into classrooms and the classrooms are grouped into schools. This is a typical three-level hierarchy: the pupils are the level-one units while the classrooms and the schools are the level-two and three units respectively. Models that specifically take this kind of hierarchy in the data into account are called multilevel models.

In recent years a few particular models for analysing certain types of multilevel data have been described in the literature (Goldstein and McDonald, 1988), while a general multilevel model that includes the particular models as special cases was developed by these authors. This model is discussed in the following section. Thereafter, two methods to estimate the parameters in a multilevel model are presented. The first is the marginal maximum likelihood method (MML method) in which the parameter estimates may be obtained by means of expected maximisation. This is an iterative procedure, also called the EM algorithm. The second method is the Fisher scoring method, also an iterative procedure, that is used to obtain maximum likelihood estimates of the parameters. The two methods will be described for two situations: one is when no constraints are imposed on the parameters; the other, when equality constraints are imposed on some of them. Thereafter a brief discussion is given of some existing work in this field, with

the emphasis on multilevel models with latent variables.

3.2 A general multilevel model

This section presents a brief discussion of the general multilevel model that was considered by Goldstein and McDonald (1988). Assume that the population under consideration has a hierarchical structure with h levels of nesting, and let the $N \times 1$ vector $\mathbf{y}$ represent the responses of a sample of N individuals on a single response variable. Let the elements of $\mathbf{y}$ be so ordered that $\mathbf{y}$ may be partitioned in h ways, each according to one of the h levels of the hierarchy. For example, in the case of three levels, say there are $n_1, n_2, \ldots, n_6$ students in six classrooms that are in two schools, the three ways of partitioning $\mathbf{y}$ are (a) $(y_{111}\ y_{112} \cdots y_{11n_1}; y_{121}\ y_{122} \cdots y_{12n_2}; \cdots; y_{231}\ y_{232} \cdots y_{23n_6})'$ for the student level, (b) $(\mathbf{y}'_{11}\ \mathbf{y}'_{12}\ \mathbf{y}'_{13}; \mathbf{y}'_{21}\ \mathbf{y}'_{22}\ \mathbf{y}'_{23})'$ for the classroom level, and (c) $(\mathbf{y}'_1\ \mathbf{y}'_2)'$ for the school level.

The linear model is written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$$

and in the case of p unknown coefficients, $\mathbf{X}$ is an $N \times p$ design matrix, and $\boldsymbol{\beta}$ the $p \times 1$ vector consisting of the unknown coefficients. To suit the hierarchical structure however, this model may be rewritten after partitioning $\mathbf{X}$ and $\boldsymbol{\beta}$ in a suitable form for the hierarchical design. According to Goldstein and McDonald (1988) this general linear multilevel model is written as

$$\mathbf{y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \sum_{k=1}^{h} \mathbf{X}_k\boldsymbol{\beta}_k. \tag{3.1}$$

In this model provision is made for identifying the fixed parameters as well as the random parameters that are regarded random at the different levels. The first term in

(3.1) contains the fixed part of the model with $\boldsymbol{\beta}_0$ a $p_0 \times 1$ vector of fixed parameters. The second term in (3.1), which is a summation of h components, contains the random part of the model. In this part, $\boldsymbol{\beta}_k$ contains the parameters that are random at the k-th level of the hierarchy. If there are n_k units at the k-th level, each $\boldsymbol{\beta}_k$ is partitioned into n_k subvectors, and if there are p_k random parameters at level k, each of the n_k subvectors has p_k components.

Assumptions made regarding the random parameters in model (3.1) are, first, that the n_k subvectors of $\boldsymbol{\beta}_k$, each of the order $p_k \times 1$, are independently and identically distributed with zero mean and covariance matrix $\boldsymbol{\Omega}_k$, of the order $p_k \times p_k$. Consequently it follows that the covariance matrix of $\boldsymbol{\beta}_k$ is given by

$$\text{Cov}(\boldsymbol{\beta}_k, \boldsymbol{\beta}_k') = \mathbf{I}_{n_k} \otimes \boldsymbol{\Omega}_k.$$

Secondly, it is assumed that random parameters at different levels of the hierarchy are uncorrelated; or, equivalently stated, it is assumed that

$$\text{Cov}(\boldsymbol{\beta}_k, \boldsymbol{\beta}_m') = \mathbf{0}\,, \quad k \neq m.$$

An extension of the univariate model in (3.1) will now be given by considering p-variate vectors of observations in the case of a two-level hierarchical structure. Here also, Goldstein and McDonald (1988) will serve as reference.

Let $\mathbf{y}$ indicate the $Np \times 1$ observed vector where it is assumed that there are M groups (level two-units) and n_i individuals (level one-units) within the i-th group. Further let $N = \sum_{i=1}^{M} n_i$ be the total number of level one-units. A general multivariate two-level model is now given by the expression

$$\mathbf{y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 \tag{3.2}$$

where the p-variate vectors of error terms associated with the two levels of the hierarchy are incorporated into the $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ vectors, and the first term contains the fixed parameters.

Taking into account the way the vector $\mathbf{y}$ is partitioned, the model in (3.2) may be rewritten in extended notation as

$$\begin{pmatrix} \mathbf{y}_{11} \\ \vdots \\ \mathbf{y}_{1n_1} \\ \mathbf{y}_{21} \\ \vdots \\ \mathbf{y}_{2n_2} \\ \vdots \\ \mathbf{y}_{M1} \\ \vdots \\ \mathbf{y}_{Mn_M} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_{01}\boldsymbol{\beta}_{01} \\ \mathbf{X}_{02}\boldsymbol{\beta}_{02} \\ \vdots \\ \mathbf{X}_{0M}\boldsymbol{\beta}_{0M} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\beta}_{111} \\ \vdots \\ \boldsymbol{\beta}_{11n_1} \\ \boldsymbol{\beta}_{121} \\ \vdots \\ \boldsymbol{\beta}_{12n_2} \\ \vdots \\ \boldsymbol{\beta}_{1M1} \\ \vdots \\ \boldsymbol{\beta}_{1Mn_M} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\beta}_{21} \\ \vdots \\ \boldsymbol{\beta}_{21} \\ \boldsymbol{\beta}_{22} \\ \vdots \\ \boldsymbol{\beta}_{22} \\ \vdots \\ \boldsymbol{\beta}_{2M} \\ \vdots \\ \boldsymbol{\beta}_{2M} \end{pmatrix}$$

where $\mathbf{y}_{ij}$ is the $p \times 1$ observed vector for the j-th level one unit in the i-th level-two unit.

From the above extended form of the model it follows that for the i-th level-two unit the model may be written as

$$\begin{pmatrix} \mathbf{y}_{i1} \\ \vdots \\ \mathbf{y}_{in_i} \end{pmatrix} = \mathbf{X}_{0i}\boldsymbol{\beta}_{0i} + \begin{pmatrix} \boldsymbol{\beta}_{1i1} \\ \vdots \\ \boldsymbol{\beta}_{1in_i} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\beta}_{2i} \\ \vdots \\ \boldsymbol{\beta}_{2i} \end{pmatrix}$$

or

$$\mathbf{y}_i = \mathbf{X}_{0i}\boldsymbol{\beta}_{0i} + \mathbf{X}_{1i}\boldsymbol{\beta}_{1i} + \mathbf{X}_{2i}\boldsymbol{\beta}_{2i} \tag{3.3}$$

where, in this expression,

$$\mathbf{X}_{1i} = \mathbf{I}_{n_i} \otimes \mathbf{I}_p = \mathbf{I}_{pn_i}$$

and

$$\mathbf{X}_{2i} = \mathbf{j}_{n_i} \otimes \mathbf{I}_p.$$

In (3.3) the $pn_i \times 1$ random vector $\boldsymbol{\beta}_{1i}$ consists of the n_i subvectors $\boldsymbol{\beta}_{1i1} \ldots \boldsymbol{\beta}_{1in_i}$, each of the order $p \times 1$, for which independent and identical distributions with common covariance matrix $\boldsymbol{\Omega}_1$, of the order $p \times p$, are assumed. The $p \times 1$ random vector $\boldsymbol{\beta}_{2i}$ is assumed to be distributed with $p \times p$ covariance matrix $\boldsymbol{\Omega}_2$. It is also assumed that the random parameters at the different levels of the hierarchy are independently distributed.

Using the above assumptions, the covariance matrix of $\mathbf{y}_i$ can now be written as

$$\begin{aligned}
\mathbf{W}_i &= \text{Cov}(\mathbf{y}_i, \mathbf{y}_i') \\
&= \mathbf{X}_{1i}\text{Cov}(\boldsymbol{\beta}_{1i}, \boldsymbol{\beta}_{1i}')\mathbf{X}_{1i}' + \mathbf{X}_{2i}\text{Cov}(\boldsymbol{\beta}_{2i}, \boldsymbol{\beta}_{2i}')\mathbf{X}_{2i}' \\
&= \mathbf{I}_{n_i} \otimes \boldsymbol{\Omega}_1 + (\mathbf{j}_{n_i} \otimes \mathbf{I}_p)\boldsymbol{\Omega}_2(\mathbf{j}_{n_i} \otimes \mathbf{I}_p)' \\
&= \mathbf{I}_{n_i} \otimes \boldsymbol{\Omega}_1 + \mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes \boldsymbol{\Omega}_2. \qquad (3.4)
\end{aligned}$$

The matrix $\mathbf{W}_i$ is of the order $pn_i \times pn_i$ and unless there is an equal number of level-one units in all the level-two units, the $\mathbf{W}_i$ $(i = 1, 2, \ldots, M)$ will be of a different order.

Goldstein and McDonald (1988) also show that multilevel models may be applied where some or all of the random parameters depend on unobservable or latent variables. They define a 2-level common factor model by writing $\boldsymbol{\beta}_{1ij}$ and $\boldsymbol{\beta}_{2i}$ in (3.2) as

$$\boldsymbol{\beta}_{1ij} = \mathbf{B}\mathbf{z}_{ij} + \mathbf{e}_{ij}$$

and

$$\boldsymbol{\beta}_{2i} = \mathbf{A}\mathbf{w}_i + \mathbf{u}_i$$

yielding the model for the ij-th observation

$$\mathbf{y}_{ij} = \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{X}_{1ij}(\mathbf{B}\mathbf{z}_{ij} + \mathbf{e}_{ij}) + \mathbf{X}_{2i}(\mathbf{A}\mathbf{w}_i + \mathbf{u}_i).$$

Of course, different definitions of $\mathbf{X}_1$ and $\mathbf{X}_2$ in (3.2) will lead to different model specifications, for example, choosing $\mathbf{X}_1$ and $\mathbf{X}_2$ as in (3.3), the model becomes

$$\mathbf{y}_{ij} = \mathbf{X}_{0ij}\boldsymbol{\beta}_0 + \mathbf{B}\mathbf{z}_{ij} + \mathbf{e}_{ij} + \mathbf{A}\mathbf{w}_i + \mathbf{u}_i$$

which may be regarded as a generalisation of the model for simultaneous factor analysis in several groups.

Now that a general linear multilevel model has been introduced, we shall proceed with the introduction of possible ways to estimate the parameters in such a model. The first method that will be described is the marginal maximum likelihood method (the MML method) and the EM (expected maximisation) algorithm. The MML method provides the equations for use in an iterative algorithm, the EM algorithm, to obtain the MML estimates of the unknown parameters.

3.3 Parameter estimation using Marginal Maximum Likelihood

The use of this method in the estimation of parameters, has been successful and is described in the literature in a number of papers - see for example Dempster, Rubin and Tsutakawa (1981), Bock (1990) and Bock and Aitkin (1981).

This section is based on a theoretical description of the marginal maximum likelihood method found in Du Toit (1993). Consider a hierarchical structure with two levels. Assume a random sample of p-variate observations has been drawn with M level-two units and n_i level-one units within the i-th level-two unit. Let $\mathbf{y}_i$ be the $pn_i \times 1$ vector of observations obtained from the i-th level-two unit, and assume that each $\mathbf{y}_i$ $(i = 1, 2, \ldots, M)$ can be described by the general model in (3.3). Also let $\mathbf{b}_i$ be a vector containing all the random parameters in the model and let $g(\mathbf{b}_i)$ be the density function of $\mathbf{b}_i$. Finally, let $f(\mathbf{y}_i|\mathbf{b}_i)$ be the density function of $\mathbf{y}_i$ conditional on $\mathbf{b}_i$.

The density function of $\mathbf{y}_i$ is now written in terms of the joint probability density function (pdf) of $\mathbf{y}_i$ and $\mathbf{b}_i$ in the standard manner, namely

$$h(\mathbf{y}_i) = \int_{\mathbf{b}_i} h(\mathbf{y}_i, \mathbf{b}_i) d\mathbf{b}_i.$$

The joint density can, however, be expressed in terms of the conditional density if we use the identity

$$f(\mathbf{y}_i|\mathbf{b}_i) = \frac{h(\mathbf{y}_i, \mathbf{b}_i)}{g(\mathbf{b}_i)}.$$

The density of $\mathbf{y}_i$ can now be written in the form

$$h(\mathbf{y}_i) = \int_{\mathbf{b}_i} f(\mathbf{y}_i|\mathbf{b}_i) g(\mathbf{b}_i) d\mathbf{b}_i.$$

The likelihood function for the M level two units can be expressed as

$$L = \prod_{i=1}^{M} \int_{\mathbf{b}_i} f(\mathbf{y}_i|\mathbf{b}_i) g(\mathbf{b}_i) d\mathbf{b}_i$$

and the natural logarithm of the likelihood function as

$$\ell nL = \sum_{i=1}^{M} \ell n \int_{\mathbf{b}_i} f(\mathbf{y}_i|\mathbf{b}_i) g(\mathbf{b}_i) d\mathbf{b}_i. \tag{3.5}$$

The parameter estimators are obtained by setting the partial derivatives of ℓnL equal to zero. Expressions for these partial derivatives are provided in the following proposition:

Proposition 3.1

Let $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ be vectors containing the unknown parameters in $g(\mathbf{b}_i)$ and $f(\mathbf{y}_i|\mathbf{b}_i)$ respectively.

The partial derivative of ℓnL with respect to a typical element of $\boldsymbol{\theta}$, say θ_ℓ, is given by

$$\frac{\partial \ell nL}{\partial \theta_\ell} = \sum_{i=1}^{M} \mathrm{E}_c \left\{ \frac{\partial \ell n\ g(\mathbf{b}_i)}{\partial \theta_\ell} \right\} \tag{3.6}$$

and the partial derivative of ℓnL with respect to a typical element of $\boldsymbol{\tau}$, say τ_ℓ, is given by

$$\frac{\partial \ell nL}{\partial \tau_\ell} = \sum_{i=1}^{M} \mathrm{E}_c \left\{ \frac{\partial \ell n\ f(\mathbf{y}_i|\mathbf{b}_i)}{\partial \tau_\ell} \right\} \tag{3.7}$$

where E_c indicates the conditional expected value of $\mathbf{b}_i$ given $\mathbf{y}_i$.

Proof

From (3.5) it follows that

$$\frac{\partial \ell nL}{\partial \theta_\ell} = \sum_{i=1}^{M} \frac{\partial}{\partial \theta_\ell} \ell n \int_{\mathbf{b}_i} f(\mathbf{y}_i|\mathbf{b}_i) g(\mathbf{b}_i) d\mathbf{b}_i \tag{3.8}$$

and

$$\frac{\partial \ell nL}{\partial \tau_\ell} = \sum_{i=1}^{M} \frac{\partial}{\partial \tau_\ell} \ell n \int_{\mathbf{b}_i} f(\mathbf{y}_i|\mathbf{b}_i) g(\mathbf{b}_i) d\mathbf{b}_i. \tag{3.9}$$

The derivative of the natural logarithm of a function has a simple form, namely

$$\frac{\partial \ell n h(x)}{\partial x} = \frac{1}{h(x)} \frac{\partial h(x)}{\partial x}. \tag{3.10}$$

This result is now substituted into (3.8) and (3.9) which are then rewritten as

$$\frac{\partial \ell nL}{\partial \theta_\ell} = \sum_{i=1}^{M} \frac{1}{h(\mathbf{y}_i)} \int_{\mathbf{b}_i} f(\mathbf{y}_i|\mathbf{b}_i) \frac{\partial g(\mathbf{b}_i)}{\partial \theta_\ell} d\mathbf{b}_i \tag{3.11}$$

and

$$\frac{\partial \ell nL}{\partial \tau_\ell} = \sum_{i=1}^{M} \frac{1}{h(\mathbf{y}_i)} \int_{\mathbf{b}_i} \frac{\partial f(\mathbf{y}_i|\mathbf{b}_i)}{\partial \tau_\ell} g(\mathbf{b}_i) d\mathbf{b}_i. \tag{3.12}$$

If we rearrange (3.10) to form the expression

$$\frac{\partial h(x)}{\partial x} = h(x) \frac{\partial \ell n h(x)}{\partial x}$$

and we use this in (3.11) and (3.12), we obtain the following expressions for the derivatives of ℓnL, namely

$$\frac{\partial \ell nL}{\partial \theta_\ell} = \sum_{i=1}^{M} \frac{1}{h(\mathbf{y}_i)} \int_{\mathbf{b}_i} \frac{\partial \ell n\ g(\mathbf{b}_i)}{\partial \theta_\ell} g(\mathbf{b}_i) f(\mathbf{y}_i|\mathbf{b}_i) d\mathbf{b}_i \tag{3.13}$$

and

$$\frac{\partial \ell nL}{\partial \tau_\ell} = \sum_{i=1}^{M} \frac{1}{h(\mathbf{y}_i)} \int_{\mathbf{b}_i} \frac{\partial \ell n\ f(\mathbf{y}_i|\mathbf{b}_i)}{\partial \tau_\ell} f(\mathbf{y}_i|\mathbf{b}_i) g(\mathbf{b}_i) d\mathbf{b}_i. \tag{3.14}$$

By simply rearranging the terms in (3.13) and (3.14), these two expressions may be written as

$$\frac{\partial \ell nL}{\partial \theta_\ell} = \sum_{i=1}^{M} \int_{\mathbf{b}_i} \frac{\partial \ell n\ g(\mathbf{b}_i)}{\partial \theta_\ell} . \frac{g(\mathbf{b}_i) f(\mathbf{y}_i|\mathbf{b}_i)}{h(\mathbf{y}_i)} d\mathbf{b}_i \tag{3.15}$$

and

$$\frac{\partial \ell nL}{\partial \tau_\ell} = \sum_{i=1}^{M} \int_{\mathbf{b}_i} \frac{\partial \ell n\ f(\mathbf{y}_i|\mathbf{b}_i)}{\partial \tau_\ell} . \frac{g(\mathbf{b}_i) f(\mathbf{y}_i|\mathbf{b}_i)}{h(\mathbf{y}_i)} d\mathbf{b}_i. \tag{3.16}$$

These two equations contain identical second terms that may be simplified as

$$\frac{g(\mathbf{b}_i) f(\mathbf{y}_i|\mathbf{b}_i)}{h(\mathbf{y}_i)} = \frac{h(\mathbf{y}_i, \mathbf{b}_i)}{h(\mathbf{y}_i)} = p(\mathbf{b}_i|\mathbf{y}_i) \tag{3.17}$$

which is the conditional density of $\mathbf{b}_i$ given $\mathbf{y}_i$, also referred to as the *posterior* pdf of $\mathbf{b}_i$.

As a final step in obtaining the gradient of ℓnL with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$, we substitute (3.17) into (3.15) and (3.16), which leads to

$$\begin{aligned}\frac{\partial \ell nL}{\partial \theta_\ell} &= \sum_{i=1}^{M} \int_{\mathbf{b}_i} \frac{\partial \ell n\ g(\mathbf{b}_i)}{\partial \theta_\ell} p(\mathbf{b}_i|\mathbf{y}_i) d\mathbf{b}_i \\ &= \sum_{i=1}^{M} \mathrm{E}_c \left\{ \frac{\partial \ell n\ g(\mathbf{b}_i)}{\partial \theta_\ell} \right\}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial \ell nL}{\partial \tau_\ell} &= \sum_{i=1}^{M} \int_{\mathbf{b}_i} \frac{\partial \ell n\ f(\mathbf{y}_i|\mathbf{b}_i)}{\partial \tau_\ell} p(\mathbf{b}_i|\mathbf{y}_i) d\mathbf{b}_i \\ &= \sum_{i=1}^{M} \mathrm{E}_c \left\{ \frac{\partial \ell n\ f(\mathbf{y}_i|\mathbf{b}_i)}{\partial \tau_\ell} \right\}\end{aligned}$$

and consequently the proposition is proved. □

To obtain the MML estimators of the parameters in $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$, expressions (3.6) and (3.7) are set equal to zero, that is

$$\sum_{i=1}^{M} \mathrm{E}_c \left\{ \frac{\partial \ell n\ g(\mathbf{b}_i)}{\partial \theta_\ell} \right\} = 0 \tag{3.18}$$

and

$$\sum_{i=1}^{M} \mathrm{E}_c \left\{ \frac{\partial \ell n\ f(\mathbf{y}_i|\mathbf{b}_i)}{\partial \tau_\ell} \right\} = 0 \tag{3.19}$$

and the estimators $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\tau}}$ are solved from these equations.

When (3.18) and (3.19) do not provide expressions for $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\tau}}$ in closed form, one has to use an iterative procedure to obtain these parameter estimates. A procedure that proved to work well in the current situation is the expected maximisation algorithm, or in short, the EM algorithm. A description of it will now be given.

In (3.18) and (3.19) the conditional expected value has to be determined for functions that contain the fixed parameters $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ as well as the random parameters $\mathbf{b}_i$ and the observed vector $\mathbf{y}_i$. Assuming now that $\mathbf{b}_i$ and $\mathbf{y}_i$ are normally distributed, (3.18) and (3.19) will contain $\mathbf{b}_i$ in the form of moments of $p(\mathbf{b}_i|\mathbf{y}_i)$ which are only of first and second order - i.e. $\mathbf{b}_i$ will appear in (3.18) and (3.19) only in the form $\mathrm{E}_c(\mathbf{b}_i)$ and $\mathrm{Cov}_c(\mathbf{b}_i, \mathbf{b}_i')$ where Cov_c is the conditional covariance matrix of $\mathbf{b}_i$ given $\mathbf{y}_i$. Expressions for $\mathrm{E}_c(\mathbf{b}_i)$ and $\mathrm{Cov}_c(\mathbf{b}_i, \mathbf{b}_i')$ may be obtained from the joint distribution of $\mathbf{y}_i$ and $\mathbf{b}_i$ (see e.g. Morrison (1990)). Since $\mathbf{y}_i$ and $\mathbf{b}_i$ are both assumed to be normally distributed, it is apparent that they jointly also follow a normal distribution which is given by

$$\begin{pmatrix} \mathbf{y}_i \\ \mathbf{b}_i \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} \mathrm{E}(\mathbf{y}_i) \\ \mathrm{E}(\mathbf{b}_i) \end{pmatrix}, \begin{pmatrix} \mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_i') & \mathrm{Cov}(\mathbf{y}_i, \mathbf{b}_i') \\ \mathrm{Cov}(\mathbf{b}_i, \mathbf{y}_i') & \mathrm{Cov}(\mathbf{b}_i, \mathbf{b}_i') \end{pmatrix} \right\}.$$

The moments of the conditional distribution of $\mathbf{b}_i$ given $\mathbf{y}_i$ may now be obtained by the expressions

$$\mathrm{E}_c(\mathbf{b}_i) = \mathrm{E}(\mathbf{b}_i) + \mathrm{Cov}(\mathbf{b}_i, \mathbf{y}_i') \left[\mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_i')\right]^{-1} (\mathbf{y}_i - \mathrm{E}(\mathbf{y}_i)) \tag{3.20}$$

and

$$\mathrm{Cov}_c(\mathbf{b}_i, \mathbf{b}_i') = \mathrm{Cov}(\mathbf{b}_i, \mathbf{b}_i') - \mathrm{Cov}(\mathbf{b}_i, \mathbf{y}_i') \left[\mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_i')\right]^{-1} \mathrm{Cov}(\mathbf{y}_i, \mathbf{b}_i'). \tag{3.21}$$

It is evident that (3.20) and (3.21) are functions of the data vector $\mathbf{y}_i$ and of the unknown *fixed* parameters in the model.

One may now proceed in the following iterative manner to obtain close approximations of the unknown parameters in the model. To start the algorithm, the vectors $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ need to be assigned some arbitrary initial values. The closer these values are to the final solution, the faster will the algorithm converge. However, if no prior knowledge is available as to the estimators $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\tau}}$, one may start for simplicity's sake with zeros and ones as initial values (e.g. ones for variances).

When these initial values have been assigned to $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\tau}}$, equations (3.20) and (3.21) are evaluated to obtain initial estimates of $\mathrm{E}_c(\mathbf{b}_i)$ and $\mathrm{Cov}_c(\mathbf{b}_i, \mathbf{b}_i')$. This is called the expected step (E-step). The estimates are now substituted into (3.18) and (3.19) to obtain a new set of values for $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\tau}}$, and this is called the maximisation step (M-step). That is the end of the first EM iteration. These new values of the parameter estimates are now again substituted into (3.20) and (3.21) in the E-step to obtain a next approximation of $\mathrm{E}_c(\mathbf{b}_i)$ and $\mathrm{Cov}_c(\mathbf{b}_i, \mathbf{b}_i')$. To complete the second iteration, the values are used in the M-step to obtain the next $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\tau}}$ from (3.18) and (3.19). This procedure is now repeated until convergence of the parameter estimates $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\tau}}$ is reached.

The EM algorithm described above can easily be applied in practice to obtain close approximations to the marginal maximum likelihood estimators of the parameters, although the method lacks the utilization of second order derivatives for determining the standard errors of the estimators. A method that does provide the standard errors of the estimators will be described next.

3.4 Parameter estimation using the Fisher scoring method

The Fisher scoring method is an iterative method of estimating parameters in a model and may be used to obtain the parameter estimates that optimise different discrepancy functions. A description of this method may be found in Browne and Du Toit (1992). The theory presented in this section is solely based on this reference.

Examples of discrepancy functions that are typically optimised using this method are generalised least squares and maximum likelihood discrepancy functions. The discussion that follows will show the use of the Fisher scoring method to obtain the estimate of the parameter vector that maximises the likelihood function under normality assumptions.

Consider an identical two-level structure such as the one in the previous section where the MML method was discussed. Let $\mathbf{y}$ be the $Np \times 1$ vector of observations (where $N = \sum_{i=1}^{M} n_i$) which is partitioned into M subvectors representing the level-two units, i.e.

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_M \end{pmatrix}.$$

Suppose that these M subvectors, $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M$, represent the experimental units where each $\mathbf{y}_i$ is assumed to have a normal distribution with an expected value and covariance matrix denoted respectively by

$$\mathrm{E}(\mathbf{y}_i) = \boldsymbol{\xi}_i$$

and

$$\mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_i') = \boldsymbol{\Sigma}_i.$$

Note that both these quantities are functions of the parameter vector $\boldsymbol{\gamma}$ and that they may be written as $\boldsymbol{\xi}_i = \boldsymbol{\xi}_i(\boldsymbol{\gamma})$ and $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_i(\boldsymbol{\gamma})$. It will be assumed that both these functions are twice continuously differentiable with respect to the elements of $\boldsymbol{\gamma}$. In the present case where maximum likelihood estimators will be obtained, the density function for the i-th experimental unit $\mathbf{y}_i$ is

$$f(\mathbf{y}_i) = (2\pi)^{-\frac{pn_i}{2}} |\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} \exp\{-\frac{1}{2} \mathrm{tr} \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\xi}_i)(\mathbf{y}_i - \boldsymbol{\xi}_i)'\}. \tag{3.22}$$

The likelihood function for the sample of M experimental units which is defined as the product of the M density functions, follows as

$$L = \prod_{i=1}^{M} f(\mathbf{y}_i).$$

Maximum likelihood estimators of the parameters are obtained by maximising L with respect to the parameter vector $\boldsymbol{\gamma}$. However, since the natural logarithmic function is a monotonic increasing function, maximising ℓnL will yield the same parameter estimators.

Using (3.22) and the definition of L, it follows that

$$\ell nL = -\frac{1}{2} \sum_{i=1}^{M} \{pn_i \ell n(2\pi) + \ell n|\boldsymbol{\Sigma}_i| + \mathrm{tr}[\boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\xi}_i)(\mathbf{y}_i - \boldsymbol{\xi}_i)']\}. \tag{3.23}$$

Omitting the constant term in the above expression and changing the sign results in a function whose *minimum* will yield the maximum likelihood estimators. This function is called the discrepancy function and is denoted by F where

$$F(\boldsymbol{\gamma}) = \frac{1}{2} \sum_{i=1}^{M} \{\ell n|\boldsymbol{\Sigma}_i| + \mathrm{tr}[\boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\xi}_i)(\mathbf{y}_i - \boldsymbol{\xi}_i)']\} \tag{3.24}$$

is a function of the parameter vector $\boldsymbol{\gamma}$.

The Fisher scoring method makes use of first order and second order derivative information in the form of the gradient and expected Hessian of the discrepancy function in the minimisation process.

Let the gradient vector of $F(\gamma)$ be denoted by $\mathbf{g}$ where

$$\mathbf{g}(\gamma) = \frac{\partial F(\gamma)}{\partial \gamma}. \tag{3.25}$$

The elements of the gradient vector are given by the expression (Du Toit, 1993)

$$[\mathbf{g}(\gamma)]_k = -\sum_{i=1}^{M} \left\{ \text{tr}\left[\mathbf{Q}_i \frac{\partial \boldsymbol{\xi}_i}{\partial \gamma_k}\right] + \frac{1}{2}\text{tr}\left[\mathbf{P}_i \frac{\partial \boldsymbol{\Sigma}_i}{\partial \gamma_k}\right] \right\} \tag{3.26}$$

where

$$\mathbf{Q}_i = (\mathbf{y}_i - \boldsymbol{\xi}_i)'\boldsymbol{\Sigma}_i^{-1}$$

and

$$\mathbf{P}_i = \boldsymbol{\Sigma}_i^{-1}((\mathbf{y}_i - \boldsymbol{\xi}_i)(\mathbf{y}_i - \boldsymbol{\xi}_i)' - \boldsymbol{\Sigma}_i)\boldsymbol{\Sigma}_i^{-1}.$$

Instead of the Hessian matrix of $F(\gamma)$, which is defined as the matrix of second order derivatives of $F(\gamma)$ with respect to the parameter vector, the expected Hessian is used as an approximation since the Hessian matrix may be difficult to obtain (Lee and Jennrich, 1979). Let the expected Hessian matrix be denoted by $\mathbf{H}$ where

$$\begin{aligned} \mathbf{H}(\gamma) &= \text{E}\left(\frac{\partial^2 F(\gamma)}{\partial \gamma \partial \gamma'}\right) \\ &\approx \frac{\partial^2 F(\gamma)}{\partial \gamma \partial \gamma'}. \end{aligned}$$

The elements of $\mathbf{H}(\gamma)$ are given by the expression (Du Toit, 1993)

$$[\mathbf{H}(\boldsymbol{\gamma})]_{k\ell} = \sum_{i=1}^{M} \left\{ \mathrm{tr}\left[\frac{\partial \boldsymbol{\xi}_i}{\partial \gamma_k} \boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\xi}_i}{\partial \gamma_\ell} \right] + \frac{1}{2} \mathrm{tr}\left[\boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i}{\partial \gamma_k} \boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i}{\partial \gamma_\ell} \right] \right\}. \tag{3.27}$$

In the iterative procedure, (3.26) and (3.27) are used to obtain an increment vector that is added to the current estimate of $\boldsymbol{\gamma}$, to obtain a new estimate that will be closer to the minimiser of the discrepancy function. If $\hat{\boldsymbol{\gamma}}_t$ is the estimate of $\boldsymbol{\gamma}$ obtained at the t-th iteration, the quantities $\mathbf{g}_t = \mathbf{g}(\hat{\boldsymbol{\gamma}}_t)$ and $\mathbf{H}_t = \mathbf{H}(\hat{\boldsymbol{\gamma}}_t)$ are evaluated and the increment vector is obtained by

$$\boldsymbol{\delta}_t = -\mathbf{H}_t^{-1}\mathbf{g}_t. \tag{3.28}$$

The next approximation, $\hat{\boldsymbol{\gamma}}_{t+1}$, is now obtained by the expression

$$\hat{\boldsymbol{\gamma}}_{t+1} = \hat{\boldsymbol{\gamma}}_t + \alpha_t \boldsymbol{\delta}_t \tag{3.29}$$

where α_t is a parameter that is chosen to ensure that $F(\hat{\boldsymbol{\gamma}}_{t+1}) \leq F(\hat{\boldsymbol{\gamma}}_t)$ and where $0 < \alpha_t \leq 1$.

An advantage of this method is that at convergence, the minimiser $\hat{\boldsymbol{\gamma}}$ of the discrepancy function is obtained as well as an estimate of the covariance matrix of these parameter estimators. This covariance matrix is given by $N^{-1}\mathbf{H}^{-1}$ where $\mathbf{H}^{-1}$ replaces $\mathbf{H}_t^{-1}$ at the point of convergence (Browne and Du Toit, 1992).

Two other methods of obtaining the minimiser $\hat{\boldsymbol{\gamma}}$ of the discrepancy function which also provide the estimated covariance matrix of the parameter estimators are the Newton-Raphson and Gauss-Newton methods. They differ only slightly from the Fisher scoring method. The Newton-Raphson method uses, instead of the approximate Hessian as in the Fisher scoring method, the complete Hessian of the discrepancy function. It will therefore be more time consuming when applied in practice, since more computations

will be required. In the Gauss-Newton method the weight matrix is not respecified at each iteration. An estimate of the weight matrix is obtained and used, without changing its value, in the iteration process.

3.5 Contraint estimation in the MML method

Nowhere in the literature has the MML method been used to estimate parameters in a model when equality constraints are imposed on some of the parameters. In this section it is shown how the MML method described in Section 3.3 may be adjusted to obtain the parameter estimators under such circumstances.

Consider a two-level hierarchical structure and a model with some parameters to be estimated subject to equality constraints. Let the $q \times 1$ parameter vector $\boldsymbol{\gamma}$ be partitioned into two subvectors, as was similarly done in the section where the unconstrained parameter estimation was discussed. In the present situation, however, let $\boldsymbol{\theta}^\star$ denote the $t_1 \times 1$ parameter subvector that contains the parameters in $g(\mathbf{b}_i)$, and let $\boldsymbol{\tau}^\star$ denote the $t_2 \times 1$ parameter subvector that contains the parameters in $f(\mathbf{y}_i|\mathbf{b}_i)$. In terms of the constraints imposed on the parameters, consideration will here be given only to the case where r_1 equality constraints are imposed on the parameters in $\boldsymbol{\theta}$ $(q_1 \times 1)$, where $\boldsymbol{\theta} \subseteq \boldsymbol{\theta}^\star$ and r_2 equality constraints are imposed on the parameters in $\boldsymbol{\tau}$ $(q_2 \times 1)$, where $\boldsymbol{\tau} \subseteq \boldsymbol{\tau}^\star$. Constraint functions containing parameters from both $\boldsymbol{\theta}^\star$ and $\boldsymbol{\tau}^\star$ will not be considered here.

Write the two sets of constraints as

$$\mathbf{c}_\theta(\boldsymbol{\theta}) = \mathbf{0} \tag{3.30}$$

and

$$\mathbf{c}_\tau(\boldsymbol{\tau}) = \mathbf{0} \tag{3.31}$$

where $\mathbf{c}_\theta$ and $\mathbf{c}_\tau$ are respectively $r_1 \times 1$ and $r_2 \times 1$ vector valued constraint functions.

Assume that these functions are continuously differentiable with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ respectively.

In many practical applications it may happen that the constraint functions are not linear. Consequently complicated derivations of formulae and computations in practice will have to be performed. This, however, is unnecessary since it is possible to derive linear constraint functions that approximate (3.30) and (3.31). Such functions are obtained by expressing (3.30) and (3.31) as first order Taylor series.

In order to show how this linearisation is accomplished, let $\mathbf{L}_\theta$ be the $r_1 \times q_1$ Jacobian matrix of $\mathbf{c}_\theta$. Consequently it follows that

$$\mathbf{L}_\theta(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}'}\mathbf{c}_\theta(\boldsymbol{\theta}).$$

Let $\boldsymbol{\theta}_0$ be a $q_1 \times 1$ arbitrarily known vector. The constraint function $\mathbf{c}_\theta$ can now be approximated by the linear function

$$\mathbf{c}_\theta(\boldsymbol{\theta}) \approx \mathbf{c}_{\theta_0} + \mathbf{L}_{\theta_0}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \tag{3.32}$$

where $\mathbf{c}_{\theta_0} = \mathbf{c}_\theta(\boldsymbol{\theta}_0)$ and $\mathbf{L}_{\theta_0} = \mathbf{L}_\theta(\boldsymbol{\theta}_0)$.

Substitution of (3.32) into (3.30) will show that the linear approximation to the constraints can be written as

$$\mathbf{L}_{\theta_0}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = -\mathbf{c}_{\theta_0}. \tag{3.33}$$

Exactly analogous to the linearisation of the constraints $\mathbf{c}_\theta = \mathbf{0}$, it can be shown that the constraints $\mathbf{c}_\tau = \mathbf{0}$ may be written as the linear approximation

$$\mathbf{L}_{\tau_0}(\boldsymbol{\tau} - \boldsymbol{\tau}_0) = -\mathbf{c}_{\tau_0} \tag{3.34}$$

where $\boldsymbol{\tau}_0$ is a $q_2 \times 1$ arbitrarily known vector, $\mathbf{L}_{\tau_0} = \mathbf{L}_\tau(\boldsymbol{\tau}_0)$ is the $r_2 \times q_2$ Jacobian matrix of $\mathbf{c}_\tau$ at the point $\boldsymbol{\tau} = \boldsymbol{\tau}_0$ and $\mathbf{c}_{\tau_0} = \mathbf{c}_\tau(\boldsymbol{\tau}_0)$.

Now that the constraint functions have been rewritten and expressed in forms that are generally easier to work with, namely as linear functions, it will be shown how the parameter estimators $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\tau}}$ are obtained which maximise the log-likelihood function while simultaneously satisfying these linear constraints.

The method of Lagrange multipliers is used to add the constraints to the log-likelihood function to form a new function for maximisation, say $L^\star$, where

$$L^\star = \ell nL + \boldsymbol{\lambda}'_\theta[\mathbf{c}_{\theta_0} + \mathbf{L}_{\theta_0}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)] + \boldsymbol{\lambda}'_\tau[\mathbf{c}_{\tau_0} + \mathbf{L}_{\tau_0}(\boldsymbol{\tau} - \boldsymbol{\tau}_0)] \tag{3.35}$$

and $\boldsymbol{\lambda}_\theta$ $(r_1 \times 1)$ and $\boldsymbol{\lambda}_\tau$ $(r_2 \times 1)$ are vectors of Lagrange multipliers.

To obtain $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\tau}}$ which maximise $L^\star$, the gradient of $L^\star$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ as well as to $\boldsymbol{\lambda}_\theta$ and $\boldsymbol{\lambda}_\tau$ will be obtained. These gradients will then be set equal to zero and solved for $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\tau}}$, $\hat{\boldsymbol{\lambda}}_\theta$ and $\hat{\boldsymbol{\lambda}}_\tau$.

The gradient of $L^\star$ with respect to the elements of $\boldsymbol{\theta}$ is given by the expression

$$\begin{aligned} \frac{\partial L^\star}{\partial \boldsymbol{\theta}'} &= \frac{\partial \ell nL}{\partial \boldsymbol{\theta}'} + \frac{\partial}{\partial \boldsymbol{\theta}'}\boldsymbol{\lambda}'_\theta \mathbf{L}_{\theta_0}\boldsymbol{\theta} \\ &= \frac{\partial \ell nL}{\partial \boldsymbol{\theta}'} + \boldsymbol{\lambda}'_\theta \mathbf{L}_{\theta_0} \end{aligned} \tag{3.36}$$

since

$$\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\theta}'} = \mathbf{I}.$$

Similarly it follows that the gradient of $L^\star$ with respect to the elements of $\boldsymbol{\tau}$ is obtained from the expression

$$\frac{\partial L^\star}{\partial \boldsymbol{\tau}'} = \frac{\partial \ell n L}{\partial \boldsymbol{\tau}'} + \boldsymbol{\lambda}_\tau' \mathbf{L}_{\tau_0}. \tag{3.37}$$

The first terms in (3.36) and (3.37) were evaluated earlier (see Proposition 3.1). Substitution of (3.6) and (3.7) into (3.36) and (3.37) respectively, and setting the latter two expressions equal to zero, leads to

$$\sum_{i=1}^{M} \mathrm{E}_c \left\{ \frac{\partial \ell n\, g(\mathbf{b}_i)}{\partial \boldsymbol{\theta}'} \right\} + \hat{\boldsymbol{\lambda}}_\theta' \mathbf{L}_{\theta_0} = \mathbf{0}' \tag{3.38}$$

and

$$\sum_{i=1}^{M} \mathrm{E}_c \left\{ \frac{\partial \ell n\, f(\mathbf{y}_i|\mathbf{b}_i)}{\partial \boldsymbol{\tau}'} \right\} + \hat{\boldsymbol{\lambda}}_\tau' \mathbf{L}_{\tau_0} = \mathbf{0}'. \tag{3.39}$$

Now that the gradient of $L^\star$ has been obtained with respect to the two parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ and set equal to zero, the gradient of $L^\star$ will now be obtained with respect to the two vectors of Lagrange multipliers $\boldsymbol{\lambda}_\theta$ and $\boldsymbol{\lambda}_\tau$. These two vectors appear only in the second and third terms of equation (3.35) and therefore the derivative of $L^\star$ with respect to $\boldsymbol{\lambda}_\theta$ and $\boldsymbol{\lambda}_\tau$ follow respectively as

$$\frac{\partial L^\star}{\partial \boldsymbol{\lambda}_\theta} = \mathbf{c}_{\theta_0} + \mathbf{L}_{\theta_0}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

and

$$\frac{\partial L^{\star}}{\partial \boldsymbol{\lambda}_{\tau}} = \mathbf{c}_{\tau_0} + \mathbf{L}_{\tau_0}(\boldsymbol{\tau} - \boldsymbol{\tau}_0).$$

These expressions, if set equal to zero, give

$$\mathbf{c}_{\theta_0} + \mathbf{L}_{\theta_0}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbf{0} \tag{3.40}$$

and

$$\mathbf{c}_{\tau_0} + \mathbf{L}_{\tau_0}(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_0) = \mathbf{0} \tag{3.41}$$

which give the two sets of linear constraints that were derived earlier in (3.33) and (3.34).

To obtain the desired solution $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\tau}}$ of $\boldsymbol{\tau}$, i.e. the parameter estimators that satisfy the constraints as well as maximise $L^{\star}$, it will now be necessary to solve (3.38) and (3.40) simultaneously for $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\lambda}}_{\theta}$, as well as to solve (3.39) and (3.41) simultaneously for $\hat{\boldsymbol{\tau}}$ and $\hat{\boldsymbol{\lambda}}_{\tau}$.

The following propositions will show that if (3.38) and (3.39) can be expressed in a certain form, expressions for $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\tau}}$, $\hat{\boldsymbol{\lambda}}_{\theta}$ and $\hat{\boldsymbol{\lambda}}_{\tau}$ can easily be obtained.

Proposition 3.2

Let $\mathbf{Y}$ be a $q_1 \times q_1$ non-singular matrix and $\mathbf{z}$ a $q_1 \times 1$ vector. If (3.38) is expressed in the form

$$\mathbf{Y}\hat{\boldsymbol{\theta}} + \mathbf{L}'_{\theta_0}\hat{\boldsymbol{\lambda}}_{\theta} = \mathbf{z}$$

the solutions for $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\lambda}}_\theta$ are obtained from the expressions

$$\hat{\boldsymbol{\theta}} = \mathbf{Y}^{-1}(\mathbf{z} - \mathbf{L}'_{\theta_0}\hat{\boldsymbol{\lambda}}_\theta) \tag{3.42}$$

and

$$\hat{\boldsymbol{\lambda}}_\theta = \mathbf{S}\mathbf{w} \tag{3.43}$$

where

$$\mathbf{S} = (\mathbf{L}_{\theta_0}\mathbf{Y}^{-1}\mathbf{L}'_{\theta_0})^{-1},$$

$$\mathbf{w} = \mathbf{L}_{\theta_0}\mathbf{Y}^{-1}\mathbf{z} - \mathbf{x},$$

$$\mathbf{x} = \mathbf{L}_{\theta_0}\boldsymbol{\theta}_0 - \mathbf{c}_{\theta_0}.$$

Proof

Suppose there exist a $q_1 \times q_1$ non-singular matrix $\mathbf{Y}$ and a $q_1 \times 1$ vector $\mathbf{z}$ such that (3.38) can be rewritten in the form

$$\mathbf{Y}\hat{\boldsymbol{\theta}} + \mathbf{L}'_{\theta_0}\hat{\boldsymbol{\lambda}}_\theta = \mathbf{z}. \tag{3.44}$$

Rewrite (3.40) as

$$\mathbf{L}_{\theta_0}\hat{\boldsymbol{\theta}} = \mathbf{x} \tag{3.45}$$

where

$$\mathbf{x} = \mathbf{L}_{\theta_0}\boldsymbol{\theta}_0 - \mathbf{c}_{\theta_0}$$

is an $r_1 \times 1$ known vector.

To obtain a simultaneous solution for $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\lambda}}_\theta$ from (3.44) and (3.45), the latter two expressions are combined to yield the single expression

$$\begin{pmatrix} \mathbf{Y} & \mathbf{L}'_{\theta_0} \\ \mathbf{L}_{\theta_0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\lambda}}_\theta \end{pmatrix} = \begin{pmatrix} \mathbf{z} \\ \mathbf{x} \end{pmatrix}. \tag{3.46}$$

Straightforward matrix algebra may now be applied to the above expression, which yields the solution

$$\begin{pmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\lambda}}_\theta \end{pmatrix} = \begin{pmatrix} \mathbf{Y} & \mathbf{L}'_{\theta_0} \\ \mathbf{L}_{\theta_0} & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{z} \\ \mathbf{x} \end{pmatrix}. \tag{3.47}$$

However, since the inversion of a partitioned matrix is required, it is possible to write this solution in a more simplified form - particularly helpful in practical applications, since the matrix to be inverted may be extremely large. This will cause the estimation procedure to be extremely time consuming, especially when the estimates are iteratively obtained.

Standard results on partitioned matrices, which may be found in Morrison (1990), are used to write

$$\begin{pmatrix} \mathbf{Y} & \mathbf{L}'_{\theta_0} \\ \mathbf{L}_{\theta_0} & \mathbf{0} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{Y}^{-1} - \mathbf{Y}^{-1}\mathbf{L}'_{\theta_0}\mathbf{S}\mathbf{L}_{\theta_0}\mathbf{Y}^{-1} & \mathbf{Y}^{-1}\mathbf{L}'_{\theta_0}\mathbf{S} \\ \mathbf{S}\mathbf{L}_{\theta_0}\mathbf{Y}^{-1} & -\mathbf{S} \end{pmatrix}$$

where

$$\mathbf{S} = (\mathbf{L}_{\theta_0}\mathbf{Y}^{-1}\mathbf{L}'_{\theta_0})^{-1}.$$

Substitution of this result into (3.47) leads to

$$\begin{pmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\lambda}}_\theta \end{pmatrix} = \begin{pmatrix} \mathbf{Y}^{-1}\mathbf{z} - \mathbf{Y}^{-1}\mathbf{L}'_{\theta_0}\mathbf{S}(\mathbf{L}_{\theta_0}\mathbf{Y}^{-1}\mathbf{z} - \mathbf{x}) \\ \mathbf{S}(\mathbf{L}_{\theta_0}\mathbf{Y}^{-1}\mathbf{z} - \mathbf{x}) \end{pmatrix}$$
$$= \begin{pmatrix} \mathbf{Y}^{-1}(\mathbf{z} - \mathbf{L}'_{\theta_0}\mathbf{S}\mathbf{w}) \\ \mathbf{S}\mathbf{w} \end{pmatrix} \tag{3.48}$$

where

$$\mathbf{w} = \mathbf{L}_{\theta_0}\mathbf{Y}^{-1}\mathbf{z} - \mathbf{x}.$$

Expressions (3.42) and (3.43) now follow directly from expression (3.48), which proves the proposition. □

Proposition 3.3

Let $\tilde{\mathbf{Y}}$ be a $q_2 \times q_2$ non-singular matrix and let $\tilde{\mathbf{z}}$ be a $q_2 \times 1$ vector. Then, if (3.39) is expressed in the form

$$\tilde{\mathbf{Y}}\hat{\boldsymbol{\tau}} + \mathbf{L}'_{\tau_0}\hat{\boldsymbol{\lambda}}_\tau = \tilde{\mathbf{z}},$$

$\hat{\boldsymbol{\tau}}$ and $\hat{\boldsymbol{\lambda}}_\tau$ may be obtained from

$$\hat{\boldsymbol{\tau}} = \tilde{\mathbf{Y}}^{-1}(\tilde{\mathbf{z}} - \mathbf{L}'_{\tau_0}\hat{\boldsymbol{\lambda}}_\tau) \tag{3.49}$$

and

$$\hat{\boldsymbol{\lambda}}_\tau = \tilde{\mathbf{S}}\tilde{\mathbf{w}} \tag{3.50}$$

where

$$\tilde{\mathbf{S}} = (\mathbf{L}_{\tau_0}\tilde{\mathbf{Y}}^{-1}\mathbf{L}'_{\tau_0})^{-1},$$

$$\tilde{\mathbf{w}} = \mathbf{L}_{\tau_0}\tilde{\mathbf{Y}}^{-1}\tilde{\mathbf{z}} - \tilde{\mathbf{x}},$$

$$\tilde{\mathbf{x}} = \mathbf{L}_{\tau_0}\boldsymbol{\tau}_0 - \mathbf{c}_{\tau_0}.$$

Proof

The proof of this proposition will be omitted since it is exactly analogous to the proof of Proposition 3.2.

It is possible that $\mathbf{Y}$ and $\mathbf{z}$ in Proposition 3.2 and $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{z}}$ in Proposition 3.3 may contain elements of the parameter vectors $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\tau}}$ respectively. If so, the expressions for the parameter estimators are not in closed form, and these estimates should be obtained by means of an iterative algorithm. The algorithm used to obtain the parameter estimates in the unconstrained case, namely the EM algorithm, will be adjusted and used also when constraints are imposed.

Generally the EM algorithm, adjusted for estimation subject to constraints, will proceed with the following basic steps: after each EM iteration, the values obtained in the M-step for $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\tau}}$ are used as starting values for a separate iterative process aimed at obtaining values for $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\tau}}$ that satisfy the constraints; when the iterations for the constraints are completed, the values obtained for $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\tau}}$ are used in the E-step of the next EM iteration.

These basic steps will now be used to write down in more detail the steps which should be followed in the adjusted EM algorithm. Assigning initial values to the parameter vectors $\hat{\boldsymbol{\theta}}^{\star}$ and $\hat{\boldsymbol{\tau}}^{\star}$ is the first step in the EM algorithm. The E-step in the first EM iteration follows now by substituting $\hat{\boldsymbol{\theta}}^{\star}$ and $\hat{\boldsymbol{\tau}}^{\star}$ into (3.20) and (3.21) in order to obtain initial estimates for $E_c(\mathbf{b}_i)$ and $\mathrm{Cov}_c(\mathbf{b}_i, \mathbf{b}_i')$. The estimates of these moments are now used in the M-step to obtain first approximations to $\boldsymbol{\theta}^{\star}$ and $\boldsymbol{\tau}^{\star}$. The equations for this step are (3.18) and (3.19). For the parameters on which no constraints are imposed, the first EM iteration is now completed. For the constrained parameters ($\boldsymbol{\theta}$ and $\boldsymbol{\tau}$), however, the approximate values obtained in the M-step are used as starting values, and $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\tau}}$ are

repeatedly calculated using equations (3.42) and (3.49) until the constraints are smaller than some prescribed value. Now the first approximations to the full parameter vectors $\boldsymbol{\theta}^{\star}$ and $\boldsymbol{\tau}^{\star}$ have been calculated, and this completes the first EM iteration. The values obtained for $\hat{\boldsymbol{\theta}}^{\star}$ and $\hat{\boldsymbol{\tau}}^{\star}$ are now used as starting values for the second EM iteration. This process is continued until convergence is met by both the parameter estimates and constraints, or until a desired number of iterations has been completed.

3.6 Constraint estimation in the Fisher scoring method

The Fisher scoring method described in Section 3.4 to obtain maximum likelihood estimators of the parameters may also be adjusted to estimate the parameters when equality constraints are imposed on some of them. In this section it will be shown how this adjustment is made to obtain the parameter estimators under such circumstances. This section is also based on the work by Browne and Du Toit (1992).

Suppose the set of r equality constraints on the parameters in the model is written as

$$\mathbf{c}(\boldsymbol{\gamma}) = \mathbf{0} \tag{3.51}$$

where $\mathbf{c}$ is an $r \times 1$ vector valued function and is assumed to be continuously differentiable with respect to the parameters in $\boldsymbol{\gamma}$.

The constraint function in (3.51) will in many practical applications not be linear, and in order to simplify derivations and calculations in such situations, this function will now be linearised.

A linear constraint function, as an approximation to $\mathbf{c}(\boldsymbol{\gamma})$ in (3.51), may be obtained by making use of a first order Taylor expansion of the non-linear constraint function. Let the Jacobian matrix of $\mathbf{c}$ be denoted by the $r \times q$ matrix $\mathbf{L}$ where

$$\mathbf{L}(\boldsymbol{\gamma}) = \frac{\partial}{\partial \boldsymbol{\gamma}'}\mathbf{c}(\boldsymbol{\gamma}).$$

The linear approximation of the constraint function follows now as

$$\mathbf{c}(\boldsymbol{\gamma}) \approx \mathbf{c}_t + \mathbf{L}_t(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}_t) \tag{3.52}$$

where $\mathbf{c}_t = \mathbf{c}(\hat{\boldsymbol{\gamma}}_t)$, $\mathbf{L}_t = \mathbf{L}(\hat{\boldsymbol{\gamma}}_t)$ and $\hat{\boldsymbol{\gamma}}_t$ is the t-th approximation to $\boldsymbol{\gamma}$ in the iteration procedure.

If expression (3.52) is now substituted into expression (3.51), it follows that the non-linear constraints are approximated by the linear constraints given by

$$\mathbf{L}_t(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}_t) = -\mathbf{c}_t. \tag{3.53}$$

In order to obtain the increment vector to be used in the iteration procedure, let $\boldsymbol{\lambda}_t$ be an $r \times 1$ vector containing Lagrange multipliers and let $\mathbf{D}_t$ be an arbitrary $r \times r$ diagonal matrix. The increment vector is now

$$\begin{pmatrix} \boldsymbol{\delta}_t \\ \boldsymbol{\lambda}_t \end{pmatrix} = \begin{pmatrix} \mathbf{H}_t + \mathbf{L}_t'\mathbf{D}_t\mathbf{L}_t & \mathbf{L}_t' \\ \mathbf{L}_t & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} -(\mathbf{g}_t + \mathbf{L}_t'\mathbf{D}_t\mathbf{c}_t) \\ -\mathbf{c}_t \end{pmatrix} \tag{3.54}$$

and the approximation of the minimiser of $F(\boldsymbol{\gamma})$ at the $(t+1)$-st iteration is obtained using this $\boldsymbol{\delta}_t$ and $\hat{\boldsymbol{\gamma}}_t$ in expression (3.29). The function for choosing the step size parameter α_t, however, is now different in the sense that it progresses to a point where the constraints are satisfied. This is accomplished by initially taking α_t=1 and, if necessary, halving it successively until

$$F(\hat{\gamma}_{t+1}) + |\boldsymbol{\lambda}_t|'|\mathbf{c}_{t+1}| < F(\hat{\gamma}_t) + |\boldsymbol{\lambda}_t|'|\mathbf{c}_t|$$

where $|\boldsymbol{\lambda}_t|$ is a vector formed by taking the absolute values of the elements of $\boldsymbol{\lambda}_t$.

When the procedure described above converges at $\hat{\gamma}_t = \hat{\gamma}$, the increment vector will be zero at that point, i.e. $\boldsymbol{\delta}_t = \mathbf{0}$ in (3.54). This implies that the equality

$$-\hat{\mathbf{g}} = \hat{\mathbf{L}}'\hat{\boldsymbol{\lambda}}$$

holds, which is a necessary condition for $\hat{\gamma}$ to be a minimum of $F(\gamma)$ subject to the constraints in (3.51).

This procedure now also provides a method of obtaining estimates of the covariance matrix of the parameter estimators as well as the covariance matrix of the vector of Lagrange multipliers. As in the unconstrained case, these covariance matrices are obtained from the inverted information matrix. In the present situation these matrices are obtained from the inverted matrix that appears in expression (3.54). To indicate how these matrices may be determined, write the inverted matrix in (3.54) at convergence as

$$\begin{pmatrix} \mathbf{H}_t + \mathbf{L}_t'\mathbf{D}_t\mathbf{L}_t & \mathbf{L}_t' \\ \mathbf{L}_t & \mathbf{0} \end{pmatrix}^{-1} = \begin{pmatrix} \hat{\mathbf{S}}_{\gamma\gamma} & \hat{\mathbf{S}}_{\gamma\lambda} \\ \hat{\mathbf{S}}_{\lambda\gamma} & \hat{\mathbf{S}}_{\lambda\lambda} \end{pmatrix}.$$

The estimated covariance matrix of the parameter estimators is now given by the expression

$$\text{Cov}(\hat{\gamma}, \hat{\gamma}') = N^{-1}\hat{\mathbf{S}}_{\gamma\gamma} \tag{3.55}$$

while the estimated covariance matrix of the Lagrange multipliers may be obtained from the expression

$$\text{Cov}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\lambda}}') = -N^{-1}(\hat{\mathbf{S}}_{\lambda\lambda} + \hat{\mathbf{D}}). \tag{3.56}$$

3.7 Proposed estimation procedure

Two methods of estimating parameters in multilevel models have been discussed. In the first of them, the MML method, an iterative algorithm called the EM algorithm may be used to obtain the parameter estimates. The second is the iterative Fisher scoring method. Both have advantages and disadvantages when implemented in practice; but in combination they provide a very useful method of obtaining estimates of parameters and their approximate standard errors.

An advantage of the EM algorithm is that the parameter vector is split into two subvectors; consequently the vectors and matrices appearing in the equations for this method are of smaller dimension. This means less computation in practice and therefore less computation time to complete the iterations. A disadvantage is that it does not make use of second order derivative information; consequently convergence of the process is very slow. Also, no estimates of the standard errors are available.

On the other hand, the Fisher scoring method does make use of second order derivative information, and so provides standard error estimates and converges more rapidly. But its equations use vectors and matrices of much larger dimension, and these require larger storage and more computation time.

Following from the above arguments, it seems to be good practice to combine the two methods to obtain parameter estimates and approximate standard errors. Since the iterations for the EM algorithm are much faster than those for the Fisher scoring

algorithm, a good way to start the process is to initialise the parameter vector to some values, e.g. zeros and ones, and use the EM algorithm to obtain close approximations of the parameter estimators. This may be achieved after, say, 50 to 100 EM iterations. The close approximations of the parameter vector may now be used as initial values in the Fisher scoring algorithm. With them as starting point, the Fisher scoring algorithm should converge within a very few iterations (less than 20).

3.8 Existing work in this field

Various topics in the field of multivariate multilevel modelling have received attention, also specifically when latent variables are included. Of the latter type, the work of Muthén (1989), McDonald and Goldstein (1989) and McDonald (1994) will now briefly be considered.

The work of Muthén (1989):

Muthén (1989) recognizes that, when engaged in latent variable modelling, heterogeneity between several populations may exist in that each population may have a different set of parameter values. He then discusses three methodologies for uncovering various forms of population heterogeneity. They are: Regular multiple-group latent variable modelling, multiple indicators multiple causes (MIMIC) modelling and multilevel modelling.

In the case of regular multiple-group latent variable modelling, the following factor analysis model in G groups is assumed:

$$\mathbf{y}_j = \boldsymbol{\nu}_j + \boldsymbol{\Lambda}\boldsymbol{\eta}_j + \boldsymbol{\epsilon}_j \quad , \quad j = 1, 2, \ldots, G$$

where $\boldsymbol{\nu}_j$ and $\boldsymbol{\Lambda}$ respectively contain intercept and slope parameters.

If it is assumed that $\mathrm{E}(\boldsymbol{\eta}_j)=\boldsymbol{\alpha}_j$, $\mathrm{Cov}(\boldsymbol{\eta}_j\boldsymbol{\eta}_j')=\boldsymbol{\Phi}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}_j\boldsymbol{\epsilon}_j')=\boldsymbol{\Theta}$, and that

$$\mathrm{E}(\mathbf{y}_j) = \boldsymbol{\nu}_j + \boldsymbol{\Lambda}\boldsymbol{\alpha}_j = \boldsymbol{\mu}_j$$

$$\mathrm{Cov}(\mathbf{y}_i\mathbf{y}_i') = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} = \boldsymbol{\Sigma},$$

it follows that the variable means are not invariant across groups while the covariance matrix is.

For a mixture of normal distributions with a common covariance matrix, denoted by $\boldsymbol{\Sigma}$, and mixture proportions w_j, the mean and covariance matrix (for the mixture distribution) follow by generalizing a two-group result as (the subscript M represents the mixture distribution)

$$\boldsymbol{\mu}_M = \sum_{j=1}^{G} w_j\boldsymbol{\mu}_j$$

and

$$\boldsymbol{\Sigma}_M = \boldsymbol{\Sigma} + \sum_{j=1}^{G} w_j(\boldsymbol{\mu}_j - \boldsymbol{\mu}_M)(\boldsymbol{\mu}_j - \boldsymbol{\mu}_M)'.$$

It is pointed out that the second term in the expression for $\boldsymbol{\Sigma}_M$ is in general such that the model that holds for $\boldsymbol{\Sigma}$ does not hold for $\boldsymbol{\Sigma}_M$ because of across-group heterogeneity. This means that even when a covariance structure model holds for each group, it may not hold for the mixture.

Two examples are given - one with and one without invariance of the measurement intercepts. In both cases distorted results are obtained, indicating that regular multiple-group latent variable modelling cannot always satisfactorily uncover population heterogeneity.

Next, Muthén (1989) proposes MIMIC modelling to capture population heterogeneity. These models incorporate a set of regressor variables to predict the latent variables and the set of observed response variables - the latent variables play an intervening role between the predictor and response variables.

The inclusion of predictor variables makes heterogeneity detection and modelling possible. It is shown that heterogeneity can be studied in two ways, each time making use of grouping variables among the predictor variables. In the first case, across-group variation in factor means is allowed for, while the second approach allows for across-group heterogeneity in measurement intercepts.

Two examples are given where MIMIC modelling was applied. In one example it is shown that this approach solves the problem of groups with too small sample sizes to produce stable correlations, by using a set of dummy predictor variables representing the groups, thus allowing for variation in factor means across the groups and also for variation in measurement intercepts.

The second example considered a model with a general and specific factors, and was carried out in two steps, effectively resulting in the analysis of a pooled within tetrachoric correlation matrix. This MIMIC-based analysis, with dummy predictor variables to account for group differences, was compared and found superior to a regular tetrachoric analysis with no predictors.

The two examples using MIMIC modelling also do not seem to satisfactorily uncover heterogeneity among different groups.

While the regular multiple-group latent variable modelling and MIMIC modelling have a fixed effects approach, an alternative modelling procedure is proposed by Muthén (1989) that incorporates random parameters. Such parameters are viewed as continuous random variables rather than varying over a finite number of groups.

Whereas the random parameter approach is well-established in regression with fixed regressors, the same cannot be said for the case of random and latent regressors, such as in factor analysis. The third methodology, namely multilevel modelling, is a contribution towards this field of research.

Since population heterogeneity often gives rise to a hierarchical structure in the data, models for such data - multilevel models - have become increasingly important. They relax both the assumption of identical distributions and of independent observations, the two parts of the well known i.i.d. assumption in classical modelling.

To illustrate the multilevel approach, Muthén (1989) considers a model he terms the Muthén-Satorra varying factor means model. This model was motivated by applications where heterogeneity could be expected for the levels of the factors.

For individual i in group j, the model is

$$\mathbf{y}_{ij} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_{ij} + \boldsymbol{\epsilon}_{ij}$$

where $\mathrm{E}(\boldsymbol{\epsilon}_{ij})=\mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}_{ij}, \boldsymbol{\epsilon}'_{ij})=\boldsymbol{\Theta}$, and

$$\boldsymbol{\eta}_{ij} = \boldsymbol{\alpha}_j + \boldsymbol{\omega}_{ij}$$

$$\boldsymbol{\alpha}_j = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\mathbf{z}_j + \boldsymbol{\delta}_{\alpha_j}$$

where $\boldsymbol{\alpha}_j$ and $\boldsymbol{\omega}_{ij}$ are respectively a group-level and individual-level random component, and $\mathbf{z}_j$ is a vector of observed group-level variables. It is also noted by Muthén (1989) that across-group variation in measurement intercepts can be allowed for by replacing $\boldsymbol{\nu}$ by $\boldsymbol{\nu}_j$, expressed as a function of $\mathbf{z}_j$ and an error term.

The likelihood for hierarchical data under the above model is given; also for the special case of balanced data (equal N_i's across groups) and no $\mathbf{z}$-vector.

It is shown that the likelihood expression can be reformulated which permits optimization via software such as LISCOMP.

An example is provided for balanced data and where LISCOMP was used to carry out

the analysis. It was found that the above model fitted sufficiently, and it was therefore unnecessary to include across-group variation in the measurement intercepts in the model. It is noted that a regular structural analysis of the regular sample covariance matrix gave similar results to those of the within covariance matrix. Such an analysis is of course incapable of uncovering between-group variation.

The work of McDonald and Goldstein (1989):

McDonald and Goldstein (1989) introduce a general two-level model for multivariate data, written for individual i in group j as the $q \times 1$ vector

$$\mathbf{y}_{ij} = \mathbf{y}_{2j} + \mathbf{y}_{1ij}$$

where it is assumed that $\text{Cov}(\mathbf{y}_{2j}, \mathbf{y}'_{1ij})=\mathbf{0}$ and, if $\mathbf{x}_j$ is a $p \times 1$ measurement vector characterizing the j-th group, then $\text{Cov}(\mathbf{x}_j, \mathbf{y}'_{1ij})=\mathbf{0}$.

Let

$$\mathbf{\Sigma}_2 = \text{Cov}(\mathbf{y}_{2j}, \mathbf{y}'_{2j})$$

and

$$\mathbf{\Sigma}_1 = \text{Cov}(\mathbf{y}_{1ij}, \mathbf{y}'_{1ij}).$$

Then the model implies that

$$\begin{aligned} \mathbf{\Sigma}_{yy} &= \text{Cov}(\mathbf{y}_{ij}, \mathbf{y}'_{ij}) \\ &= \mathbf{\Sigma}_2 + \mathbf{\Sigma}_1. \end{aligned}$$

These covariance matrices, however, could be structured if a general two-level common factor model is defined as

$$\begin{pmatrix} \mathbf{x}_j \\ \mathbf{y}_{2j} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Lambda}_x \\ \boldsymbol{\Lambda}_2 \end{pmatrix} \mathbf{v}_2 + \begin{pmatrix} \mathbf{e}_{xj} \\ \mathbf{e}_{2j} \end{pmatrix}$$

and

$$\mathbf{y}_{1ij} = \boldsymbol{\Lambda}_1 \mathbf{v}_1 + \mathbf{e}_{1ij}$$

where $\text{Cov}(\mathbf{v}_2, \mathbf{v}_2')=\boldsymbol{\Phi}_2$ and $\text{Cov}(\mathbf{v}_1, \mathbf{v}_1')=\boldsymbol{\Phi}_1$ and

$$\text{Cov}(\mathbf{e}_{xj}, \mathbf{e}_{xj}') = \boldsymbol{\Psi}_x$$

$$\text{Cov}(\mathbf{e}_{2j}, \mathbf{e}_{2j}') = \boldsymbol{\Psi}_2$$

$$\text{Cov}(\mathbf{e}_{1ij}, \mathbf{e}_{1ij}') = \boldsymbol{\Psi}_1.$$

This model specification now implies the following structured covariance matrices:

$$\boldsymbol{\Sigma}_{xx} = \boldsymbol{\Lambda}_x \boldsymbol{\Phi}_2 \boldsymbol{\Lambda}_x' + \boldsymbol{\Psi}_x$$

$$\boldsymbol{\Sigma}_{xy} = \boldsymbol{\Lambda}_x \boldsymbol{\Phi}_2 \boldsymbol{\Lambda}_2'$$

$$\boldsymbol{\Sigma}_2 = \boldsymbol{\Lambda}_2 \boldsymbol{\Phi}_2 \boldsymbol{\Lambda}_2' + \boldsymbol{\Psi}_2$$

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1 \boldsymbol{\Lambda}_1' + \boldsymbol{\Psi}_1$$

and consequently

$$\boldsymbol{\Sigma}_{yy} = (\boldsymbol{\Lambda}_2 \ \boldsymbol{\Lambda}_1) \begin{pmatrix} \boldsymbol{\Phi}_2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Phi}_1 \end{pmatrix} \begin{pmatrix} \boldsymbol{\Lambda}_2' \\ \boldsymbol{\Lambda}_1' \end{pmatrix} + \boldsymbol{\Psi}_2 + \boldsymbol{\Psi}_1.$$

The log-likelihood function for the general model is given, and expressed in a form that requires only matrix inversion or computation of determinants of matrices of the order $\max(p, q)$. It is also shown that in the balanced case, the log-likelihood function can be expressed in terms of a convenient set of sufficient statistics.

Further, the first order derivatives of the log-likelihood with respect to the model parameters are provided. Again, results are presented separately for the balanced case.

The work of McDonald (1994):

McDonald (1994) extends the RAM model of McArdle and McDonald (1984) - a model for path analysis with latent variables - to a model for analyzing two-level data. He defines a two-level RAM model as the level-two model

$$\begin{pmatrix} \mathbf{x}_j \\ \mathbf{y}_{2j} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{xx} & \mathbf{A}_{x2} \\ \mathbf{A}_{2x} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{x}_j \\ \mathbf{y}_{2j} \end{pmatrix} + \begin{pmatrix} \mathbf{u}_{xj} \\ \mathbf{u}_{2j} \end{pmatrix} +$$

with the level-one model

$$\mathbf{y}_{1ij} = \mathbf{A}_1 \mathbf{y}_{1ij} + \mathbf{e}_{1ij}.$$

Here, the vectors $\mathbf{x}_j$ (measures on the level-two units), $\mathbf{y}_{2j}$ and $\mathbf{y}_{1ij}$ contain observable as well as latent variables.

The above model statements yield

$$\mathrm{Cov}(\mathbf{y}_{1ij}, \mathbf{y}'_{1ij}) = (\mathbf{I} - \mathbf{A}_1)^{-1} \mathbf{S}_1 (\mathbf{I} - \mathbf{A}_1)'^{-1}$$

and

$$\mathrm{Cov}\left[\begin{pmatrix} \mathbf{x}_j \\ \mathbf{y}_{2j} \end{pmatrix}, (\mathbf{x}'_j \; \mathbf{y}'_{2j}) \right] = \begin{bmatrix} \mathbf{I} - \mathbf{A}_{xx} & -\mathbf{A}_{x2} \\ -\mathbf{A}_{2x} & \mathbf{I} - \mathbf{A}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{S}_{xx} & \mathbf{S}_{x2} \\ \mathbf{S}_{2x} & \mathbf{S}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I} - \mathbf{A}_{xx} & -\mathbf{A}_{x2} \\ -\mathbf{A}_{2x} & \mathbf{I} - \mathbf{A}_{22} \end{bmatrix}^{-1}$$

where

$$\mathrm{Cov}(\mathbf{e}_{1ij}, \mathbf{e}'_{1ij}) = \mathbf{S}_1$$

$$\text{Cov}\left[\begin{pmatrix}\mathbf{u}_{xj}\\ \mathbf{u}_{2j}\end{pmatrix},(\mathbf{u}'_{xj}\ \mathbf{u}'_{2j})\right]=\begin{bmatrix}\mathbf{S}_{xx} & \mathbf{S}_{x2}\\ \mathbf{S}_{2x} & \mathbf{S}_{22}\end{bmatrix}.$$

A program, BIRAM, has been developed to apply the above model. For recursive models, the program allows a reparameterization at both levels to obtain standardized coefficients with correct standard errors.

3.9 Summary

The concept of hierarchically structured populations and data obtained from them, are introduced in this chapter. A general multilevel model for the analysis of such data is then discussed in detail. The univariate case is then extended to a general multivariate two-level model. Two estimation procedures for estimating the unknown parameters in this two-level model, are presented. The first is the method of marginal maximum likelihood. This method is used to obtain general equations which lend itself to be used in an iterative EM algorithm to obtain the parameter estimates. The second method is the well known Fisher scoring method. The general equations for the maximum likelihood parameter estimators are derived for use in the iterative Fisher scoring algorithm. It is shown next how both estimation procedures may be adjusted to estimate the parameters when constraints are imposed on them. Subsequently, a two-stage procedure which makes use of both the MML method and Fisher scoring method is proposed for practical use. The last section of the chapter provides a summary of some work that has been done in the analysis of multilevel models, specifically when latent variables are included.

CHAPTER 4

BILEVEL FACTOR ANALYSIS MODELS AND EXPECTED MAXIMISATION

4.1 Introduction

In this chapter a two-level factor analysis model will be defined and it will be shown how its parameters may be estimated by means of expected maximisation. Two situations, namely exploratory and confirmatory analysis will be considered. Model identification, a general problem in factor analysis models, and methods to ensure identified models are considered. The final section of the chapter gives a practical application of this method of parameter estimation using real data.

4.2 Factor analysis models for hierarchical data structures

A general two-level multivariate model is defined by Goldstein and McDonald (1988) and is discussed in Chapter 3. McDonald and Goldstein (1989) consider a two-level model for linear structural relations. McDonald (1993) shows that there is a sense in which this is a special (but degenerate) case of Goldstein and McDonald (1988). We note that Goldstein and McDonald (1989), and likewise Longford and Muthén (1992) and the present work, does not allow a "slopes as outcomes" model, with factor loadings random over level two. McDonald and Goldstein (1989) give, under normality assumptions, the likelihood and its first order derivatives as a basis for determining the parameter estimates. They show that, in the balanced case, the sample mean vectors and covariance matrices are minimal sufficient statistics, and that these (or functions of them) are maximum likelihood estimates of the unrestricted parameter matrices. They also derive a likelihood ratio test to test restricted parameter matrices against a general alternative (in the balanced case).

Longford and Muthén (1992) consider a two-level factor analysis model that is a special case of the model of McDonald and Goldstein (1989). They assume a general unbalanced design, whereas McDonald and Goldstein (1989) concentrated on efficient estimation in balanced designs. Longford and Muthén (1992) give the log-likelihood function and rewrite it in a computationally more efficient form, and also present it for the balanced case. First and second order partial derivatives of the log-likelihood (general unbalanced case) are derived and they suggest that a Fisher-scoring algorithm should be used instead of a Newton-Raphson algorithm, since the expectation of the second order derivative matrix is substantially simpler than the matrix of exact derivatives. They also note that the algorithm can be adapted to handle constraints by applying the chain-rule or the method of Lagrange multipliers. They base their hypothesis testing and model checking on how much the deviance (-2 log-likelihood) of a restricted model differs from that of the saturated model - the difference is a chi-squared variate.

4.3 A two-level factor analysis model

This section summarizes the two-level factor analysis model of Longford and Muthén (1992).

In many cases the classical factor analysis model is applied to inappropriate data because of the assumption of independence of the vectors of observations. This assumption may not be entirely true in hierarchically structured data where, for example, students are observed within classrooms. In such a situation it is often reasonable to assume that the students in a classroom are more similar, since they share a common environment. It may therefore be necessary to model this within-group homogeneity (or between-group variation) by a group-level (or between-group) covariance structure, but also to model the within-group variation by an individual-level (or within-group) covariance structure (Longford and Muthén, 1992).

A two-level factor analysis model that assumes a common factor structure at each level

(group- and individual-level) was discussed by Longford and Muthén (1992). That same model will be considered in this chapter as a special case, and it will be shown how it fits into the general model given by (3.3).

Consider the same two-level hierarchical structure that was introduced in Chapter 3, where there are M groups and in group i there are n_i observations. Let each p-variate observation made on an individual, say the j-th individual in the i-th group, be denoted by the $p \times 1$ vector $\mathbf{y}_{ij}$. These vectors are assumed to be normally distributed random vectors of observations where, in total, there are $N = \sum_{i=1}^{M} n_i$ observations.

The model described by Longford and Muthén (1992) assumes that, conditional on the group mean $\mathbf{m}_i$, observations made on individuals within each group have a common factor structure. This means that the observations in a specific group, say group i, are independently and identically distributed and follow the model

$$\mathbf{y}_{ij} = \mathbf{m}_i + \mathbf{\Lambda}_1 \mathbf{d}_{1,ij} + \mathbf{e}_{1,ij} , \quad j = 1, 2, \ldots, n_i. \tag{4.1}$$

It is further assumed that the M mean vectors, $\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_M$, also have an underlying factor structure and consequently that they are independently and identically distributed according to the mathematical model

$$\mathbf{m}_i = \boldsymbol{\mu} + \mathbf{\Lambda}_2 \mathbf{d}_{2,i} + \mathbf{e}_{2,i} , \quad i = 1, 2, \ldots, M.$$

An alternative form to specify the two-level model would be to write

$$\mathbf{y}_{ij} = \boldsymbol{\mu} + \mathbf{\Lambda}_2 \mathbf{d}_{2,i} + \mathbf{e}_{2,i} + \mathbf{\Lambda}_1 \mathbf{d}_{1,ij} + \mathbf{e}_{1,ij} , \quad i = 1, 2, \ldots, M; \;\; j = 1, 2, \ldots, n_i. \tag{4.2}$$

Assumptions regarding the random parameters in this model are the following: at level

one it is assumed that $\mathbf{d}_{1,ij} \sim N(\mathbf{0}, \mathbf{\Phi}_1)$ and $\mathbf{e}_{1,ij} \sim N(\mathbf{0}, \mathbf{D}_1)$, while at the second level it is assumed that $\mathbf{d}_{2,i} \sim N(\mathbf{0}, \mathbf{\Phi}_2)$ and $\mathbf{e}_{2,i} \sim N(\mathbf{0}, \mathbf{D}_2)$. Further, it is assumed that these random parameters are mutually independent normal random samples.

The vector $\boldsymbol{\mu}$, the two $\mathbf{\Lambda}$-matrices, the two $\mathbf{\Phi}$-matrices and the two $\mathbf{D}$-matrices contain the fixed parameters of the model. $\boldsymbol{\mu}$ represents an overall mean effect, while the $\mathbf{\Lambda}$'s contain the factor loadings with $\mathbf{\Lambda_1}$ a $p \times r_1$ matrix and $\mathbf{\Lambda_2}$ a $p \times r_2$ matrix, indicating r_1 and r_2 factors in the within- and between-group structures respectively. The $\mathbf{\Phi}$'s represent factor covariance matrices while the $\mathbf{D}$'s are assumed to be diagonal matrices representing the error variances.

To write down the model for the i-th group, let the vector of observations for this group be denoted by $\mathbf{y}_i$, where $\mathbf{y}_i$ is the $pn_i \times 1$ vector defined as

$$\mathbf{y}_i = \begin{pmatrix} \mathbf{y}_{i1} \\ \mathbf{y}_{i2} \\ \vdots \\ \mathbf{y}_{in_i} \end{pmatrix}.$$

In a similar fashion define $\mathbf{d}_{1,i}$ of order $r_1 n_i \times 1$ as

$$\mathbf{d}_{1,i} = \begin{pmatrix} \mathbf{d}_{1,i1} \\ \mathbf{d}_{1,i2} \\ \vdots \\ \mathbf{d}_{1,in_i} \end{pmatrix}$$

and define $\mathbf{e}_{1,i}$ of order $pn_i \times 1$ as

$$\mathbf{e}_{1,i} = \begin{pmatrix} \mathbf{e}_{1,i1} \\ \mathbf{e}_{1,i2} \\ \vdots \\ \mathbf{e}_{1,in_i} \end{pmatrix}.$$

For the i-th group, the model (4.2) can now be written as

$$\mathbf{y}_i = \mathbf{j}_{n_i} \otimes \mathbf{m}_i + (\mathbf{I}_{n_i} \otimes \boldsymbol{\Lambda}_1)\mathbf{d}_{1,i} + \mathbf{e}_{1,i}. \tag{4.3}$$

To show how this model fits into the general framework provided in Chapter 3, note that the first term in (4.3) may be rewritten as

$$\begin{aligned} \mathbf{j}_{n_i} \otimes \mathbf{m}_i &= \mathbf{j}_{n_i} \otimes \boldsymbol{\mu} + \mathbf{j}_{n_i} \otimes (\boldsymbol{\Lambda}_2\mathbf{d}_{2,i} + \mathbf{e}_{2,i}) \\ &= (\mathbf{j}_{n_i} \otimes \mathbf{I}_p)\boldsymbol{\mu} + (\mathbf{j}_{n_i} \otimes \mathbf{I}_p)(\boldsymbol{\Lambda}_2\mathbf{d}_{2,i} + \mathbf{e}_{2,i}). \end{aligned}$$

If the last part of (4.3) is also rewritten in a slightly different form, namely as

$$(\mathbf{I}_{n_i} \otimes \boldsymbol{\Lambda}_1)\mathbf{d}_{1,i} + \mathbf{e}_{1,i} = \mathbf{I}_{pn_i}[(\mathbf{I}_{n_i} \otimes \boldsymbol{\Lambda}_1)\mathbf{d}_{1,i} + \mathbf{e}_{1,i}] \ ,$$

it is evident that the two-level factor analysis model given by (4.3) is a special case of the general multivariate two-level model given by (3.3), with the random vectors $\boldsymbol{\beta}_{1i}$ and $\boldsymbol{\beta}_{2i}$ defined by

$$\boldsymbol{\beta}_{1i} = (\mathbf{I}_{n_i} \otimes \boldsymbol{\Lambda}_1)\mathbf{d}_{1,i} + \mathbf{e}_{1,i}$$

which contains the n_i subvectors $\boldsymbol{\beta}_{1i1} \ldots \boldsymbol{\beta}_{1in_i}$ where

$$\boldsymbol{\beta}_{1ij} = \boldsymbol{\Lambda}_1 \mathbf{d}_{1,ij} + \mathbf{e}_{1,ij} \ , \quad j = 1, 2, \ldots, n_i$$

and

$$\boldsymbol{\beta}_{2i} = \boldsymbol{\Lambda}_2 \mathbf{d}_{2,i} + \mathbf{e}_{2,i}.$$

The expected value and covariance matrix of the n_i observations in the i-th group can now be obtained for the model in (4.3), namely

$$\mathrm{E}(\mathbf{y}_i) = \mathbf{j}_{n_i} \otimes \boldsymbol{\mu}$$

and the covariance matrix for $\mathbf{y}_i$ may be obtained using the general expression given by (3.4), namely

$$\begin{aligned} \mathbf{W}_i &= \mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_i') \\ &= \mathbf{I}_{n_i} \otimes \mathbf{V}_1 + \mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes \mathbf{V}_2 \end{aligned} \tag{4.4}$$

where

$$\begin{aligned} \mathbf{V}_1 &= \mathrm{Cov}(\boldsymbol{\beta}_{1ij}, \boldsymbol{\beta}_{1ij}') \\ &= \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1 \boldsymbol{\Lambda}_1' + \mathbf{D}_1 \end{aligned}$$

and

$$\begin{aligned}\mathbf{V}_2 &= \text{Cov}(\boldsymbol{\beta}_{2i}, \boldsymbol{\beta}'_{2i}) \\ &= \boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2\boldsymbol{\Lambda}'_2 + \mathbf{D}_2.\end{aligned}$$

It is clear from (4.4) that $\mathbf{W}_i$ depends on n_i for determining its order. Consequently, the $\mathbf{W}_i$'s will be of different order unless all the n_i are equal, which is called the balanced case, implying that there is an equal number of individuals in each group.

4.4 The model parameters

The model description in the previous section shows that each p-variate vector of observations is expressed as a factor model with fixed slope parameters ($\boldsymbol{\Lambda}_1$), and random measurement intercepts ($\mathbf{m}_i$) that allow for across-group variation. The model in (4.1) therefore has the two vectors containing random parameters, $\mathbf{m}_i$ and $\mathbf{d}_{1,ij}$. The fixed parameters in the model are the elements of $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}_1$, $\boldsymbol{\Phi}_1$, $\mathbf{D}_1$, $\boldsymbol{\Lambda}_2$, $\boldsymbol{\Phi}_2$ and $\mathbf{D}_2$. In total there are therefore $q^\star$ fixed parameters that need estimation, where

$$\begin{aligned}q^\star &= p + pr_1 + r_1(r_1+1)/2 + p + pr_2 + r_2(r_2+1)/2 + p \\ &= 3p + p(r_1+r_2) + r_1(r_1+1)/2 + r_2(r_2+1)/2.\end{aligned}$$

In order to ensure that the estimators for $\boldsymbol{\Phi}_1$ and $\boldsymbol{\Phi}_2$ are positive definite matrices, $\boldsymbol{\Phi}_1$ and $\boldsymbol{\Phi}_2$ are written as the product of a lower triangular matrix with its transpose - the number of parameters is therefore not affected. In particular, we write

$$\boldsymbol{\Phi}_1 = \mathbf{U}_1\mathbf{U}'_1$$

and

$$\mathbf{\Phi}_2 = \mathbf{U}_2\mathbf{U}_2'$$

where $\mathbf{U}_1$ and $\mathbf{U}_2$ are respectively an $r_1 \times r_1$ and an $r_2 \times r_2$ lower triangular matrix. In a practical application, estimates for the $r_1(r_1+1)/2$ and $r_2(r_2+1)/2$ elements of $\mathbf{U}_1$ and $\mathbf{U}_2$ respectively are obtained, and then the estimates for $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$ are calculated using the above expressions, yielding positive definite estimates for $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$.

It should be noted that no constraints are placed on $\mathbf{D}_1$ and $\mathbf{D}_2$ to ensure positive definiteness. See, for example, Martin and McDonald (1975) in this regard.

4.5 Model identification and constraints

Similar to the identification problem that exists in classical factor analysis, the parameters in the bilevel factor analysis model are not uniquely defined unless r_1^2 and r_2^2 independent restrictions are imposed on the elements of $(\mathbf{\Lambda}_1, \mathbf{\Phi}_1)$ and $(\mathbf{\Lambda}_2, \mathbf{\Phi}_2)$ respectively. Identification problems arise from the relative scale of factor variances at the two levels. Choices include: (i) setting factor variances to unity, (ii) equating a loading to unity at both levels, (iii) equating loadings across levels and setting factor variances to unity at one level (McDonald, 1994), and (iv) employing correlation structures at one or both levels (McDonald, 1994). The choice made here to represent $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$ by their Cholesky factors preclude (iii) and (iv). In exploratory analysis one could employ similar constraints as in the classical case, namely to fix the scale of orthogonal factors by constraining $\mathbf{\Phi}_1 = \mathbf{I}$ and $\mathbf{\Phi}_2 = \mathbf{I}$ (choice (i)), and then the remaining $r_1(r_1-1)/2$ and $r_2(r_2-1)/2$ identification conditions are imposed on the elements of $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ respectively by constraining $\mathbf{\Lambda}_1'\mathbf{D}_1^{-1}\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2'\mathbf{D}_2^{-1}\mathbf{\Lambda}_2$ to be diagonal matrices. In classical factor analysis, this method of removing the indeterminacies leads to a mathematical convenient way of parameter estimation, involving eigenvalues and eigenvectors. Such convenience, however, leading to efficient parameter estimation, is not possible in the bilevel model since it involves non-linear constraints at both levels. This method will

also lead to a solution that will generally not be interpretable, and subsequent rotations will have to be employed.

In a confirmatory model, the indeterminacies are removed by specifying certain parameter values in advance. Most of the times the factor variances are fixed at unity to fix the scale of the factors, and the remaining identification conditions are then imposed by fixing elements of the loading matrices at zero. Another method to fix the scale of the factors is to fix a loading on each factor at unity, and leave the factor covariance matrix free for estimation.

Since not only the number of fixed parameters is sufficient to ensure identification, but also their position, some guidelines will now be provided that may assist in deciding which parameters to fix at pre-specified values.

Some existing work in this field include contributions by Howe (1955), Jöreskog (1969), Jennrich (1978) and Dunn (1973). These contributions were summarised by Jöreskog (1979) and the conditions for a unique solution were given for four cases. These were: (i) orthogonal solution with fixed zero elements, (ii) orthogonal solution with arbitrary fixed elements, (iii) oblique solution with fixed zero elements and (iv) oblique solution with arbitrary fixed elements.

We shall only consider case (iii) since the practical examples we shall present are like that. Also, in practice in general it is the most interesting and used case. Suppose now we have a $p \times m$ factor matrix and that the m factor variances are left free for estimation. This means that each column of $\mathbf{\Lambda}$ should contain a parameter that is fixed at unity to fix the scales of the factors. These fixed parameters should be in different rows. This leaves at least $m^2 - m = m(m-1)$ more parameters that need to be fixed (at zero) and can be accomplished by fixing $m - 1$ parameters in each column. It may happen that the positions of these fixed zeroes will result in the third condition of Jöreskog (1979), as set out in Section 2.3, not being satisfied. A trivial example of this would be if there are m zeroes in the same row - which is non-practical since it means that the relevant

variable is not associated to any of the factors. The third condition will also not be satisfied if more than one zero is specified in $m-1$ rows, and the zeroes are in the same columns and they are the only zeroes in at least one of the columns.

It seems that when zeroes are specified, one should try not to specify too many in the same row, and their positions in the rows should as far as possible not be the same.

4.6 The MML estimators of the free parameters

As indicated in Chapter 3, the MML method depends on two density functions, each being a function of a subset of the unknown parameters, in order to obtain the gradient of ℓnL. These two functions are $g(\mathbf{b}_i)$ and $f(\mathbf{y}_i|\mathbf{b}_i)$, where $\mathbf{b}_i$ is a vector containing the random parameters, and is therefore defined as

$$\mathbf{b}_i = \begin{pmatrix} \mathbf{m}_i \\ \mathbf{d}_{1,i} \end{pmatrix}.$$

The two density functions, $g(\mathbf{b}_i)$ and $f(\mathbf{y}_i|\mathbf{b}_i)$, are normal density functions according to the assumptions. If we write $\mathrm{E}(\mathbf{b}_i) = \boldsymbol{\beta}$ and $k_i = p + r_1 n_i$ it follows that they are given by

$$\begin{aligned}
g(\mathbf{b}_i) &= (2\pi)^{-\frac{k_i}{2}} \left| \begin{matrix} \mathbf{V}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n_i} \otimes \boldsymbol{\Phi}_1 \end{matrix} \right|^{-\frac{1}{2}} \exp\left\{ -\tfrac{1}{2}\mathrm{tr} \begin{pmatrix} \mathbf{V}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n_i} \otimes \boldsymbol{\Phi}_1 \end{pmatrix}^{-1} (\mathbf{b}_i - \boldsymbol{\beta})(\mathbf{b}_i - \boldsymbol{\beta})' \right\} \\
&= (2\pi)^{-\frac{k_i}{2}} |\mathbf{V}_2|^{-\frac{1}{2}} |\mathbf{I}_{n_i} \otimes \boldsymbol{\Phi}_1|^{-\frac{1}{2}} \exp\left\{ -\tfrac{1}{2}\mathrm{tr} \begin{pmatrix} \mathbf{V}_2^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n_i} \otimes \boldsymbol{\Phi}_1^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{m}_i - \boldsymbol{\mu} \\ \mathbf{d}_{1,i} \end{pmatrix} ((\mathbf{m}_i - \boldsymbol{\mu})' \; ; \; \mathbf{d}_{1,i}') \right\} \\
&= (2\pi)^{-\frac{k_i}{2}} |\mathbf{V}_2|^{-\frac{1}{2}} |\boldsymbol{\Phi}_1|^{-\frac{n_i}{2}} \exp\left\{ -\tfrac{1}{2} ((\mathbf{m}_i - \boldsymbol{\mu})' \; ; \; \mathbf{d}_{1,i}') \begin{pmatrix} \mathbf{V}_2^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n_i} \otimes \boldsymbol{\Phi}_1^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{m}_i - \boldsymbol{\mu} \\ \mathbf{d}_{1,i} \end{pmatrix} \right\} \\
&= (2\pi)^{-\frac{k_i}{2}} |\mathbf{V}_2|^{-\frac{1}{2}} |\boldsymbol{\Phi}_1|^{-\frac{n_i}{2}} \exp\left\{ -\tfrac{1}{2}(\mathbf{m}_i - \boldsymbol{\mu})'\mathbf{V}_2^{-1}(\mathbf{m}_i - \boldsymbol{\mu}) - \tfrac{1}{2}\sum_{j=1}^{n_i} \mathbf{d}_{1,ij}'\boldsymbol{\Phi}_1^{-1}\mathbf{d}_{1,ij} \right\}
\end{aligned}$$

and

$$f(\mathbf{y}_i|\mathbf{b}_i) = (2\pi)^{-\frac{pn_i}{2}}|\mathbf{I}_{n_i} \otimes \mathbf{D}_1|^{-\frac{1}{2}}\exp\left\{ -\tfrac{1}{2}\mathrm{tr}(\mathbf{I}_{n_i} \otimes \mathbf{D}_1)^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_{\mathbf{y}_i})(\mathbf{y}_i - \boldsymbol{\mu}_{\mathbf{y}_i})' \right\}$$

$$= (2\pi)^{-\frac{pn_i}{2}}|\mathbf{D}_1|^{-\frac{n_i}{2}}\exp\left\{ -\tfrac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_{\mathbf{y}_i})'(\mathbf{I}_{n_i} \otimes \mathbf{D}_1^{-1})(\mathbf{y}_i - \boldsymbol{\mu}_{\mathbf{y}_i}) \right\}$$

where

$$\boldsymbol{\mu}_{\mathbf{y}_i} = \mathbf{j}_{n_i} \otimes \mathbf{m}_i + (\mathbf{I}_{n_i} \otimes \boldsymbol{\Lambda}_1)\mathbf{d}_{1,i}.$$

Inspection of the above expressions will show that $g(\mathbf{b}_i)$ contains the unknown parameters $\boldsymbol{\mu}$, $\mathbf{U}_1$, $\boldsymbol{\Lambda}_2$, $\mathbf{U}_2$ and $\mathbf{D}_2$ while $f(\mathbf{y}_i|\mathbf{b}_i)$ contains $\boldsymbol{\Lambda}_1$ and $\mathbf{D}_1$. Let these two sets of parameters now be denoted by the $t_1 \times 1$ vector $\boldsymbol{\theta}$ and the $t_2 \times 1$ vector $\boldsymbol{\tau}$ where $t_1 = p(r_2 + 2) + r_2(r_2 + 1)/2 + r_1(r_1 + 1)/2$ and $t_2 = p(r_1 + 1)$, and where

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\mu} \\ \mathrm{vec}(\boldsymbol{\Lambda}_2) \\ \mathrm{vecs}(\mathbf{U}_2) \\ \mathrm{diag}(\mathbf{D}_2) \\ \mathrm{vecs}(\mathbf{U}_1) \end{pmatrix}$$

and

$$\boldsymbol{\tau} = \begin{pmatrix} \mathrm{vec}(\boldsymbol{\Lambda}_1) \\ \mathrm{diag}(\mathbf{D}_1) \end{pmatrix}.$$

Consequently $\boldsymbol{\theta}$ represents the unknown parameters in $g(\mathbf{b}_i)$ and $\boldsymbol{\tau}$ represents the unknown parameters in $f(\mathbf{y}_i|\mathbf{b}_i)$.

Use will now be made of Proposition 3.1 to obtain the partial derivatives of the log-likelihood function of the $\mathbf{y}_i$ $(i = 1, 2 \ldots, M)$ vectors with respect to the elements of $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$. First, the natural logarithms of $g(\mathbf{b}_i)$ and $f(\mathbf{y}_i|\mathbf{b}_i)$ must be obtained. Omitting the constant terms, the logarithms of these functions are given by

$$\ell n\ g(\mathbf{b}_i) = -\frac{1}{2}\ \ell n|\mathbf{V}_2| - \frac{n_i}{2}\ \ell n|\boldsymbol{\Phi}_1| - \frac{1}{2}(\mathbf{m}_i - \boldsymbol{\mu})'\mathbf{V}_2^{-1}(\mathbf{m}_i - \boldsymbol{\mu}) - \frac{1}{2}\sum_{j=1}^{n_i}\mathbf{d}'_{1,ij}\boldsymbol{\Phi}_1^{-1}\mathbf{d}_{1,ij}$$

and

$$\begin{aligned}\ell n\ f(\mathbf{y}_i|\mathbf{b}_i) &= -\frac{n_i}{2}\ell n|\mathbf{D}_1| - \frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_{\mathbf{y}_i})'(\mathbf{I}_{n_i} \otimes \mathbf{D}_1^{-1})(\mathbf{y}_i - \boldsymbol{\mu}_{\mathbf{y}_i}) \\ &= -\frac{n_i}{2}\ell n|\mathbf{D}_1| - \frac{1}{2}\sum_{\ell=1}^{n_i}(\mathbf{y}_{i\ell} - \boldsymbol{\mu}_{\mathbf{y}_{i\ell}})'\mathbf{D}_1^{-1}(\mathbf{y}_{i\ell} - \boldsymbol{\mu}_{\mathbf{y}_{i\ell}})\end{aligned}$$

where

$$\boldsymbol{\mu}_{\mathbf{y}_{i\ell}} = \mathbf{m}_i + \boldsymbol{\Lambda}_1\mathbf{d}_{1,i\ell}.$$

The next step will be to obtain the partial derivatives - see (3.6) and (3.7) - of $\ell n\ g(\mathbf{b}_i)$ and $\ell n\ f(\mathbf{y}_i|\mathbf{b}_i)$ with respect to the elements of $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ respectively. Since both these functions are natural logarithms of normal density functions, well known results - cf. expression (3.26) - can be used to obtain the derivatives. In particular, the expressions for the derivative of $\ell n\ g(\mathbf{b}_i)$ and $\ell n\ f(\mathbf{y}_i|\mathbf{b}_i)$ with respect to a general parameter in $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ respectively, are

$$\begin{aligned}\frac{\partial \ell n\ g(\mathbf{b}_i)}{\partial \theta_\ell} &= \frac{1}{2}\mathrm{tr}\left[\mathbf{V}_2^{-1}\left\{(\mathbf{m}_i-\boldsymbol{\mu})(\mathbf{m}_i-\boldsymbol{\mu})'-\mathbf{V}_2\right\}\mathbf{V}_2^{-1}\frac{\partial \mathbf{V}_2}{\partial \theta_\ell}\right] \\ &\quad +\mathrm{tr}\left[(\mathbf{m}_i-\boldsymbol{\mu})'\mathbf{V}_2^{-1}\frac{\partial \boldsymbol{\mu}}{\partial \theta_\ell}\right] \\ &\quad +\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Phi}_1^{-1}\left(\sum_{j=1}^{n_i}\mathbf{d}_{1,ij}\mathbf{d}_{1,ij}'-n_i\boldsymbol{\Phi}_1\right)\boldsymbol{\Phi}_1^{-1}\frac{\partial \boldsymbol{\Phi}_1}{\partial \theta_\ell}\right] \qquad (4.5)\end{aligned}$$

and

$$\begin{aligned}\frac{\partial \ell n\ f(\mathbf{y}_i|\mathbf{b}_i)}{\partial \tau_\ell} &= \frac{1}{2}\sum_{j=1}^{n_i}\mathrm{tr}\left[\mathbf{D}_1^{-1}\left\{(\mathbf{y}_{ij}-\boldsymbol{\mu}_{\mathbf{y}_{ij}})(\mathbf{y}_{ij}-\boldsymbol{\mu}_{\mathbf{y}_{ij}})'-\mathbf{D}_1\right\}\mathbf{D}_1^{-1}\frac{\partial \mathbf{D}_1}{\partial \tau_\ell}\right] \\ &\quad +\sum_{j=1}^{n_i}\mathrm{tr}\left[(\mathbf{y}_{ij}-\boldsymbol{\mu}_{\mathbf{y}_{ij}})'\mathbf{D}_1^{-1}\frac{\partial \boldsymbol{\mu}_{\mathbf{y}_{ij}}}{\partial \tau_\ell}\right]. \qquad (4.6)\end{aligned}$$

Expressions (4.5) and (4.6) give the general form of the derivatives with respect to the unknown parameters. The following propositions provide identities necessary for obtaining the MML estimators of the parameters.

Proposition 4.1

$$\mathrm{E}_c\left[(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})'\right] = \mathrm{Cov}_c(\mathbf{m}_i, \mathbf{m}_i') + (\mathrm{E}_c(\mathbf{m}_i) - \boldsymbol{\mu})\,(\mathrm{E}_c(\mathbf{m}_i) - \boldsymbol{\mu})'$$

Proof

Rewrite $(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})'$ in the equivalent form

$$[(\mathbf{m}_i - \mathrm{E}_c(\mathbf{m}_i)) + (\mathrm{E}_c(\mathbf{m}_i) - \boldsymbol{\mu})][(\mathbf{m}_i - \mathrm{E}_c(\mathbf{m}_i)) + (\mathrm{E}_c(\mathbf{m}_i) - \boldsymbol{\mu})]'.$$

The conditional expected value of this expression is

$$\begin{aligned} & \mathrm{E}_c[\mathbf{m}_i - \mathrm{E}_c(\mathbf{m}_i)][\mathbf{m}_i - \mathrm{E}_c(\mathbf{m}_i)]' + [\mathrm{E}_c(\mathbf{m}_i) - \boldsymbol{\mu}][\mathrm{E}_c(\mathbf{m}_i) - \boldsymbol{\mu}]' \\ & +\mathrm{E}_c\left(\mathbf{m}_i - \mathrm{E}_c(\mathbf{m}_i)\right)\left(\mathrm{E}_c(\mathbf{m}_i) - \boldsymbol{\mu}\right)' + \mathrm{E}_c\left(\mathrm{E}_c(\mathbf{m}_i) - \boldsymbol{\mu}\right)\left(\mathbf{m}_i - \mathrm{E}_c(\mathbf{m}_i)\right)' \\ = \;& \mathrm{Cov}_c(\mathbf{m}_i, \mathbf{m}_i') + (\mathrm{E}_c(\mathbf{m}_i) - \boldsymbol{\mu})\,(\mathrm{E}_c(\mathbf{m}_i) - \boldsymbol{\mu})' \end{aligned}$$

where $\mathrm{Cov}_c(\mathbf{m}_i, \mathbf{m}_i')$ denotes the conditional covariance matrix of $\mathbf{b}_i$ given $\mathbf{y}_i$.

The two cross products have vanished, since

$$\begin{aligned} & \mathrm{E}_c\left[\mathbf{m}_i\mathrm{E}_c(\mathbf{m}_i)' + \mathrm{E}_c(\mathbf{m}_i)\boldsymbol{\mu}' - \mathbf{m}_i\boldsymbol{\mu}' - \mathrm{E}_c(\mathbf{m}_i)\mathrm{E}_c(\mathbf{m}_i)'\right] \\ = \;& \mathrm{E}_c(\mathbf{m}_i)\mathrm{E}_c(\mathbf{m}_i)' + \mathrm{E}_c(\mathbf{m}_i)\boldsymbol{\mu}' - \mathrm{E}_c(\mathbf{m}_i)\boldsymbol{\mu}' - \mathrm{E}_c(\mathbf{m}_i)\mathrm{E}_c(\mathbf{m}_i)' \\ = \;& \mathbf{0}. \end{aligned}$$

A similar derivation will show that the other cross product term also vanishes. This proves the proposition. □

Proposition 4.2

$$\mathrm{E}_c(\mathbf{m}_i\mathbf{d}'_{1,ij}) = \mathrm{Cov}_c(\mathbf{m}_i\mathbf{d}'_{1,ij}) + \mathrm{E}_c(\mathbf{m}_i)\mathrm{E}_c(\mathbf{d}_{1,ij})'$$

Proof

$$\begin{aligned}\mathrm{E}_c(\mathbf{m}_i\mathbf{d}'_{1,ij}) &= \mathrm{E}_c\left[\mathbf{m}_i - \mathrm{E}_c(\mathbf{m}_i) + \mathrm{E}_c(\mathbf{m}_i)\right]\left[\mathbf{d}_{1,ij} - \mathrm{E}_c(\mathbf{d}_{1,ij}) + \mathrm{E}_c(\mathbf{d}_{1,ij})\right]' \\ &= \mathrm{E}_c\left(\mathbf{m}_i - \mathrm{E}_c(\mathbf{m}_i)\right)\left(\mathbf{d}_{1,ij} - \mathrm{E}_c(\mathbf{d}_{1,ij})\right)' + \mathrm{E}_c(\mathbf{m}_i)\mathrm{E}_c(\mathbf{d}_{1,ij})'\end{aligned}$$

from which the required result follows. □

Proposition 4.3

$$\mathrm{E}_c(\mathbf{d}_{1,ij}\mathbf{d}'_{1,ij}) = \mathrm{Cov}_c(\mathbf{d}_{1,ij}, \mathbf{d}'_{1,ij}) + \mathrm{E}_c(\mathbf{d}_{1,ij})\mathrm{E}_c(\mathbf{d}_{1,ij})'$$

Proof

$$\begin{aligned}\mathrm{E}_c(\mathbf{d}_{1,ij}\mathbf{d}'_{1,ij}) &= \mathrm{E}_c\left(\mathbf{d}_{1,ij} - \mathrm{E}_c(\mathbf{d}_{1,ij}) + \mathrm{E}_c(\mathbf{d}_{1,ij})\right)\left(\mathbf{d}_{1,ij} - \mathrm{E}_c(\mathbf{d}_{1,ij}) + \mathrm{E}_c(\mathbf{d}_{1,ij})\right)' \\ &= \mathrm{E}_c\left(\mathbf{d}_{1,ij} - \mathrm{E}_c(\mathbf{d}_{1,ij})\right)\left(\mathbf{d}_{1,ij} - \mathrm{E}_c(\mathbf{d}_{1,ij})\right)' + \mathrm{E}_c(\mathbf{d}_{1,ij})\mathrm{E}_c(\mathbf{d}_{1,ij})'\end{aligned}$$

from which the required result follows. □

Proposition 4.4

$$\mathrm{E}_c(\mathbf{m}_i\mathbf{m}_i') = \mathrm{Cov}_c(\mathbf{m}_i, \mathbf{m}_i') + \mathrm{E}_c(\mathbf{m}_i)\mathrm{E}_c(\mathbf{m}_i)'$$

Proof

$$\begin{aligned}\mathrm{E}_c(\mathbf{m}_i\mathbf{m}_i') &= \mathrm{E}_c\left(\mathbf{m}_i - \mathrm{E}_c(\mathbf{m}_i) + \mathrm{E}_c(\mathbf{m}_i)\right)\left(\mathbf{m}_i - \mathrm{E}_c(\mathbf{m}_i) + \mathrm{E}_c(\mathbf{m}_i)\right)' \\ &= \mathrm{E}_c\left(\mathbf{m}_i - \mathrm{E}_c(\mathbf{m}_i)\right)\left(\mathbf{m}_i - \mathrm{E}_c(\mathbf{m}_i)\right)' + \mathrm{E}_c(\mathbf{m}_i)\mathrm{E}_c(\mathbf{m}_i)'\end{aligned}$$

from which the required result follows. □

In addition to Propositions 4.1 to 4.4, use will also be made of the well-known identities

$$\mathrm{tr}[\mathbf{AB}] = \mathrm{tr}[\mathbf{A}'\mathbf{B}']$$

for $\mathbf{A}$ and $\mathbf{B}$ of suitable order, and

$$\mathrm{tr}[\mathbf{A}\mathbf{J}_{\ell k}] = [\mathbf{A}]_{k\ell}$$

in the derivation of the MML estimators.

The following five propositions will provide expressions for the MML estimators of the five natural subsets of the parameter vector.

Proposition 4.5

The MML estimator for $\boldsymbol{\mu}$ is given by the expression

$$\hat{\boldsymbol{\mu}} = \frac{1}{M}\sum_{i=1}^{M} \mathrm{E}_c(\mathbf{m}_i). \tag{4.7}$$

Proof

Consider the derivative of $\ell n\ g(\mathbf{b}_i)$ with respect to a typical element of $\boldsymbol{\mu}$, say μ_k $(k = 1, 2 \ldots, p)$. This derivative, using (4.5), is given by

$$\begin{aligned}
\frac{\partial \ell n\ g(\mathbf{b}_i)}{\partial \mu_k} &= \mathrm{tr}\left[(\mathbf{m}_i - \boldsymbol{\mu})'\mathbf{V}_2^{-1}\frac{\partial \boldsymbol{\mu}}{\partial \mu_k}\right] \\
&= \mathrm{tr}\left[(\mathbf{m}_i - \boldsymbol{\mu})'\mathbf{V}_2^{-1}\mathbf{J}_{k1}\right] \\
&= \left[\mathbf{V}_2^{-1}(\mathbf{m}_i - \boldsymbol{\mu})\right]_{k1}
\end{aligned}$$

so that the derivative of $\ell n\ g(\mathbf{b}_i)$ with respect to the vector $\boldsymbol{\mu}$ becomes

$$\frac{\partial \ell n\ g(\mathbf{b}_i)}{\partial \boldsymbol{\mu}} = \mathbf{V}_2^{-1}(\mathbf{m}_i - \boldsymbol{\mu}).$$

This result has to be substituted into (3.6) to obtain the gradient of $\ell n\ L$ with respect to $\boldsymbol{\mu}$. Therefore

$$\frac{\partial \ell n\ L}{\partial \mu} = \sum_{i=1}^{M} \mathrm{E}_c(\mathbf{V}_2^{-1}(\mathbf{m}_i - \boldsymbol{\mu}))$$

and setting this expression equal to zero and solving for $\boldsymbol{\mu}$, gives

$$\sum_{i=1}^{M} \mathrm{E}_c(\hat{\boldsymbol{\mu}}) = \sum_{i=1}^{M} \mathrm{E}_c(\mathbf{m}_i)$$

which proves the proposition. □

Proposition 4.6

The MML estimator for $\mathbf{\Lambda}_2$ is given by the expression

$$\hat{\mathbf{\Lambda}}_2 = \frac{1}{M}\sum_{i=1}^{M}\{\mathrm{Cov}_c(\mathbf{m}_i, \mathbf{m}_i') + (\mathrm{E}_c(\mathbf{m}_i) - \boldsymbol{\mu})(\mathrm{E}_c(\mathbf{m}_i) - \boldsymbol{\mu})'\}\,\hat{\mathbf{V}}_2^{-1}\hat{\mathbf{\Lambda}}_2. \tag{4.8}$$

Proof

Consider the derivative of $\ell n\ g(\mathbf{b}_i)$ with respect to a typical element of $\mathbf{\Lambda}_2$, say $[\mathbf{\Lambda}_2]_{k\ell}$ $(k = 1, 2, \ldots, p; \ell = 1, 2, \ldots, r_2)$. Using (4.5) to obtain this derivative yields

$$\frac{\partial \ell n\ g(\mathbf{b}_i)}{\partial [\mathbf{\Lambda}_2]_{k\ell}} = \frac{1}{2}\mathrm{tr}\left[\mathbf{V}_2^{-1}\{(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})' - \mathbf{V}_2\}\,\mathbf{V}_2^{-1}\frac{\partial \mathbf{V}_2}{\partial [\mathbf{\Lambda}_2]_{k\ell}}\right].$$

However, $\mathbf{V}_2 = \mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{\Lambda}_2' + \mathbf{D}_2$, so that

$$\frac{\partial \mathbf{V}_2}{\partial [\mathbf{\Lambda}_2]_{k\ell}} = \frac{\partial \mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{\Lambda}_2'}{\partial [\mathbf{\Lambda}_2]_{k\ell}} = \mathbf{J}_{k\ell}\mathbf{\Phi}_2\mathbf{\Lambda}_2' + \mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{J}_{\ell k}$$

and the derivative of $\ell n\ g(\mathbf{b}_i)$ therefore becomes

$$\begin{aligned}
\frac{\partial \ell n\ g(\mathbf{b}_i)}{\partial [\mathbf{\Lambda}_2]_{k\ell}} &= \frac{1}{2}\mathrm{tr}[\mathbf{V}_2^{-1}\{\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})' - \mathbf{V}_2\}\mathbf{V}_2^{-1}(\mathbf{J}_{k\ell}\mathbf{\Phi}_2\mathbf{\Lambda}_2' + \mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{J}_{\ell k})] \\
&= \mathrm{tr}[\mathbf{V}_2^{-1}\{(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})' - \mathbf{V}_2\}\mathbf{V}_2^{-1}\mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{J}_{\ell k}] \\
&= [\mathbf{V}_2^{-1}\{(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})' - \mathbf{V}_2\}\mathbf{V}_2^{-1}\mathbf{\Lambda}_2\mathbf{\Phi}_2]_{k\ell}.
\end{aligned}$$

The above result leads to the conclusion that the $p \times r_2$ matrix of derivatives is

$$\frac{\partial \ell n\, g(\mathbf{b}_i)}{\partial \boldsymbol{\Lambda}_2} = \mathbf{V}_2^{-1}\{(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})' - \mathbf{V}_2\}\mathbf{V}_2^{-1}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2.$$

Substituting this result into (3.6) leads to the following expression for the gradient of $\ell n\, L$ with respect to $\boldsymbol{\Lambda}_2$:

$$\frac{\partial \ell n\, L}{\partial \boldsymbol{\Lambda}_2} = \sum_{i=1}^{M} \mathrm{E}_c[\mathbf{V}_2^{-1}\{(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m} - \boldsymbol{\mu})' - \mathbf{V}_2\}\mathbf{V}_2^{-1}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2].$$

To obtain the maximum likelihood estimator of $\boldsymbol{\Lambda}_2$, this expression is set equal to zero and solved for $\boldsymbol{\Lambda}_2$.

Therefore

$$\sum_{i=1}^{M} \mathrm{E}_c[\hat{\mathbf{V}}_2^{-1}\{(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})' - \hat{\mathbf{V}}_2\}\hat{\mathbf{V}}_2^{-1}\hat{\boldsymbol{\Lambda}}_2\hat{\boldsymbol{\Phi}}_2] = \mathbf{0}$$

or, since $\hat{\mathbf{V}}_2^{-1}$ and $\hat{\boldsymbol{\Phi}}_2$ are constant with respect to the conditional expected value operator and $\hat{\mathbf{V}}_2^{-1} \neq \mathbf{0}$ and $\hat{\boldsymbol{\Phi}}_2 \neq \mathbf{0}$, it follows that

$$\sum_{i=1}^{M} \mathrm{E}_c[\{(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})' - \hat{\mathbf{V}}_2\}\hat{\mathbf{V}}_2^{-1}\hat{\boldsymbol{\Lambda}}_2] = \mathbf{0}$$

or, equivalently,

$$\sum_{i=1}^{M} \mathrm{E}_c[\hat{\boldsymbol{\Lambda}}_2] = \sum_{i=1}^{M} \mathrm{E}_c[(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})'\hat{\mathbf{V}}_2^{-1}\hat{\boldsymbol{\Lambda}}_2]$$

and therefore

$$\hat{\mathbf{\Lambda}}_2 = \frac{1}{M}\sum_{i=1}^{M} \mathrm{E}_c[(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})']\hat{\mathbf{V}}_2^{-1}\hat{\mathbf{\Lambda}}_2.$$

Use of Proposition 4.1 and the above completes the proof. □

Proposition 4.7

The MML estimator for $\mathbf{U}_2$ is given by the expression

$$\hat{\mathbf{U}}_2 = \left(\hat{\mathbf{\Lambda}}_2'\hat{\mathbf{\Lambda}}_2\right)^{-1}\hat{\mathbf{\Lambda}}_2'\frac{1}{M}\sum_{i=1}^{M}\left\{\mathrm{Cov}_c(\mathbf{m}_i, \mathbf{m}_i') + (\mathrm{E}_c(\mathbf{m}_i) - \boldsymbol{\mu})(\mathrm{E}_c(\mathbf{m}_i) - \boldsymbol{\mu})'\right\}\hat{\mathbf{V}}_2^{-1}\hat{\mathbf{\Lambda}}_2\hat{\mathbf{U}}_2. \quad (4.9)$$

Proof

Consider the derivative of $\ell n\ g(\mathbf{b}_i)$ with respect to a typical element of $\mathbf{U}_2$, say $[\mathbf{U}_2]_{k\ell}$ $(k = 1, 2, \ldots, r_2; \ell = 1, 2, \ldots, k)$. Using (4.5) to obtain this derivative yields

$$\frac{\partial \ell n\ g(\mathbf{b}_i)}{\partial [\mathbf{U}_2]_{k\ell}} = \frac{1}{2}\mathrm{tr}\left[\mathbf{V}_2^{-1}\left\{(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})' - \mathbf{V}_2\right\}\mathbf{V}_2^{-1}\frac{\partial \mathbf{V}_2}{\partial [\mathbf{U}_2]_{k\ell}}\right].$$

However, $\mathbf{V}_2 = \mathbf{\Lambda}_2\mathbf{U}_2\mathbf{U}_2'\mathbf{\Lambda}_2' + \mathbf{D}_2$, so that

$$\frac{\partial \mathbf{V}_2}{\partial [\mathbf{U}_2]_{k\ell}} = \frac{\partial \mathbf{\Lambda}_2\mathbf{U}_2\mathbf{U}_2'\mathbf{\Lambda}_2'}{\partial [\mathbf{U}_2]_{k\ell}} = \mathbf{\Lambda}_2\mathbf{J}_{k\ell}\mathbf{U}_2'\mathbf{\Lambda}_2' + \mathbf{\Lambda}_2\mathbf{U}_2\mathbf{J}_{\ell k}\mathbf{\Lambda}_2'$$

and the derivative of $\ell n\ g(\mathbf{b}_i)$ therefore becomes

$$\frac{\partial \ell n\ g(\mathbf{b}_i)}{\partial [\mathbf{U}_2]_{k\ell}} = \frac{1}{2}\mathrm{tr}[\mathbf{V}_2^{-1}\{\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})' - \mathbf{V}_2\}\mathbf{V}_2^{-1}(\mathbf{\Lambda}_2\mathbf{J}_{k\ell}\mathbf{U}_2'\mathbf{\Lambda}_2' + \mathbf{\Lambda}_2\mathbf{U}_2\mathbf{J}_{\ell k}\mathbf{\Lambda}_2')]$$

$$= \text{tr}[\boldsymbol{\Lambda}_2'\mathbf{V}_2^{-1}\{(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})' - \mathbf{V}_2\}\mathbf{V}_2^{-1}\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{\ell k}]$$

$$= [\boldsymbol{\Lambda}_2'\mathbf{V}_2^{-1}\{(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})' - \mathbf{V}_2\}\mathbf{V}_2^{-1}\boldsymbol{\Lambda}_2\mathbf{U}_2]_{k\ell}.$$

The above result leads to the conclusion that the $r_2 \times r_2$ lower triangular matrix of derivatives is

$$\frac{\partial \ell n\ g(\mathbf{b}_i)}{\partial \mathbf{U}_2} = \boldsymbol{\Lambda}_2'\mathbf{V}_2^{-1}\{(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})' - \mathbf{V}_2\}\mathbf{V}_2^{-1}\boldsymbol{\Lambda}_2\mathbf{U}_2.$$

Substituting this result into (3.6) leads to the following expression for the gradient of $\ell n\ L$ with respect to $\mathbf{U}_2$:

$$\frac{\partial \ell n\ L}{\partial \mathbf{U}_2} = \sum_{i=1}^{M} \mathrm{E}_c[\boldsymbol{\Lambda}_2'\mathbf{V}_2^{-1}\{(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m} - \boldsymbol{\mu})' - \mathbf{V}_2\}\mathbf{V}_2^{-1}\boldsymbol{\Lambda}_2\mathbf{U}_2].$$

To obtain the maximum likelihood estimator of $\mathbf{U}_2$, this expression is set equal to zero and solved for $\mathbf{U}_2$.

Therefore

$$\sum_{i=1}^{M} \mathrm{E}_c[\hat{\boldsymbol{\Lambda}}_2'\hat{\mathbf{V}}_2^{-1}\{(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})' - \hat{\mathbf{V}}_2\}\hat{\mathbf{V}}_2^{-1}\hat{\boldsymbol{\Lambda}}_2\hat{\mathbf{U}}_2] = \mathbf{0}$$

or, since $\hat{\mathbf{V}}_2^{-1}$ and $\hat{\boldsymbol{\Lambda}}_2$ are constant with respect to the conditional expected value operator and $\hat{\mathbf{V}}_2^{-1} \neq \mathbf{0}$ and $\hat{\boldsymbol{\Lambda}}_2 \neq \mathbf{0}$, it follows that

$$\sum_{i=1}^{M} \mathrm{E}_c[\{(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})' - \hat{\mathbf{V}}_2\}\hat{\mathbf{V}}_2^{-1}\hat{\boldsymbol{\Lambda}}_2\hat{\mathbf{U}}_2] = \mathbf{0}$$

or, equivalently,

$$\sum_{i=1}^{M} \mathrm{E}_c[\hat{\mathbf{\Lambda}}_2 \hat{\mathbf{U}}_2] = \sum_{i=1}^{M} \mathrm{E}_c[(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})' \hat{\mathbf{V}}_2^{-1} \hat{\mathbf{\Lambda}}_2 \hat{\mathbf{U}}_2]$$

and therefore

$$\hat{\mathbf{U}}_2 = \left(\hat{\mathbf{\Lambda}}_2' \hat{\mathbf{\Lambda}}_2\right)^{-1} \hat{\mathbf{\Lambda}}_2' \frac{1}{M} \sum_{i=1}^{M} \mathrm{E}_c[(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})'] \hat{\mathbf{V}}_2^{-1} \hat{\mathbf{\Lambda}}_2 \hat{\mathbf{U}}_2.$$

Use of Proposition 4.1 and the above completes the proof. □

Proposition 4.8

The MML estimator for $\mathbf{D}_2$ is given by the expression

$$\hat{\mathbf{D}}_2 = \frac{1}{M} \sum_{i=1}^{M} \mathrm{Diag}\left[\mathrm{Cov}_c(\mathbf{m}_i, \mathbf{m}_i') + (\mathrm{E}_c(\mathbf{m}_i) - \boldsymbol{\mu})(\mathrm{E}_c(\mathbf{m}_i) - \boldsymbol{\mu})'\right] - \mathrm{Diag}\left[\hat{\mathbf{\Lambda}}_2 \hat{\mathbf{\Phi}}_2 \hat{\mathbf{\Lambda}}_2'\right]. \tag{4.10}$$

Proof

Consider the derivative of $\ell n\ g(\mathbf{b}_i)$ with respect to the elements in $\mathbf{D}_2$. For a typical element in $\mathbf{D}_2$, say $[\mathbf{D}_2]_{kk}$ $(k = 1, 2, \ldots, p)$, the derivative of $\ell n\ g(\mathbf{b}_i)$ is (cf. (4.5))

$$\frac{\partial \ell n\ g(\mathbf{b}_i)}{\partial [\mathbf{D}_2]_{kk}} = \frac{1}{2} \mathrm{tr}\left[\mathbf{V}_2^{-1} \left\{(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})' - \mathbf{V}_2\right\} \mathbf{V}_2^{-1} \frac{\partial \mathbf{D}_2}{\partial [\mathbf{D}_2]_{kk}}\right].$$

This expression, however, can be simplified by using the reparameterisation theorem (Browne, 1991). It may be applied here since $\mathbf{D}_2$ and $\mathbf{D}_2^{-1}$ have the same structural form - namely they are both diagonal matrices - and consequently the derivative may be written as (if $\mathbf{D}_2^{\star} = \mathbf{D}_2^{-1}$)

$$\begin{aligned}\frac{\partial \ell n\ g(\mathbf{b}_i)}{\partial [\mathbf{D}_2^\star]_{kk}} &= -\frac{1}{2}\mathrm{tr}\left[\{(\mathbf{m}_i-\boldsymbol{\mu})(\mathbf{m}_i-\boldsymbol{\mu})'-\mathbf{V}_2\}\frac{\partial \mathbf{D}_2^\star}{\partial [\mathbf{D}_2^\star]_{kk}}\right] \\ &= -\frac{1}{2}\mathrm{tr}\left[\{(\mathbf{m}_i-\boldsymbol{\mu})(\mathbf{m}_i-\boldsymbol{\mu})'-\mathbf{V}_2\}\mathbf{J}_{kk}\right]\end{aligned}$$

so that the derivative with respect to the p unknown parameters in $\mathbf{D}_2^\star$ becomes

$$\frac{\partial \ell n\ g(\mathbf{b}_i)}{\partial \mathbf{D}_2^\star} = -\frac{1}{2}\mathrm{Diag}\left[(\mathbf{m}_i-\boldsymbol{\mu})(\mathbf{m}_i-\boldsymbol{\mu})'-\mathbf{V}_2\right].$$

This expression now has to be substituted into (3.6) yielding

$$\frac{\partial \ell n\ L}{\partial \mathbf{D}_2^\star} = \sum_{i=1}^{M}\mathrm{E}_c\left\{-\frac{1}{2}\mathrm{Diag}\left[(\mathbf{m}_i-\boldsymbol{\mu})(\mathbf{m}_i-\boldsymbol{\mu})'-\mathbf{V}_2\right]\right\}.$$

Setting this expression equal to zero and solving for $\mathbf{D}_2^\star$ will give the maximum likelihood estimator for $\mathbf{D}_2^\star$. However, one may solve directly for $\mathbf{D}_2$ since $\widehat{\mathbf{D}_2^{-1}} = \hat{\mathbf{D}}_2^{-1}$ (the transformation from $\mathbf{D}_2$ to $\mathbf{D}_2^{-1}$ is one-to-one).

The solution for $\mathbf{D}_2$ therefore has to satisfy

$$\sum_{i=1}^{M}\mathrm{E}_c\left\{\mathrm{Diag}[\hat{\mathbf{V}}_2]\right\} = \sum_{i=1}^{M}\mathrm{E}_c\left\{\mathrm{Diag}\left[(\mathbf{m}_i-\boldsymbol{\mu})(\mathbf{m}_i-\boldsymbol{\mu})'\right]\right\}$$

which, using Proposition 4.1, completes the proof of this proposition. □

Proposition 4.9

The MML estimator for $\mathbf{U}_1$ is given by the expression

$$\hat{\mathbf{U}}_1 = \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{n_i} \mathrm{E}_c\left(\mathbf{d}_{1,ij}\mathbf{d}'_{1,ij}\right)\hat{\mathbf{\Phi}}_1^{-1}\hat{\mathbf{U}}_1. \tag{4.11}$$

Proof

Consider the derivative of $\ell n\ g(\mathbf{b}_i)$ with respect to a typical element of $\mathbf{U}_1$, say $[\mathbf{U}_1]_{k\ell}$ $(k = 1, 2, \ldots, r_1; \ell = 1, 2, \ldots, k)$. Using (4.5) to obtain this derivative yields

$$\frac{\partial \ell n\ g(\mathbf{b}_i)}{\partial [\mathbf{U}_1]_{k\ell}} = \frac{1}{2}\mathrm{tr}\left[\mathbf{\Phi}_1^{-1}\left(\sum_{j=1}^{n_i}\mathbf{d}_{1,ij}\mathbf{d}'_{1,ij} - n_i\mathbf{\Phi}_1\right)\mathbf{\Phi}_1^{-1}\frac{\partial \mathbf{U}_1\mathbf{U}_1'}{\partial [\mathbf{U}_1]_{k\ell}}\right].$$

Evaluating the partial derivative in the above expression, yields

$$\begin{aligned}\frac{\partial \ell n\ g(\mathbf{b}_i)}{\partial [\mathbf{U}_1]_{k\ell}} &= \mathrm{tr}\left[\mathbf{\Phi}_1^{-1}\left(\sum_{j=1}^{n_i}\mathbf{d}_{1,ij}\mathbf{d}'_{1,ij} - n_i\mathbf{\Phi}_1\right)\mathbf{\Phi}_1^{-1}\mathbf{U}_1\mathbf{J}_{\ell k}\right] \\ &= \left[\mathbf{\Phi}_1^{-1}\left(\sum_{j=1}^{n_i}\mathbf{d}_{1,ij}\mathbf{d}'_{1,ij} - n_i\mathbf{\Phi}_1\right)\mathbf{\Phi}_1^{-1}\mathbf{U}_1\right]_{k\ell}.\end{aligned}$$

The above result leads to the conclusion that the $r_1 \times r_1$ lower triangular matrix of derivatives is

$$\frac{\partial \ell n\ g(\mathbf{b}_i)}{\partial \mathbf{U}_1} = \mathbf{\Phi}_1^{-1}\left(\sum_{j=1}^{n_i}\mathbf{d}_{1,ij}\mathbf{d}'_{1,ij} - n_i\mathbf{\Phi}_1\right)\mathbf{\Phi}_1^{-1}\mathbf{U}_1.$$

Substituting this result into (3.6) leads to the following expression for the gradient of $\ell n\ L$ with respect to $\mathbf{U}_1$:

$$\frac{\partial \ell n\ L}{\partial \mathbf{U}_1} = \sum_{i=1}^{M}\mathrm{E}_c\left[\mathbf{\Phi}_1^{-1}\left(\sum_{j=1}^{n_i}\mathbf{d}_{1,ij}\mathbf{d}'_{1,ij} - n_i\mathbf{\Phi}_1\right)\mathbf{\Phi}_1^{-1}\mathbf{U}_1\right].$$

To obtain the maximum likelihood estimator of $\mathbf{U}_2$, this expression is set equal to zero and solved for $\mathbf{U}_1$.

Therefore

$$\sum_{i=1}^{M} \mathrm{E}_c \left[\hat{\mathbf{\Phi}}_1^{-1} \left(\sum_{j=1}^{n_i} \mathbf{d}_{1,ij} \mathbf{d}'_{1,ij} - n_i \hat{\mathbf{\Phi}}_1 \right) \hat{\mathbf{\Phi}}_1^{-1} \hat{\mathbf{U}}_1 \right] = \mathbf{0}$$

or, since $\hat{\mathbf{\Phi}}_1$ is constant with respect to the conditional expected value operator and $\hat{\mathbf{\Phi}}_1^{-1} \neq \mathbf{0}$, it follows that

$$\sum_{i=1}^{M} \sum_{j=1}^{n_i} \mathrm{E}_c \left(\mathbf{d}_{1,ij} \mathbf{d}'_{1,ij} \right) \hat{\mathbf{\Phi}}_1^{-1} \hat{\mathbf{U}}_1 - N \hat{\mathbf{U}}_1 = \mathbf{0}$$

which completes the proof. □

So far, expressions for the MML estimators of the elements of $\boldsymbol{\theta}$ (i.e. the elements of $\boldsymbol{\mu}$, $\mathbf{\Lambda}_2$, $\mathbf{U}_2$, $\mathbf{D}_2$ and $\mathbf{U}_1$) have been obtained. The next two propositions will provide expressions for the maximum likelihood estimators of the elements of $\boldsymbol{\tau}$ (i.e. the elements of $\mathbf{\Lambda}_1$ and $\mathbf{D}_1$) that are contained in $\ell n\, f(\mathbf{y}_i|\mathbf{b}_i)$.

Proposition 4.10

The MML estimator for $\mathbf{\Lambda}_1$ is given by the expression

$$\begin{aligned} \hat{\mathbf{\Lambda}}_1 \;=\; & \left[\sum_{i=1}^{M} \sum_{j=1}^{n_i} \left\{ \mathbf{y}_{ij} \mathrm{E}_c(\mathbf{d}_{1,ij})' - \mathrm{Cov}_c(\mathbf{m}_i, \mathbf{d}'_{1,ij}) - \mathrm{E}_c(\mathbf{m}_i) \mathrm{E}_c(\mathbf{d}_{1,ij})' \right\} \right] \times \\ & \left[\sum_{i=1}^{M} \sum_{j=1}^{n_i} \left\{ \mathrm{Cov}_c(\mathbf{d}_{1,ij}, \mathbf{d}'_{1,ij}) + \mathrm{E}_c(\mathbf{d}_{1,ij}) \mathrm{E}_c(\mathbf{d}_{1,ij})' \right\} \right]^{-1}. \end{aligned} \tag{4.12}$$

Proof

Consider the partial derivative of $\ell n\ f(\mathbf{y}_i|\mathbf{b}_i)$ with respect to a typical element of $\mathbf{\Lambda}_1$, say $[\mathbf{\Lambda}_1]_{k\ell}$ $(k = 1, 2, \ldots, p; \ell = 1, 2, \ldots, r_1)$. Using (4.6) this derivative is (note that $\mathbf{\Lambda}_1$ is contained in $\boldsymbol{\mu}_{\mathbf{y}_{ij}}$)

$$\frac{\partial \ell n\ f(\mathbf{y}_i|\mathbf{b}_i)}{\partial [\mathbf{\Lambda}_1]_{k\ell}} = \sum_{j=1}^{n_i} \text{tr}\left[(\mathbf{y}_{ij} - \boldsymbol{\mu}_{\mathbf{y}_{ij}})' \mathbf{D}_1^{-1} \frac{\partial \boldsymbol{\mu}_{\mathbf{y}_{ij}}}{\partial [\mathbf{\Lambda}_1]_{k\ell}}\right] \tag{4.13}$$

However, $\boldsymbol{\mu}_{\mathbf{y}_{ij}} = \mathbf{m}_i + \mathbf{\Lambda}_1 \mathbf{d}_{1,ij}$, so that

$$\frac{\partial \boldsymbol{\mu}_{\mathbf{y}_{ij}}}{\partial [\mathbf{\Lambda}_1]_{k\ell}} = \frac{\partial \mathbf{\Lambda}_1}{\partial [\mathbf{\Lambda}_1]_{k\ell}} \mathbf{d}_{1,ij} = \mathbf{J}_{k\ell} \mathbf{d}_{1,ij}$$

and substituting this result into (4.13) gives

$$\begin{aligned} \frac{\partial \ell n\ f(\mathbf{y}_i|\mathbf{b}_i)}{\partial [\mathbf{\Lambda}_1]_{k\ell}} &= \sum_{j=i}^{n_i} \text{tr}\left[(\mathbf{y}_{ij} - \boldsymbol{\mu}_{\mathbf{y}_{ij}})' \mathbf{D}_1^{-1} \mathbf{J}_{k\ell} \mathbf{d}_{1,ij}\right] \\ &= \sum_{j=1}^{n_i} \text{tr}\left[\mathbf{D}_1^{-1} (\mathbf{y}_{ij} - \boldsymbol{\mu}_{\mathbf{y}_{ij}}) \mathbf{d}'_{1,ij} \mathbf{J}_{\ell k}\right]. \end{aligned}$$

Therefore, the $p \times r_1$ matrix of derivatives is

$$\frac{\partial \ell n\ f(\mathbf{y}_i|\mathbf{b}_i)}{\partial \mathbf{\Lambda}_1} = \sum_{j=1}^{n_i} \mathbf{D}_1^{-1} (\mathbf{y}_{ij} - \boldsymbol{\mu}_{\mathbf{y}_{ij}}) \mathbf{d}'_{1,ij}.$$

To obtain the gradient of $\ell n\ L$, the above expression needs to be substituted into (3.7), which gives

$$\frac{\partial \ell n\ L}{\partial \mathbf{\Lambda}_1} = \sum_{i=1}^{M} \text{E}_c \left\{ \sum_{j=1}^{n_i} \mathbf{D}_1^{-1} (\mathbf{y}_{ij} - \boldsymbol{\mu}_{\mathbf{y}_{ij}}) \mathbf{d}'_{1,ij} \right\}$$

and setting this expression equal to zero, gives

$$\begin{aligned}\mathbf{0} &= \sum_{i=1}^{M} \mathrm{E}_c \left\{ \sum_{j=1}^{n_i} \hat{\mathbf{D}}_1^{-1} (\mathbf{y}_{ij} - \mathbf{m}_i - \hat{\mathbf{\Lambda}}_1 \mathbf{d}_{1,ij}) \mathbf{d}'_{1,ij} \right\} \\ &= \sum_{i=1}^{M} \sum_{j=1}^{n_i} \left\{ \mathbf{y}_{ij} \mathrm{E}_c(\mathbf{d}'_{1,ij}) - \mathrm{E}_c(\mathbf{m}_i \mathbf{d}'_{1,ij}) - \hat{\mathbf{\Lambda}}_1 \mathrm{E}_c(\mathbf{d}_{1,ij} \mathbf{d}'_{1,ij}) \right\}\end{aligned}$$

so that

$$\hat{\mathbf{\Lambda}}_1 \sum_{i=1}^{M} \sum_{j=1}^{n_i} \mathrm{E}_c(\mathbf{d}_{1,ij} \mathbf{d}'_{1,ij}) = \sum_{i=1}^{M} \sum_{j=1}^{n_i} \left\{ \mathbf{y}_{ij} \mathrm{E}_c(\mathbf{d}'_{1,ij}) - \mathrm{E}_c(\mathbf{m}_i \mathbf{d}'_{1,ij}) \right\}$$

and consequently

$$\hat{\mathbf{\Lambda}}_1 = \left[\sum_{i=1}^{M} \sum_{j=1}^{n_i} \left\{ \mathbf{y}_{ij} \mathrm{E}_c(\mathbf{d}'_{1,ij}) - \mathrm{E}_c(\mathbf{m}_i \mathbf{d}'_{1,ij}) \right\} \right] \left[\sum_{i=1}^{M} \sum_{j=1}^{n_i} \mathrm{E}_c(\mathbf{d}_{1,ij} \mathbf{d}'_{1,ij}) \right]^{-1}.$$

Propositions 4.2 and 4.3 are now used to complete the proof of this proposition. □

Proposition 4.11

The MML estimator for $\mathbf{D}_1$ is given by the expression

$$\begin{aligned}\hat{\mathbf{D}}_1 &= \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\text{Diag}(\mathbf{y}_{ij}\mathbf{y}_{ij}') + \frac{1}{N}\sum_{i=1}^{M} n_i \text{Diag}\left[\text{Cov}_c(\mathbf{m}_i,\mathbf{m}_i') + \text{E}_c(\mathbf{m}_i)\text{E}_c(\mathbf{m}_i)'\right] \\ &+\text{Diag}\left[\boldsymbol{\Lambda}_1\left[\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\left\{\text{Cov}_c(\mathbf{d}_{1,ij},\mathbf{d}_{1,ij}') + \text{E}_c(\mathbf{d}_{1,ij})\text{E}_c(\mathbf{d}_{1,ij})'\right\}\right]\boldsymbol{\Lambda}_1'\right] \\ &-\frac{2}{N}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\text{Diag}\left[\mathbf{y}_{ij}\text{E}_c(\mathbf{m}_i)'\right] - \frac{2}{N}\text{Diag}\left[\left\{\sum_{i=1}^{M}\sum_{j=1}^{n_i}\mathbf{y}_{ij}\text{E}_c(\mathbf{d}_{1,ij})'\right\}\boldsymbol{\Lambda}_1'\right] \\ &+\frac{2}{N}\text{Diag}\left[\sum_{i=1}^{M}\sum_{j=1}^{n_i}\left\{\text{Cov}_c(\mathbf{m}_i,\mathbf{d}_{1,ij}') + \text{E}_c(\mathbf{m}_i)\text{E}_c(\mathbf{d}_{1,ij})'\right\}\boldsymbol{\Lambda}_1'\right]. \qquad (4.14)\end{aligned}$$

Proof

Since $\mathbf{D}_1$ is diagonal, we choose to use the reparameterisation theorem as in Proposition 4.8 where an expression for $\hat{\mathbf{D}}_2$ was obtained.

Therefore let $\mathbf{D}_1^\star = \mathbf{D}_1^{-1}$. The partial derivative of $\ell n\ f(\mathbf{y}_i|\mathbf{b}_i)$ with respect to a typical element in $\mathbf{D}_1^\star$, say $[\mathbf{D}_1^\star]_{kk}$ $(k = 1, 2, \ldots, p)$, is given by

$$\begin{aligned}\frac{\partial \ell n\ f(\mathbf{y}_i|\mathbf{b}_i)}{\partial[\mathbf{D}_1^\star]_{kk}} &= -\frac{1}{2}\sum_{j=1}^{n_i}\text{tr}\left[\left\{(\mathbf{y}_{ij}-\boldsymbol{\mu}_{\mathbf{y}_{ij}})(\mathbf{y}_{ij}-\boldsymbol{\mu}_{\mathbf{y}_{ij}})' - \mathbf{D}_1\right\}\frac{\partial \mathbf{D}_1^\star}{\partial[\mathbf{D}_1^\star]_{kk}}\right] \\ &= -\frac{1}{2}\sum_{j=1}^{n_i}\text{tr}\left[\left\{(\mathbf{y}_{ij}-\boldsymbol{\mu}_{\mathbf{y}_{ij}})(\mathbf{y}_{ij}-\boldsymbol{\mu}_{\mathbf{y}_{ij}})' - \mathbf{D}_1\right\}\mathbf{J}_{kk}\right].\end{aligned}$$

The $p \times p$ diagonal matrix of derivatives is therefore

$$\frac{\partial \ell n\ f(\mathbf{y}_i|\mathbf{b}_i)}{\partial \mathbf{D}_1^\star} = -\frac{1}{2}\sum_{j=1}^{n_i} \text{Diag}\left[(\mathbf{y}_{ij} - \boldsymbol{\mu}_{\mathbf{y}_{ij}})(\mathbf{y}_{ij} - \boldsymbol{\mu}_{\mathbf{y}_{ij}})' - \mathbf{D}_1\right]$$

and substitution of this result into (3.7) gives

$$\frac{\partial \ell n\ L}{\partial \mathbf{D}_1^\star} = \sum_{i=1}^{M} \text{E}_c\left\{-\frac{1}{2}\sum_{j=1}^{n_i} \text{Diag}\left[(\mathbf{y}_{ij} - \boldsymbol{\mu}_{\mathbf{y}_{ij}})(\mathbf{y}_{ij} - \boldsymbol{\mu}_{\mathbf{y}_{ij}})' - \mathbf{D}_1\right]\right\}.$$

To obtain the maximum likelihood estimator for $\mathbf{D}_1^\star$ we now have to set the above expression equal to zero and solve for $\mathbf{D}_1^\star$ (or for $\mathbf{D}_1$).

Therefore

$$-\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{n_i} \text{E}_c\left\{\text{Diag}\left[(\mathbf{y}_{ij} - \boldsymbol{\mu}_{\mathbf{y}_{ij}})(\mathbf{y}_{ij} - \boldsymbol{\mu}_{\mathbf{y}_{ij}})' - \hat{\mathbf{D}}_1\right]\right\} = \mathbf{0}$$

and consequently

$$\sum_{i=1}^{M}\sum_{j=1}^{n_i} \text{E}_c\left\{\hat{\mathbf{D}}_1\right\} = \sum_{i=1}^{M}\sum_{j=1}^{n_i} \text{Diag}\left[\text{E}_c\left\{(\mathbf{y}_{ij} - \boldsymbol{\mu}_{\mathbf{y}_{ij}})(\mathbf{y}_{ij} - \boldsymbol{\mu}_{\mathbf{y}_{ij}})'\right\}\right]$$

which leads to

$$\hat{\mathbf{D}}_1 = \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{n_i} \text{Diag}\left[\text{E}_c\left\{(\mathbf{y}_{ij} - \boldsymbol{\mu}_{\mathbf{y}_{ij}})(\mathbf{y}_{ij} - \boldsymbol{\mu}_{\mathbf{y}_{ij}})'\right\}\right].$$

However, $\boldsymbol{\mu}_{\mathbf{y}_{ij}} = \mathbf{m}_i + \boldsymbol{\Lambda}_1\mathbf{d}_{1,ij}$, so that

$$(\mathbf{y}_{ij} - \boldsymbol{\mu}_{\mathbf{y}_{ij}})(\mathbf{y}_{ij} - \boldsymbol{\mu}_{\mathbf{y}_{ij}})' \;=\; (\mathbf{y}_{ij} - \mathbf{m}_i - \boldsymbol{\Lambda}_1\mathbf{d}_{1,ij})(\mathbf{y}_{ij} - \mathbf{m}_i - \boldsymbol{\Lambda}_1\mathbf{d}_{1,ij})'$$

$$= \mathbf{y}_{ij}\mathbf{y}_{ij}' + \mathbf{m}_i\mathbf{m}_i' + \boldsymbol{\Lambda}_1\mathbf{d}_{1,ij}\mathbf{d}_{1,ij}'\boldsymbol{\Lambda}_1' - \mathbf{y}_{ij}\mathbf{m}_i' - \mathbf{y}_{ij}\mathbf{d}_{1,ij}'\boldsymbol{\Lambda}_1'$$
$$-\mathbf{m}_i\mathbf{y}_{ij}' + \mathbf{m}_i\mathbf{d}_{1,ij}'\boldsymbol{\Lambda}_1' - \boldsymbol{\Lambda}_1\mathbf{d}_{1,ij}\mathbf{y}_{ij}' + \boldsymbol{\Lambda}_1\mathbf{d}_{1,ij}\mathbf{m}_i',$$

with conditional expected value

$$\mathrm{E}_c(\mathbf{y}_{ij} - \boldsymbol{\mu}_{\mathbf{y}_{ij}})(\mathbf{y}_{ij} - \boldsymbol{\mu}_{\mathbf{y}_{ij}})' = \mathbf{y}_{ij}\mathbf{y}_{ij}' + \mathrm{E}_c(\mathbf{m}_i\mathbf{m}_i') + \boldsymbol{\Lambda}_1\mathrm{E}_c(\mathbf{d}_{1,ij}\mathbf{d}_{1,ij}')\boldsymbol{\Lambda}_1'$$

$$-\mathbf{y}_{ij}\mathrm{E}_c(\mathbf{m}_i)' - \mathbf{y}_{ij}\mathrm{E}_c(\mathbf{d}_{1,ij})'\boldsymbol{\Lambda}_1' - \mathrm{E}_c(\mathbf{m}_i)\mathbf{y}_{ij}'$$

$$+\mathrm{E}_c(\mathbf{m}_i\mathbf{d}_{1,ij}')\boldsymbol{\Lambda}_1' - \boldsymbol{\Lambda}_1\mathrm{E}_c(\mathbf{d}_{1,ij})\mathbf{y}_{ij}' + \boldsymbol{\Lambda}_1\mathrm{E}_c(\mathbf{d}_{1,ij}\mathbf{m}_i').$$

The use of Propositions 4.2, 4.3 and 4.4 completes the proof of this proposition. □

It is evident from the estimators of the unknown parameters $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}_2$, $\mathbf{U}_2$, $\mathbf{D}_2$, $\boldsymbol{\Lambda}_1$, $\mathbf{U}_1$ and $\mathbf{D}_1$, that the conditional distribution of $\mathbf{b}_i$ given $\mathbf{y}_i$ plays an important role in these parameter estimators, since they are completely determined by the moments of this distribution.

From well-known normal distribution theory, the moments of the conditional distribution function $p(\mathbf{b}_i|\mathbf{y}_i)$ can be obtained from the joint density $h(\mathbf{b}_i, \mathbf{y}_i)$ - see e.g. Morrison (1990). The moments of the joint distribution function of $\mathbf{b}_i$ and $\mathbf{y}_i$ will therefore now be given.

From the column vector that contains $\mathbf{b}_i$ and $\mathbf{y}_i$ in the following way, namely

$$\begin{pmatrix} \mathbf{b}_i \\ \mathbf{y}_i \end{pmatrix} = \begin{pmatrix} \mathbf{m}_i \\ \mathbf{d}_{1,i} \\ \mathbf{y}_i \end{pmatrix}$$

the expected value is obtained as

$$\mathrm{E}\begin{pmatrix} \mathbf{m}_i \\ \mathbf{d}_{1,i} \\ \mathbf{y}_i \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{0} \\ \mathbf{j}_{n_i} \otimes \boldsymbol{\mu} \end{pmatrix} \tag{4.15}$$

and the covariance matrix as

$$\mathrm{Cov}\left[\begin{pmatrix} \mathbf{m}_i \\ \mathbf{d}_{1,i} \\ \mathbf{y}_i \end{pmatrix}, (\mathbf{m}_i' \; \mathbf{d}_{1,i}' \; \mathbf{y}_i')\right] = \begin{pmatrix} \mathrm{Cov}(\mathbf{m}_i, \mathbf{m}_i') & \mathrm{Cov}(\mathbf{m}_i, \mathbf{d}_{1,i}') & \mathrm{Cov}(\mathbf{m}_i, \mathbf{y}_i') \\ \mathrm{Cov}(\mathbf{d}_{1,i}, \mathbf{m}_i') & \mathrm{Cov}(\mathbf{d}_{1,i}, \mathbf{d}_{1,i}') & \mathrm{Cov}(\mathbf{d}_{1,i}, \mathbf{y}_i') \\ \mathrm{Cov}(\mathbf{y}_i, \mathbf{m}_i') & \mathrm{Cov}(\mathbf{y}_i, \mathbf{d}_{1,i}') & \mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_i') \end{pmatrix}$$

In this matrix, the six non-duplicated submatrices are

i.

$$\begin{aligned} \mathrm{Cov}(\mathbf{m}_i, \mathbf{m}_i') &= \boldsymbol{\Lambda}_2 \boldsymbol{\Phi}_2 \boldsymbol{\Lambda}_2' + \mathbf{D}_2 \\ &= \mathbf{V}_2 \end{aligned}$$

ii.

$$\mathrm{Cov}(\mathbf{d}_{1,i}, \mathbf{d}_{1,i}') = \mathbf{I}_{n_i} \otimes \boldsymbol{\Phi}_1$$

iii.

$$\mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_i') = \mathbf{W}_i$$

iv.

$$\text{Cov}(\mathbf{d}_{1,i}, \mathbf{m}_i') = \mathbf{0}$$

v.

$$\begin{aligned}\text{Cov}(\mathbf{y}_i, \mathbf{m}_i') &= \text{Cov}\left[(\mathbf{j}_{n_i} \otimes \mathbf{m}_i + (\mathbf{I}_{n_i} \otimes \mathbf{\Lambda}_1)\mathbf{d}_{1,i} + \mathbf{e}_{1,i}), \mathbf{m}_i'\right] \\ &= \mathbf{j}_{n_i} \otimes \mathbf{V}_2\end{aligned}$$

vi.

$$\begin{aligned}\text{Cov}(\mathbf{y}_i, \mathbf{d}_{1,i}') &= \text{Cov}\left[(\mathbf{j}_{n_i} \otimes \mathbf{m}_i + (\mathbf{I}_{n_i} \otimes \mathbf{\Lambda}_1)\mathbf{d}_{1,i} + \mathbf{e}_{1,i}), \mathbf{d}_{1,i}'\right] \\ &= (\mathbf{I}_{n_i} \otimes \mathbf{\Lambda}_1)(\mathbf{I}_{n_i} \otimes \mathbf{\Phi}_1) \\ &= \mathbf{I}_{n_i} \otimes \mathbf{\Lambda}_1\mathbf{\Phi}_1\end{aligned}$$

From the above it follows that the covariance matrix of $(\mathbf{b}_i' \; \mathbf{y}_i')$ is given by

$$\text{Cov}\left[\begin{pmatrix} \mathbf{m}_i \\ \mathbf{d}_{1,i} \\ \mathbf{y}_i \end{pmatrix}, (\mathbf{m}_i' \; \mathbf{d}_{1,i}' \; \mathbf{y}_i')\right] = \begin{pmatrix} \mathbf{V}_2 & \mathbf{0} & (\mathbf{j}_{n_i} \otimes \mathbf{V}_2)' \\ \mathbf{0} & \mathbf{I}_{n_i} \otimes \mathbf{\Phi}_1 & (\mathbf{I}_{n_i} \otimes \mathbf{\Lambda}_1\mathbf{\Phi}_1)' \\ \mathbf{j}_{n_i} \otimes \mathbf{V}_2 & \mathbf{I}_{n_i} \otimes \mathbf{\Lambda}_1\mathbf{\Phi}_1 & \mathbf{W}_i \end{pmatrix} \tag{4.16}$$

so that the moments of the joint distribution of $\mathbf{b}_i$ and $\mathbf{y}_i$ are given by (4.15) and (4.16).

Consequently the moments of the conditional distribution of $\mathbf{b}_i$ given $\mathbf{y}_i$ is

$$\text{E}_c(\mathbf{b}_i) = \text{E}(\mathbf{b}_i) + \text{Cov}(\mathbf{b}_i, \mathbf{y}_i') \left[\text{Cov}(\mathbf{y}_i, \mathbf{y}_i')\right]^{-1} (\mathbf{y}_i - \text{E}(\mathbf{y}_i))$$

$$= \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} (\mathbf{j}_{n_i} \otimes \mathbf{V}_2)' \\ (\mathbf{I}_{n_i} \otimes \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1)' \end{pmatrix} \mathbf{W}_i^{-1} (\mathbf{y}_i - \mathbf{j}_{n_i} \otimes \boldsymbol{\mu}) \tag{4.17}$$

and

$$\begin{aligned} & \mathrm{Cov_c}(\mathbf{b}_i, \mathbf{b}_i') \\ = \; & \mathrm{Cov}(\mathbf{b}_i, \mathbf{b}_i') - \mathrm{Cov}(\mathbf{b}_i, \mathbf{y}_i') \left[\mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_i')\right]^{-1} \mathrm{Cov}(\mathbf{y}_i, \mathbf{b}_i') \\ = \; & \begin{pmatrix} \mathbf{V}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n_i} \otimes \boldsymbol{\Phi}_1 \end{pmatrix} - \begin{pmatrix} (\mathbf{j}_{n_i} \otimes \mathbf{V}_2)' \\ (\mathbf{I}_{n_i} \otimes \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1)' \end{pmatrix} \mathbf{W}_i^{-1} (\mathbf{j}_n \otimes \mathbf{V}_2 \;\; \mathbf{I}_{n_i} \otimes \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1). \end{aligned} \tag{4.18}$$

Using the above two results, it is now possible to get expressions for $\mathrm{E_c}(\mathbf{m}_i)$, $\mathrm{E_c}(\mathbf{d}_{1,i})$, $\mathrm{Cov_c}(\mathbf{m}_i, \mathbf{m}_i')$, $\mathrm{Cov_c}(\mathbf{d}_{1,i}, \mathbf{d}_{1,i}')$ and $\mathrm{Cov_c}(\mathbf{m}_i, \mathbf{d}_{1,i}')$ which are necessary in determining the maximum likelihood estimators of the unknown parameters.

From (4.17) it follows that

$$\mathrm{E_c}(\mathbf{m}_i) = \boldsymbol{\mu} + (\mathbf{j}_{n_i} \otimes \mathbf{V}_2)' \mathbf{W}_i^{-1} (\mathbf{y}_i - \mathbf{j}_{n_i} \otimes \boldsymbol{\mu}) \tag{4.19}$$

and

$$\mathrm{E_c}(\mathbf{d}_{1,i}) = (\mathbf{I}_{n_i} \otimes \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1)' \mathbf{W}_i^{-1} (\mathbf{y}_i - \mathbf{j}_{n_i} \otimes \boldsymbol{\mu})$$

of which the latter can be used to obtain an expression for the conditional expected value for the j-th individual, namely

$$\mathrm{E}_c(\mathbf{d}_{1,ij}) = (\mathbf{J}_{j1} \otimes \mathbf{\Lambda}_1\mathbf{\Phi}_1)'\mathbf{W}_i^{-1}(\mathbf{y}_i - \mathbf{j}_{n_i} \otimes \boldsymbol{\mu}). \tag{4.20}$$

From (4.18) it follows that

$$\mathrm{Cov}_c(\mathbf{m}_i, \mathbf{m}_i') = \mathbf{V}_2 - (\mathbf{j}_{n_i} \otimes \mathbf{V}_2)'\mathbf{W}_i^{-1}(\mathbf{j}_{n_i} \otimes \mathbf{V}_2), \tag{4.21}$$

$$\mathrm{Cov}_c(\mathbf{d}_{1,i}, \mathbf{d}_{1,i}') = \mathbf{I}_{n_i} \otimes \mathbf{\Phi}_1 - (\mathbf{I}_{n_i} \otimes \mathbf{\Lambda}_1\mathbf{\Phi}_1)'\mathbf{W}_i^{-1}(\mathbf{I}_{n_i} \otimes \mathbf{\Lambda}_1\mathbf{\Phi}_1)$$

and

$$\mathrm{Cov}_c(\mathbf{m}_i, \mathbf{d}_{1,i}') = -(\mathbf{j}_{n_i} \otimes \mathbf{V}_2)'\mathbf{W}_i^{-1}(\mathbf{I}_{n_i} \otimes \mathbf{\Lambda}_1\mathbf{\Phi}_1).$$

The latter two expressions can be used to obtain the conditional covariances for the j-th individual, namely

$$\mathrm{Cov}_c(\mathbf{d}_{1,ij}, \mathbf{d}_{1,ij}') = \mathbf{\Phi}_1 - (\mathbf{J}_{j1} \otimes \mathbf{\Lambda}_1\mathbf{\Phi}_1)'\mathbf{W}_i^{-1}(\mathbf{J}_{j1} \otimes \mathbf{\Lambda}_1\mathbf{\Phi}_1) \tag{4.22}$$

and

$$\mathrm{Cov}_c(\mathbf{m}_i, \mathbf{d}_{1,ij}') = -(\mathbf{j}_{n_i} \otimes \mathbf{V}_2)'\mathbf{W}_i^{-1}(\mathbf{J}_{j1} \otimes \mathbf{\Lambda}_1\mathbf{\Phi}_1). \tag{4.23}$$

In expressions (4.19) to (4.23), $\mathbf{W}_i$ appears only in its inverse form. Since this matrix is quite large ($pn_i \times pn_i$), use can be made of its partitioning to use a standard result - see e.g. Browne (1991) - to write $\mathbf{W}_i^{-1}$ also as a partitioned matrix, and thus to express (4.19) to (4.23) in computationally more efficient forms.

Recall that $\mathbf{W}_i = \mathbf{I}_{n_i} \otimes \mathbf{V}_1 + \mathbf{j}_{n_i}\mathbf{j}'_{n_i} \otimes \mathbf{V}_2$ and note that this is equivalent to

$$\mathbf{W}_i = \mathbf{I}_{n_i} \otimes \mathbf{V}_1 + (\mathbf{j}_{n_i} \otimes \mathbf{I}_p)\mathbf{V}_2(\mathbf{j}_{n_i} \otimes \mathbf{I}_p)'.$$

This expression is of the form $\mathbf{A} + \mathbf{BCB}'$ where

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

Applying this identity to $\mathbf{W}_i^{-1}$, noting that

$$\mathbf{A} = \mathbf{I}_{n_i} \otimes \mathbf{V}_1,$$

$$\mathbf{B} = \mathbf{j}_{n_i} \otimes \mathbf{I}_p$$

and

$$\mathbf{C} = \mathbf{V}_2,$$

it follows that

$$\begin{aligned} \mathbf{W}_i^{-1} &= \mathbf{I}_{n_i} \otimes \mathbf{V}_1^{-1} - \mathbf{j}_{n_i}\mathbf{j}'_{n_i} \otimes \mathbf{C}_i \\ &= \begin{pmatrix} \mathbf{S}_i & -\mathbf{C}_i & \cdots & -\mathbf{C}_i \\ -\mathbf{C}_i & \mathbf{S}_i & \cdots & -\mathbf{C}_i \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{C}_i & -\mathbf{C}_i & \cdots & \mathbf{S}_i \end{pmatrix} \end{aligned} \tag{4.24}$$

where

$$\mathbf{S}_i = \mathbf{V}_1^{-1} - \mathbf{C}_i$$

and

$$\mathbf{C}_i = \mathbf{V}_1^{-1}(\mathbf{V}_2^{-1} + n_i\mathbf{V}_1^{-1})^{-1}\mathbf{V}_1^{-1}.$$

This expression for $\mathbf{W}_i^{-1}$ requires only the inversion of $p \times p$ matrices. However, it will now be shown that (4.24) can be rewritten so that only matrices of order $r_1 \times r_1$ and $r_2 \times r_2$ need be inverted. Since

$$\mathbf{V}_1^{-1} = (\mathbf{D}_1 + \mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{\Lambda}_1')^{-1},$$

it follows that

$$\mathbf{V}_1^{-1} = \mathbf{D}_1^{-1} - \mathbf{D}_1^{-1}\mathbf{\Lambda}_1(\mathbf{\Phi}_1^{-1} + \mathbf{\Lambda}_1'\mathbf{D}_1^{-1}\mathbf{\Lambda}_1)^{-1}\mathbf{\Lambda}_1'\mathbf{D}_1^{-1}$$

so that the only inversions necessary are those of an $r_1 \times r_1$ matrix and of a diagonal matrix. In (4.24), $\mathbf{C}_i$ can be equivalently written as

$$\begin{aligned}
\left[\mathbf{V}_1(\mathbf{V}_2^{-1} + n_i\mathbf{V}_1^{-1})\mathbf{V}_1\right]^{-1} &= (\mathbf{V}_1\mathbf{V}_2^{-1}\mathbf{V}_1 + n_i\mathbf{V}_1)^{-1} \\
&= (\mathbf{V}_1\mathbf{V}_2^{-1}\mathbf{V}_1 + n_i\mathbf{V}_1\mathbf{V}_2^{-1}\mathbf{V}_2)^{-1} \\
&= \left[\mathbf{V}_1\mathbf{V}_2^{-1}(\mathbf{V}_1 + n_i\mathbf{V}_2)\right]^{-1} \\
&= (\mathbf{V}_1 + n_i\mathbf{V}_2)^{-1}\mathbf{V}_2\mathbf{V}_1^{-1}.
\end{aligned}$$

It will now be shown how $(\mathbf{V}_1 + n_i\mathbf{V}_2)^{-1}$, which is a $p \times p$ matrix, can be rewritten so that it will be necessary to invert only $r_1 \times r_1$ and $r_2 \times r_2$ matrices.

Use the structure of $\mathbf{V}_2$ and write

$$
\begin{aligned}
(\mathbf{V}_1 + n_i\mathbf{V}_2)^{-1} &= [\mathbf{V}_1 + n_i(\mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{\Lambda}_2' + \mathbf{D}_2)]^{-1} \\
&= [(\mathbf{V}_1 + n_i\mathbf{D}_2) + n_i\mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{\Lambda}_2']^{-1}
\end{aligned}
$$

Let $\tilde{\mathbf{V}} = \mathbf{V}_1 + n_i\mathbf{D}_2$, then

$$
(\mathbf{V}_1 + n_i\mathbf{V}_2)^{-1} = \tilde{\mathbf{V}}^{-1} - \tilde{\mathbf{V}}^{-1}\mathbf{\Lambda}_2(n_i^{-1}\mathbf{\Phi}_2^{-1} + \mathbf{\Lambda}_2'\tilde{\mathbf{V}}^{-1}\mathbf{\Lambda}_2)^{-1}\mathbf{\Lambda}_2'\tilde{\mathbf{V}}^{-1}
$$

while the inverse of $\tilde{\mathbf{V}}$ can be written as (using the structure of $\mathbf{V}_1$ and Khatri's result)

$$
\begin{aligned}
(\mathbf{V}_1 + n_i\mathbf{D}_2)^{-1} &= [(n_i\mathbf{D}_2 + \mathbf{D}_1) + \mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{\Lambda}_1']^{-1} \\
&= \tilde{\mathbf{D}}^{-1} - \tilde{\mathbf{D}}^{-1}\mathbf{\Lambda}_1(\mathbf{\Phi}_1^{-1} + \mathbf{\Lambda}_1'\tilde{\mathbf{D}}^{-1}\mathbf{\Lambda}_1)^{-1}\mathbf{\Lambda}_1'\tilde{\mathbf{D}}^{-1}
\end{aligned}
$$

where $\tilde{\mathbf{D}} = n_i\mathbf{D}_2 + \mathbf{D}_1$.

Consequently $\mathbf{W}_i^{-1}$, of order $pn_i \times pn_i$ can be obtained by inverting $r_1 \times r_1$ and $r_2 \times r_2$ matrices. This is a useful result since, in general, $r_1 << pn_i$ and $r_2 << pn_i$.

Expression (4.24) can now be applied to rewrite (4.19) to (4.23) in computationally more efficient expressions. Doing this, (4.19) becomes

$$
\mathrm{E}_c(\mathbf{m}_i) = \boldsymbol{\mu} + (\mathbf{j}_{n_i} \otimes \mathbf{V}_2)'\mathbf{W}_i^{-1}(\mathbf{y}_i - \mathbf{j}_{n_i} \otimes \boldsymbol{\mu})
$$

$$
\begin{aligned}
&= \boldsymbol{\mu} + (\mathbf{V}_2\ \mathbf{V}_2 \cdots \mathbf{V}_2)
\begin{pmatrix}
\mathbf{S}_i & -\mathbf{C}_i & \cdots & -\mathbf{C}_i \\
-\mathbf{C}_i & \mathbf{S}_i & \cdots & -\mathbf{C}_i \\
\vdots & \vdots & \ddots & \vdots \\
-\mathbf{C}_i & -\mathbf{C}_i & \cdots & \mathbf{S}_i
\end{pmatrix}
\begin{pmatrix}
\mathbf{y}_{i1} - \boldsymbol{\mu} \\
\mathbf{y}_{i2} - \boldsymbol{\mu} \\
\vdots \\
\mathbf{y}_{in_i} - \boldsymbol{\mu}
\end{pmatrix} \\
&= \boldsymbol{\mu} + \mathbf{V}_2(\mathbf{V}_1^{-1} - n_i\mathbf{C}_i)(\sum_{j=1}^{n_i} \mathbf{y}_{ij} - n_i\boldsymbol{\mu}) \qquad (4.25)
\end{aligned}
$$

while (4.20) becomes

$$
\begin{aligned}
\mathrm{E}_c(\mathbf{d}_{1,ij}) &= (\mathbf{J}_{j1} \otimes \boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)'\mathbf{W}_i^{-1}(\mathbf{y}_i - \mathbf{j}_{n_i} \otimes \boldsymbol{\mu}) \\
&= (\mathbf{0} \cdots \boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1' \cdots \mathbf{0})
\begin{pmatrix}
\mathbf{S}_i & \cdots & -\mathbf{C}_i \\
\vdots & \ddots & \vdots \\
-\mathbf{C}_i & \cdots & \mathbf{S}_i
\end{pmatrix}
\begin{pmatrix}
\mathbf{y}_{i1} - \boldsymbol{\mu} \\
\vdots \\
\mathbf{y}_{in_i} - \boldsymbol{\mu}
\end{pmatrix} \\
&= (-\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1'\mathbf{C}_i \cdots \boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1'\mathbf{S}_i \cdots -\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1'\mathbf{C}_i)
\begin{pmatrix}
\mathbf{y}_{i1} - \boldsymbol{\mu} \\
\vdots \\
\mathbf{y}_{in_i} - \boldsymbol{\mu}
\end{pmatrix} \\
&= \boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1' \left[\mathbf{V}_1^{-1}(\mathbf{y}_{ij} - \boldsymbol{\mu}) - \mathbf{C}_i(\sum_{j=1}^{n_i} \mathbf{y}_{ij} - n_i\boldsymbol{\mu})\right]. \qquad (4.26)
\end{aligned}
$$

Substituting the expression for $\mathbf{W}_i^{-1}$ in (4.24) into the expressions for the conditional covariance matrices given by (4.21) to (4.23), it follows that

$$
\mathrm{Cov}_c(\mathbf{m}_i, \mathbf{m}_i') = \mathbf{V}_2 - (\mathbf{j}_{n_i} \otimes \mathbf{V}_2)'\mathbf{W}_i^{-1}(\mathbf{j}_{n_i} \otimes \mathbf{V}_2)
$$

$$= \mathbf{V}_2 - (\mathbf{V}_2\ \mathbf{V}_2 \cdots \mathbf{V}_2) \begin{pmatrix} \mathbf{S}_i & \cdots & -\mathbf{C}_i \\ \vdots & \ddots & \vdots \\ -\mathbf{C}_i & \cdots & \mathbf{S}_i \end{pmatrix} \begin{pmatrix} \mathbf{V}_2 \\ \mathbf{V}_2 \\ \vdots \\ \mathbf{V}_2 \end{pmatrix}$$

$$= \mathbf{V}_2 - n_i \mathbf{V}_2 (\mathbf{V}_1^{-1} - n_i \mathbf{C}_i) \mathbf{V}_2 \tag{4.27}$$

and

$$\begin{aligned} \mathrm{Cov_c}(\mathbf{d}_{1,ij}, \mathbf{d}'_{1,ij}) &= \mathbf{\Phi}_1 - (\mathbf{J}_{j1} \otimes \mathbf{\Lambda}_1 \mathbf{\Phi}_1)' \mathbf{W}_i^{-1} (\mathbf{J}_{j1} \otimes \mathbf{\Lambda}_1 \mathbf{\Phi}_1) \\ &= \mathbf{\Phi}_1 - (\mathbf{0} \cdots \mathbf{\Phi}_1 \mathbf{\Lambda}_1' \cdots \mathbf{0}) \begin{pmatrix} \mathbf{S}_i & \cdots & -\mathbf{C}_i \\ \vdots & \ddots & \vdots \\ -\mathbf{C}_i & \cdots & \mathbf{S}_i \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{\Lambda}_1 \mathbf{\Phi}_1 \\ \vdots \\ \mathbf{0} \end{pmatrix} \\ &= \mathbf{\Phi}_1 - \mathbf{\Phi}_1 \mathbf{\Lambda}_1' \mathbf{S}_i \mathbf{\Lambda}_1 \mathbf{\Phi}_1. \end{aligned} \tag{4.28}$$

Finally,

$$\begin{aligned} \mathrm{Cov_c}(\mathbf{m}_i, \mathbf{d}'_{1,ij}) &= -(\mathbf{j}_{n_i} \otimes \mathbf{V}_2)' \mathbf{W}_i^{-1} (\mathbf{J}_{j1} \otimes \mathbf{\Lambda}_1 \mathbf{\Phi}_1) \\ &= -(\mathbf{V}_2\ \mathbf{V}_2 \cdots \mathbf{V}_2) \begin{pmatrix} \mathbf{S}_i & \cdots & -\mathbf{C}_i \\ \vdots & \ddots & \vdots \\ -\mathbf{C}_i & \cdots & \mathbf{S}_i \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{\Lambda}_1 \mathbf{\Phi}_1 \\ \vdots \\ \mathbf{0} \end{pmatrix} \\ &= -\mathbf{V}_2 (\mathbf{V}_1^{-1} - n_i \mathbf{C}_i) \mathbf{\Lambda}_1 \mathbf{\Phi}_1. \end{aligned} \tag{4.29}$$

The expressions (4.25) through (4.29) now give computationally efficient formulae for

the moments of $p(\mathbf{b}_i|\mathbf{y}_i)$ that are of primary importance in the EM-algorithm.

4.7 Constrained estimation in the exploratory case

In an exploratory analysis one could remove the $r_1^2 + r_2^2$ indeterminacies in the bilevel model by assuming orthogonal factors with unit variances on both levels (i.e. having $\mathbf{\Phi}_1 = \mathbf{I}$ and $\mathbf{\Phi}_2 = \mathbf{I}$), and also require that $\mathbf{\Lambda}_1'\mathbf{D}_1^{-1}\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2'\mathbf{D}_2^{-1}\mathbf{\Lambda}_2$ are diagonal matrices. The structures imposed on the individual-level and group-level covariance matrices in this case are therefore $\mathbf{V}_1 = \mathbf{\Lambda}_1\mathbf{\Lambda}_1' + \mathbf{D}_1$ and $\mathbf{V}_2 = \mathbf{\Lambda}_2\mathbf{\Lambda}_2' + \mathbf{D}_2$. The MML estimators of $\boldsymbol{\mu}$, $\mathbf{D}_1$ and $\mathbf{D}_2$ in this orthogonal model may be obtained using Propositions 4.5, 4.8 and 4.11, where $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$ are replaced by identity matrices. Since constraints are imposed on the elements of $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$, Propositions 4.6 and 4.10 can not be used to obtain the MML estimators of these two matrices, but need some adjustment. This will now be considered.

The constraints imposed on the elements of $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ are non-linear. The next proposition will show how such non-linear constraints may be linearised, and subsequently, the MML estimators of $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ will be obtained subject to the (approximately) linear constraints, using the Lagrange multiplier method.

Proposition 4.12

Suppose $\mathbf{\Lambda}$ is a general $p \times r$ matrix of factor loadings, and $\mathbf{D}$ is a $p \times p$ diagonal matrix with $d_{ii} > 0 \quad (i = 1, 2, \ldots, p)$. Suppose further that the elements of $\mathbf{\Lambda}$ are required to satisfy the $r(r-1)/2$ equality constraints given by $\mathbf{\Lambda}'\mathbf{D}^{-1}\mathbf{\Lambda}$ = diagonal, or

$$\mathbf{c}(\mathbf{\Lambda}) = \text{vecs}^{\star}(\mathbf{\Lambda}'\mathbf{D}^{-1}\mathbf{\Lambda}) = \mathbf{0}.$$

These non-linear equality constraints may be approximated by the set of linear equality constraints given by

$$\mathbf{L}\text{vec}(\mathbf{\Lambda}) = \text{vecs}^{\star}(\mathbf{\Lambda}_0'\mathbf{D}^{-1}\mathbf{\Lambda}_0)$$

where $\mathbf{L}$ is the matrix of first order partial derivatives of $\mathbf{c}(\mathbf{\Lambda})$ with respect to the elements of $\mathbf{\Lambda}$ at the point where $\mathbf{\Lambda} = \mathbf{\Lambda}_0$ and where $\mathbf{\Lambda}_0$ is some arbitrary known $p \times r$ matrix.

Proof

The $r(r-1)/2$ vector valued function $\mathbf{c}$ may be approximated by a first order Taylor function. This approximation is given by

$$\mathbf{c}(\mathbf{\Lambda}) \approx \mathbf{c}(\mathbf{\Lambda}_0) + \mathbf{L}\text{vec}(\mathbf{\Lambda} - \mathbf{\Lambda}_0)$$

where $\mathbf{\Lambda}_0$ is some known matrix and $\mathbf{L}$ is the $r(r-1)/2 \times pr$ known matrix of first order partial derivatives of $\mathbf{c}(\mathbf{\Lambda})$ with respect to the elements of $\mathbf{\Lambda}$ at the point where $\mathbf{\Lambda} = \mathbf{\Lambda}_0$. Consequently the matrix $\mathbf{L}$ is given by the expression

$$\mathbf{L} = \left.\frac{\partial \mathbf{c}(\mathbf{\Lambda})}{\partial \text{vec}'(\mathbf{\Lambda})}\right|_{\mathbf{\Lambda}=\mathbf{\Lambda}_0}.$$

To derive an expression for this matrix of derivatives, first consider a typical element of $\mathbf{\Lambda}$, say $[\mathbf{\Lambda}]_{k\ell}$ $(k = 1, 2, \ldots, p; \ell = 1, 2, \ldots, r_1)$. Then the column of $\mathbf{L}$ containing the derivative of $\mathbf{c}(\mathbf{\Lambda})$ with respect to $[\mathbf{\Lambda}]_{k\ell}$ is

$$\begin{aligned}
\frac{\partial \mathbf{c}(\mathbf{\Lambda})}{\partial [\mathbf{\Lambda}]_{k\ell}} &= \frac{\partial \text{vecs}^{\star}(\mathbf{\Lambda}'\mathbf{D}^{-1}\mathbf{\Lambda})}{\partial [\mathbf{\Lambda}]_{k\ell}} \\
&= \text{vecs}^{\star}(\mathbf{J}_{\ell k}\mathbf{D}^{-1}\mathbf{\Lambda} + \mathbf{\Lambda}'\mathbf{D}^{-1}\mathbf{J}_{k\ell}) \\
&= [\mathbf{J}_{\ell k}\mathbf{D}^{-1}\mathbf{\Lambda} + \mathbf{\Lambda}'\mathbf{D}^{-1}\mathbf{J}_{k\ell}]_{ij}\,, \quad i = 2, 3, \ldots, r;\ j = 1, 2, \ldots, i-1 \\
&= \text{tr}[\mathbf{J}_{ji}\mathbf{J}_{\ell k}\mathbf{D}^{-1}\mathbf{\Lambda} + \mathbf{J}_{ji}\mathbf{\Lambda}'\mathbf{D}^{-1}\mathbf{J}_{k\ell}]\,, \quad i = 2, 3, \ldots, r;\ j = 1, 2, \ldots, i-1 \\
&= \text{tr}[\mathbf{J}_{ij}\mathbf{\Lambda}'\mathbf{D}^{-1}\mathbf{J}_{k\ell} + \mathbf{J}_{ji}\mathbf{\Lambda}'\mathbf{D}^{-1}\mathbf{J}_{k\ell}]\,, \quad i = 2, 3, \ldots, r;\ j = 1, 2, \ldots, i-1 \\
&= [\mathbf{I}_r]_{\ell i}[\mathbf{\Lambda}'\mathbf{D}^{-1}]_{jk} + [\mathbf{I}_r]_{\ell j}[\mathbf{\Lambda}'\mathbf{D}^{-1}]_{ik}\,, \quad i = 2, 3, \ldots, r;\ j = 1, 2, \ldots, i-1 \\
&= [\mathbf{I}_r \otimes (\mathbf{\Lambda}'\mathbf{D}^{-1})]_{j\ell,ki} + [\mathbf{I}_r \otimes (\mathbf{\Lambda}'\mathbf{D}^{-1})]_{i\ell,kj}, i = 2, 3, \ldots, r; j = 1, 2, \ldots, i-1.
\end{aligned}$$

Fixing i and j in this expression and taking all the values of k and ℓ to obtain a specific row of $\mathbf{L}$, the expression shows that this row is the sum of two different rows of the matrix $\mathbf{I}_r \otimes (\mathbf{\Lambda}'\mathbf{D}^{-1})$. However, the second-last expression in the above derivation shows that only one term will have non-zero values, for i and j cannot simultaneously be equal to ℓ (since $i \neq j$).

To see more clearly how $\mathbf{L}$ may be determined from $\mathbf{I}_r \otimes (\mathbf{\Lambda}'\mathbf{D}^{-1})$, write

$$\mathbf{I}_r \otimes (\mathbf{\Lambda}'\mathbf{D}^{-1}) = \begin{pmatrix} \mathbf{\Lambda}'\mathbf{D}^{-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}'\mathbf{D}^{-1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{\Lambda}'\mathbf{D}^{-1} \end{pmatrix}$$

and note that a typical element of this kronecker product is written as $[\mathbf{I}_r \otimes \mathbf{\Lambda}'\mathbf{D}^{-1})]_{ab,cd}$ where $ab = 11,\ 21,\ldots r1,\ 12,\ 22\ldots,r2,\ldots,1r,\ 2r,\ldots,rr$ and $cd = 11,\ 21,\ldots p1,\ 12,\ 22\ldots,p2,\ldots,1p,\ 2p,\ldots,pp$. Inspection of the above shows that the first row of $\mathbf{L}$ (i.e. when i and j are fixed at 2 and 1 respectively) will contain in its first p positions (i.e. for $\ell = 1$ and $k = 1,2,\ldots,p$) the second row of $\mathbf{\Lambda}'\mathbf{D}^{-1}$, while in positions $p+1,\ p+2,\ldots,2p$ (i.e. for $\ell = 2$ and $k = 1,2,\ldots,p$) will be the first row of $\mathbf{\Lambda}'\mathbf{D}^{-1}$. The other $p(r-2)$ positions in the row will have values equal to zero.

Generally, in each row of $\mathbf{L}$ there will be $2p$ elements that correspond to two of the rows of $\mathbf{\Lambda}'\mathbf{D}^{-1}$, while the other $p(r-2)$ elements will be zero. The positions of the $2p$ non-zero elements are determined by the values of i and j.

Even closer inspection of the above shows that an $r(r-1)/2 \times r^2$ selection matrix may be used to select the necessary rows from $\mathbf{I}_r \otimes (\mathbf{\Lambda}'\mathbf{D}^{-1})$ to form $\mathbf{L}$. This selection matrix, say $\mathbf{H}$, turns out to be the selection matrix such that $\mathbf{H}\text{vec}(\mathbf{S}) = 2\text{vecs}^{\star}(\mathbf{S})$, where $\mathbf{S}$ is a $q \times q$ symmetric matrix and $\text{vecs}^{\star}(\mathbf{S}) = (s_{21}, s_{31}, s_{32}, s_{41}, s_{42}, s_{43}, \ldots, s_{q,q-1})'$, defined similarly as $\text{vec}(\mathbf{S})$ except that the diagonal elements are being omitted.

It has therefore now been shown that $\mathbf{L} = \mathbf{H}[\mathbf{I}_r \otimes (\mathbf{\Lambda}'\mathbf{D}^{-1})]$ at the point where $\mathbf{\Lambda} = \mathbf{\Lambda}_0$, or

$$\mathbf{L} = \mathbf{H}[\mathbf{I}_r \otimes (\mathbf{\Lambda}_0'\mathbf{D}^{-1})].$$

The first order Taylor expansion of $\mathbf{c}(\mathbf{\Lambda})$ can thus now be written as

$$\mathbf{c}(\boldsymbol{\Lambda}) \approx \mathbf{c}(\boldsymbol{\Lambda}_0) + \mathbf{H}\left[\mathbf{I}_r \otimes (\boldsymbol{\Lambda}_0'\mathbf{D}^{-1})\right] \text{vec}(\boldsymbol{\Lambda} - \boldsymbol{\Lambda}_0)$$

and if this expression is set equal to zero, it follows that the constraints can be approximated by

$$\mathbf{L}\text{vec}(\boldsymbol{\Lambda}) = \mathbf{H}\left[\mathbf{I}_r \otimes (\boldsymbol{\Lambda}_0'\mathbf{D}^{-1})\right] \text{vec}(\boldsymbol{\Lambda}_0) - \text{vecs}^\star(\boldsymbol{\Lambda}_0'\mathbf{D}^{-1}\boldsymbol{\Lambda}_0)$$

(since $\mathbf{c}(\boldsymbol{\Lambda}_0) = \text{vecs}^\star(\boldsymbol{\Lambda}_0'\mathbf{D}^{-1}\boldsymbol{\Lambda}_0)$).

It can easily be shown that $[\mathbf{I}_r \otimes (\boldsymbol{\Lambda}_0'\mathbf{D}^{-1})]\,\text{vec}(\boldsymbol{\Lambda}_0) = \text{vec}(\boldsymbol{\Lambda}_0'\mathbf{D}^{-1}\boldsymbol{\Lambda}_0)$ so that the constraints become

$$\begin{aligned}
\mathbf{L}\text{vec}(\boldsymbol{\Lambda}) &= \mathbf{H}\text{vec}(\boldsymbol{\Lambda}_0'\mathbf{D}^{-1}\boldsymbol{\Lambda}_0) - \text{vecs}^\star(\boldsymbol{\Lambda}_0'\mathbf{D}^{-1}\boldsymbol{\Lambda}_0) \\
&= 2\text{vecs}^\star(\boldsymbol{\Lambda}_0'\mathbf{D}^{-1}\boldsymbol{\Lambda}_0) - \text{vecs}^\star(\boldsymbol{\Lambda}_0'\mathbf{D}^{-1}\boldsymbol{\Lambda}_0) \\
&= \text{vecs}^\star(\boldsymbol{\Lambda}_0'\mathbf{D}^{-1}\boldsymbol{\Lambda}_0)
\end{aligned}$$

which completes the proof of this proposition. □

Proposition 4.12 will now be used to linearise the constraints imposed on the elements of $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$. The method of Lagrange multipliers will then be applied to maximise the log-likelihood function $\ell n\, L$ in estimating $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$ subject to these linear constraints.

First consider the equality constraints on $\boldsymbol{\Lambda}_1$, namely $\boldsymbol{\Lambda}_1'\mathbf{D}_1^{-1}\boldsymbol{\Lambda}_1$ = diagonal. Making use of Proposition 4.12 shows that these non-linear constraints may be approximated by

$$\mathbf{L}_1 \text{vec}(\mathbf{\Lambda}_1) = \text{vecs}^\star(\mathbf{\Lambda}_{10}' \mathbf{D}_1^{-1} \mathbf{\Lambda}_{10}) \tag{4.30}$$

where $\mathbf{\Lambda}_{10}$ is a $p \times r_1$ known matrix, and $\mathbf{L}_1$ is the $r_1(r_1-1)/2 \times pr_1$ known matrix defined by

$$\mathbf{L}_1 = \mathbf{H}_1[\mathbf{I}_{r_1} \otimes (\mathbf{\Lambda}_{10}' \mathbf{D}_1^{-1})]$$

and $\mathbf{H}_1$ is a selection matrix (see the proof of Proposition 4.12).

Consider the non-linear constraints on the elements of $\mathbf{\Lambda}_2$ where these elements should satisfy the equality constraints $\mathbf{\Lambda}_2' \mathbf{D}_2^{-1} \mathbf{\Lambda}_2$ = diagonal. Proposition 4.12 may then be used to approximate these non-linear constraints by the linear constraints given by

$$\mathbf{L}_2 \text{vec}(\mathbf{\Lambda}_2) = \text{vecs}^\star(\mathbf{\Lambda}_{20}' \mathbf{D}_2^{-1} \mathbf{\Lambda}_{20}) \tag{4.31}$$

where $\mathbf{\Lambda}_{20}$ is a $p \times r_2$ known matrix, and $\mathbf{L}_2$ is the $r_2(r_2-1)/2 \times pr_2$ known matrix defined by

$$\mathbf{L}_2 = \mathbf{H}_2[\mathbf{I}_{r_2} \otimes (\mathbf{\Lambda}_{20}' \mathbf{D}_2^{-1})]$$

and $\mathbf{H}_2$ is a selection matrix as defined in the proof of Proposition 4.12.

Let $\boldsymbol{\lambda}_1$ be an $r_1(r_1-1)/2 \times 1$ and $\boldsymbol{\lambda}_2$ an $r_2(r_2-1)/2 \times 1$ vector of Lagrange multipliers, and form the new function for maximisation, $L^\star$, where

$$L^\star = \ell n\, L + \boldsymbol{\lambda}_1' \left[\mathbf{L}_1 \text{vec}(\mathbf{\Lambda}_1) - \text{vecs}^\star(\mathbf{\Lambda}_{10}' \mathbf{D}_1^{-1} \mathbf{\Lambda}_{10})\right] + \boldsymbol{\lambda}_2' \left[\mathbf{L}_2 \text{vec}(\mathbf{\Lambda}_2) - \text{vecs}^\star(\mathbf{\Lambda}_{20}' \mathbf{D}_2^{-1} \mathbf{\Lambda}_{20})\right].$$

Expressions for $\hat{\mathbf{\Lambda}}_1$ and $\hat{\boldsymbol{\lambda}}_1$ that maximise $L^\star$ will now be derived. Therefore, the gradient of $L^\star$ with respect to the elements of $\mathbf{\Lambda}_1$ and $\boldsymbol{\lambda}_1$ is required and, respectively, these gradients are given by

$$\frac{\partial L^\star}{\partial \text{vec}'(\mathbf{\Lambda}_1)} = \frac{\partial \ell n\ L}{\partial \text{vec}'(\mathbf{\Lambda}_1)} + \frac{\partial}{\partial \text{vec}'(\mathbf{\Lambda}_1)} \boldsymbol{\lambda}_1' \left[\mathbf{L}_1 \text{vec}(\mathbf{\Lambda}_1) - \text{vecs}^\star(\mathbf{\Lambda}_{10}' \mathbf{D}_1^{-1} \mathbf{\Lambda}_{10}) \right] \tag{4.32}$$

and

$$\frac{\partial L^\star}{\partial \boldsymbol{\lambda}_1} = \mathbf{L}_1 \text{vec}(\mathbf{\Lambda}_1) - \text{vecs}^\star(\mathbf{\Lambda}_{10}' \mathbf{D}_1^{-1} \mathbf{\Lambda}_{10}). \tag{4.33}$$

These two expressions have to be set equal to zero and simultaneously solved for $\mathbf{\Lambda}_1$ and $\boldsymbol{\lambda}_1$. The resulting solution $\hat{\mathbf{\Lambda}}_1$ for $\mathbf{\Lambda}_1$ will then satisfy the constraints, since for this particular choice of $\hat{\mathbf{\Lambda}}_1$, (4.33) will be zero which gives the constraints.

The first term in (4.32) was determined in the previous section (see the proof of Proposition 4.10). The second term is

$$\frac{\partial}{\partial \text{vec}'(\mathbf{\Lambda}_1)} \boldsymbol{\lambda}_1' \mathbf{L}_1 \text{vec}(\mathbf{\Lambda}_1) = \boldsymbol{\lambda}_1' \mathbf{L}_1$$

and therefore it follows that

$$\frac{\partial L^\star}{\partial \text{vec}'(\mathbf{\Lambda}_1)} = \sum_{i=1}^{M} \text{E}_c \left[\sum_{j=1}^{n_i} \text{vec}' \left\{ \mathbf{D}_1^{-1} (\mathbf{y}_{ij} - \boldsymbol{\mu}_{\mathbf{y}_{ij}}) \mathbf{d}_{1,ij}' \right\} \right] + \boldsymbol{\lambda}_1' \mathbf{L}_1. \tag{4.34}$$

Setting (4.33) and (4.34) equal to zero, results in

$$\mathbf{L}_1 \text{vec}(\hat{\mathbf{\Lambda}}_1) = \text{vecs}^\star(\mathbf{\Lambda}_{10}' \mathbf{D}_1^{-1} \mathbf{\Lambda}_{10}), \tag{4.35}$$

which gives the constraints, and

$$\sum_{i=1}^{M} \mathrm{E}_c \left[\sum_{j=1}^{n_i} \mathrm{vec} \left\{ \mathbf{D}_1^{-1} (\mathbf{y}_{ij} - \mathbf{m}_i - \hat{\mathbf{\Lambda}}_1 \mathbf{d}_{1,ij}) \mathbf{d}'_{1,ij} \right\} \right] + \mathbf{L}'_1 \hat{\boldsymbol{\lambda}}_1 = \mathbf{0} \tag{4.36}$$

since $\boldsymbol{\mu}_{\mathbf{y}_{ij}} = \mathbf{m}_i + \mathbf{\Lambda}_1 \mathbf{d}_{1,ij}$.

It will now be indicated how (4.36) can be rewritten in a form that will make the use of Proposition 3.3 possible in obtaining the solution for $\hat{\mathbf{\Lambda}}_1$ and $\hat{\boldsymbol{\lambda}}_1$.

First, rewrite (4.36) as

$$\sum_{i=1}^{M} \sum_{j=1}^{n_i} \mathrm{vec} \left\{ \mathbf{D}_1^{-1} \left(\mathbf{y}_{ij} \mathrm{E}_c(\mathbf{d}'_{1,ij}) - \mathrm{E}_c(\mathbf{m}_i \mathbf{d}'_{1,ij}) - \hat{\mathbf{\Lambda}}_1 \mathrm{E}_c(\mathbf{d}_{1,ij} \mathbf{d}'_{1,ij}) \right) \right\} + \mathbf{L}'_1 \hat{\boldsymbol{\lambda}}_1 = \mathbf{0}$$

or, combining all terms that do not involve $\hat{\mathbf{\Lambda}}_1$ or $\hat{\boldsymbol{\lambda}}_1$,

$$\mathrm{vec} \left\{ \mathbf{D}_1^{-1} \hat{\mathbf{\Lambda}}_1 \sum_{i=1}^{M} \sum_{j=1}^{n_i} \mathrm{E}_c(\mathbf{d}_{1,ij} \mathbf{d}'_{1,ij}) \right\} - \mathbf{L}'_1 \hat{\boldsymbol{\lambda}}_1 = \mathrm{vec} \left\{ \mathbf{D}_1^{-1} \sum_{i=1}^{M} \sum_{j=1}^{n_i} \left(\mathbf{y}_{ij} \mathrm{E}_c(\mathbf{d}'_{1,ij}) - \mathrm{E}_c(\mathbf{m}_i \mathbf{d}'_{1,ij}) \right) \right\}. \tag{4.37}$$

Now, if $\mathbf{A}$ and $\mathbf{B}$ are defined, for the case of simplifying the above expression, as

$$\mathbf{A} = \sum_{i=1}^{M} \sum_{j=1}^{n_i} \mathrm{E}_c(\mathbf{d}_{1,ij} \mathbf{d}'_{1,ij})$$

and

$$\mathbf{B} = \sum_{i=1}^{M} \sum_{j=1}^{n_i} \left(\mathbf{y}_{ij} \mathrm{E}_c(\mathbf{d}'_{1,ij}) - \mathrm{E}_c(\mathbf{m}_i \mathbf{d}'_{1,ij}) \right),$$

expression (4.37) can be rewritten in the simple form

$$\text{vec}\left(\mathbf{D}_1^{-1}\hat{\mathbf{\Lambda}}_1\mathbf{A}\right) - \mathbf{L}_1'\hat{\boldsymbol{\lambda}}_1 = \text{vec}\left(\mathbf{D}_1^{-1}\mathbf{B}\right),$$

or

$$(\mathbf{A} \otimes \mathbf{D}_1^{-1})\text{vec}(\hat{\mathbf{\Lambda}}_1) - \mathbf{L}_1'\hat{\boldsymbol{\lambda}}_1 = \text{vec}(\mathbf{D}_1^{-1}\mathbf{B}). \tag{4.38}$$

This equation is of the form that permits the use of Proposition 3.3 to obtain $\hat{\mathbf{\Lambda}}_1$ and $\hat{\boldsymbol{\lambda}}_1$, since (4.38) may be written as

$$\tilde{\mathbf{Y}}\hat{\boldsymbol{\tau}} + \mathbf{L}_{\tau_0}'\hat{\boldsymbol{\lambda}}_\tau = \tilde{\mathbf{z}}$$

where

$$\tilde{\mathbf{Y}} = \mathbf{A} \otimes \mathbf{D}_1^{-1},$$

$$\hat{\boldsymbol{\tau}} = \text{vec}(\hat{\mathbf{\Lambda}}_1),$$

$$\mathbf{L}_{\tau_0}' = \mathbf{L}_1',$$

$$\hat{\boldsymbol{\lambda}}_\tau = -\hat{\boldsymbol{\lambda}}_1$$

and

$$\tilde{\mathbf{z}} = \text{vec}(\mathbf{D}_1^{-1}\mathbf{B}).$$

It now follows directly from Proposition 3.3 that

$$\text{vec}(\hat{\mathbf{\Lambda}}_1) = (\mathbf{A}^{-1} \otimes \mathbf{D}_1)(\text{vec}(\mathbf{D}_1^{-1}\mathbf{B}) - \mathbf{L}_1'\hat{\boldsymbol{\lambda}}_1) \tag{4.39}$$

and

$$\hat{\boldsymbol{\lambda}}_1 = \tilde{\mathbf{S}}\tilde{\mathbf{w}}, \tag{4.40}$$

where

$$\tilde{\mathbf{S}} = (\mathbf{L}_1(\mathbf{A}^{-1} \otimes \mathbf{D}_1)\mathbf{L}_1')^{-1},$$

$$\tilde{\mathbf{w}} = \mathbf{L}_1(\mathbf{A}^{-1} \otimes \mathbf{D}_1)\text{vec}(\mathbf{D}_1^{-1}\mathbf{B}) - \tilde{\mathbf{x}}$$

and

$$\begin{aligned} \tilde{\mathbf{x}} &= \mathbf{L}_1\text{vec}(\boldsymbol{\Lambda}_{10}) - \text{vecs}^\star(\boldsymbol{\Lambda}_{10}'\mathbf{D}_1^{-1}\boldsymbol{\Lambda}_{10}) \\ &= \text{vecs}^\star(\boldsymbol{\Lambda}_{10}'\mathbf{D}_1^{-1}\boldsymbol{\Lambda}_{10}). \end{aligned}$$

An expression for $\hat{\boldsymbol{\Lambda}}_1$ that will satisfy the constraints imposed on its elements, and $\hat{\boldsymbol{\lambda}}_1$, the vector of Lagrange multipliers, have therefore been obtained and are given by (4.39) and (4.40).

Subsequently, expressions for $\hat{\boldsymbol{\Lambda}}_2$ and $\hat{\boldsymbol{\lambda}}_2$ will be derived. The gradients of $L^\star$ with respect to the elements of $\boldsymbol{\Lambda}_2$ and $\boldsymbol{\lambda}_2$ are respectively given by

$$\frac{\partial L^\star}{\partial \text{vec}'(\boldsymbol{\Lambda}_2)} = \frac{\partial \ell n\, L}{\partial \text{vec}'(\boldsymbol{\Lambda}_2)} + \frac{\partial}{\partial \text{vec}'(\boldsymbol{\Lambda}_2)}\boldsymbol{\lambda}_2'\left[\mathbf{L}_2\text{vec}(\boldsymbol{\Lambda}_2) - \text{vecs}^\star(\boldsymbol{\Lambda}_{20}'\mathbf{D}_2^{-1}\boldsymbol{\Lambda}_{20})\right] \tag{4.41}$$

and

$$\frac{\partial L^\star}{\partial \boldsymbol{\lambda}_2} = \mathbf{L}_2\text{vec}(\boldsymbol{\Lambda}_2) - \text{vecs}^\star(\boldsymbol{\Lambda}_{20}'\mathbf{D}_1^{-1}\boldsymbol{\Lambda}_{20}). \tag{4.42}$$

As before, these two expressions have to be set equal to zero and simultaneously solved for $\boldsymbol{\Lambda}_2$ and $\boldsymbol{\lambda}_2$.

The first term in (4.41) was determined in the previous section (see the proof of Proposition 4.6), while the second term is

$$\frac{\partial}{\partial \text{vec}'(\boldsymbol{\Lambda}_2)} \boldsymbol{\lambda}_2' \mathbf{L}_2 \text{vec}(\boldsymbol{\Lambda}_2) = \boldsymbol{\lambda}_2' \mathbf{L}_2.$$

Consequently it follows that

$$\frac{\partial L^\star}{\partial \text{vec}'(\boldsymbol{\Lambda}_2)} = \sum_{i=1}^{M} \text{E}_c \left[\text{vec}' \left\{ \mathbf{V}_2^{-1} \left((\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})' - \mathbf{V}_2 \right) \mathbf{V}_2^{-1} \boldsymbol{\Lambda}_2 \right\} \right] + \boldsymbol{\lambda}_2' \mathbf{L}_2. \tag{4.43}$$

Setting (4.42) and (4.43) equal to zero, respectively yields

$$\mathbf{L}_2 \text{vec}(\hat{\boldsymbol{\Lambda}}_2) = \text{vecs}^\star \left(\boldsymbol{\Lambda}_{20}' \mathbf{D}_2^{-1} \boldsymbol{\Lambda}_{20} \right) \tag{4.44}$$

which gives the constraints, and

$$\sum_{i=1}^{M} \text{E}_c \left[\text{vec} \left\{ \mathbf{V}_2^{-1} \left((\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})' - \mathbf{V}_2 \right) \mathbf{V}_2^{-1} \hat{\boldsymbol{\Lambda}}_2 \right\} \right] + \mathbf{L}_2' \hat{\boldsymbol{\lambda}}_2 = \mathbf{0}.$$

To obtain expressions for $\hat{\boldsymbol{\Lambda}}_2$ and $\hat{\boldsymbol{\lambda}}_2$, the above expression will be rewritten in a form that will permit the use of Proposition 3.2 to obtain the estimators. It follows that it may be rewritten as

$$M \text{vec}(\mathbf{V}_2^{-1} \hat{\boldsymbol{\Lambda}}_2) - \mathbf{L}_2' \hat{\boldsymbol{\lambda}}_2 = \text{vec} \left\{ \mathbf{V}_2^{-1} \left[\sum_{i=1}^{M} \text{E}_c (\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})' \right] \mathbf{V}_2^{-1} \hat{\boldsymbol{\Lambda}}_2 \right\}$$

or, equivalently as

$$\text{vec}(\mathbf{V}_2^{-1}\hat{\mathbf{\Lambda}}_2) - \frac{1}{M}\mathbf{L}_2'\hat{\boldsymbol{\lambda}}_2 = \frac{1}{M}\text{vec}\left\{\mathbf{V}_2^{-1}\left[\sum_{i=1}^{M}\text{E}_c(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})'\right]\mathbf{V}_2^{-1}\hat{\mathbf{\Lambda}}_2\right\}. \quad (4.45)$$

In the above equation, write

$$\begin{aligned}\text{vec}(\mathbf{V}_2^{-1}\hat{\mathbf{\Lambda}}_2) &= \text{vec}(\mathbf{V}_2^{-1}\hat{\mathbf{\Lambda}}_2\mathbf{I}_{r_2}) \\ &= (\mathbf{I}_{r_2}\otimes\mathbf{V}_2^{-1})\text{vec}(\hat{\mathbf{\Lambda}}_2). \quad (4.46)\end{aligned}$$

Substitution of (4.46) into (4.45) gives

$$(\mathbf{I}_{r_2}\otimes\mathbf{V}_2^{-1})\text{vec}(\hat{\mathbf{\Lambda}}_2) - \frac{1}{M}\mathbf{L}_2'\hat{\boldsymbol{\lambda}}_2 = \text{vec}(\mathbf{U}) \quad (4.47)$$

where

$$\mathbf{U} = \frac{1}{M}\left\{\mathbf{V}_2^{-1}\left[\sum_{i=1}^{M}\text{E}_c(\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})'\right]\mathbf{V}_2^{-1}\hat{\mathbf{\Lambda}}_2\right\}.$$

Expression (4.47) is now in the form that permits the use of Proposition 3.2 to obtain equations for $\hat{\mathbf{\Lambda}}_2$ and $\hat{\boldsymbol{\lambda}}_2$. Inspection of (4.47) shows that this expression resembles (3.44) of Proposition 3.2 in that

$$\mathbf{Y} = \mathbf{I}_{r_2}\otimes\mathbf{V}_2^{-1},$$

$$\hat{\boldsymbol{\theta}} = \text{vec}(\hat{\mathbf{\Lambda}}_2),$$

$$\mathbf{L}_{\theta_0}' = \mathbf{L}_2',$$

$$\hat{\boldsymbol{\lambda}}_\theta = -M^{-1}\hat{\boldsymbol{\lambda}}_2$$

and

$$\mathbf{z} = \text{vec}(\mathbf{U}).$$

The use of Proposition 3.2 now shows that the equations for $\hat{\boldsymbol{\Lambda}}_2$ and $\hat{\boldsymbol{\lambda}}_2$ are

$$\text{vec}(\hat{\boldsymbol{\Lambda}}_2) = (\mathbf{I}_{r_2} \otimes \mathbf{V}_2)(\text{vec}(\mathbf{U}) - M^{-1}\mathbf{L}_2'\hat{\boldsymbol{\lambda}}_2) \tag{4.48}$$

and

$$\hat{\boldsymbol{\lambda}}_2 = M\mathbf{S}\mathbf{w} \tag{4.49}$$

where

$$\mathbf{S} = (\mathbf{L}_2(\mathbf{I}_{r_2} \otimes \mathbf{V}_2)\mathbf{L}_2')^{-1},$$
$$\mathbf{w} = \mathbf{L}_2(\mathbf{I}_{r_2} \otimes \mathbf{V}_2)\text{vec}(\mathbf{U}) - \mathbf{x}$$

and

$$\begin{aligned}\mathbf{x} &= \mathbf{L}_2\text{vec}(\boldsymbol{\Lambda}_{20}) - \text{vecs}^\star(\boldsymbol{\Lambda}_{20}'\mathbf{D}_2^{-1}\boldsymbol{\Lambda}_{20}) \\ &= \text{vecs}^\star(\boldsymbol{\Lambda}_{20}'\mathbf{D}_2^{-1}\boldsymbol{\Lambda}_{20}).\end{aligned}$$

An EM algorithm will now be outlined that may be used to obtain the MML estimates of the fixed parameters in an exploratory bilevel model.

Since the expressions for calculating $\hat{\boldsymbol{\Lambda}}_1$ and $\hat{\boldsymbol{\Lambda}}_2$ make use of some initial value ($\boldsymbol{\Lambda}_{10}$ and $\boldsymbol{\Lambda}_{20}$), an iterative procedure will be used to obtain $\hat{\boldsymbol{\Lambda}}_1$ and $\hat{\boldsymbol{\Lambda}}_2$ which satisfy the

constraints. These iterations are nested within the iterations of the EM algorithm. This process of iterations within iterations proceeds as follows:

Initial values need to be provided for the EM-algorithm. It was indicated in the previous section that initially one may set $\boldsymbol{\mu} = \mathbf{0}$, $\mathbf{D}_1 = \mathbf{D}_2 = \mathbf{I}_p$, while arbitrary values may be assigned to $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$. These initial values are used in the E-step to calculate the moments of the conditional distribution of $\mathbf{b}_i$ given $\mathbf{y}_i$ from (4.25) to (4.29). The moments may now be employed in (4.7), (4.8), (4.9), (4.10), (4.11), (4.12) and (4.14) to obtain new estimates for $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}_1$, $\mathbf{U}_1$, $\mathbf{D}_1$, $\boldsymbol{\Lambda}_2$, $\mathbf{U}_2$ and $\mathbf{D}_2$ - which is the M-step.

The values for $\hat{\boldsymbol{\Lambda}}_1$ and $\hat{\boldsymbol{\Lambda}}_2$, however, are used as the initial values in an iterative procedure in which $\hat{\boldsymbol{\Lambda}}_1$ and $\hat{\boldsymbol{\Lambda}}_2$ are repeatedly obtained by applying (4.39) and (4.48). The iterations are continued until the largest absolute constraint is less than 0,1 for both $\hat{\boldsymbol{\Lambda}}_1$ and $\hat{\boldsymbol{\Lambda}}_2$. The EM algorithm may now continue by using this set of parameter estimates to calculate the moments in the next E-step. The moments again are used in the next M-step to obtain the next set of estimates for $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}_1$, $\mathbf{U}_1$, $\mathbf{D}_2$, $\boldsymbol{\Lambda}_2$, $\mathbf{U}_2$ and $\mathbf{D}_2$. Again the values for $\hat{\boldsymbol{\Lambda}}_1$ and $\hat{\boldsymbol{\Lambda}}_2$ are used as initial values in an iterative procedure where $\hat{\boldsymbol{\Lambda}}_1$ and $\hat{\boldsymbol{\Lambda}}_2$ are repeatedly obtained from (4.39) and (4.48) until the largest absolute constraint is less than the largest absolute constraint of the previous EM iteration.

This process is repeated until both the estimates and the constraints have reached convergence. The criterion used for the convergence of the constraints, namely that the largest absolute constraint be less than the one of the previous EM iteration, is not implemented throughout the entire iteration procedure. As soon as the largest absolute constraint becomes practically zero (less than 10^{-6}), the EM iterations are continued when the largest absolute constraint becomes less than 10^{-6}, and no longer than the value obtained in the previous iteration.

The above procedure may be summarised by the following steps (which includes the case of no constraints):

Step 1:

Initialise the parameter estimates $\hat{\boldsymbol{\mu}} = \mathbf{0}$, $\hat{\mathbf{D}}_1 = \hat{\mathbf{D}}_2 = \mathbf{I}_p$, $\hat{\boldsymbol{\Lambda}}_1 = \hat{\boldsymbol{\Lambda}}_{10}$, $\hat{\boldsymbol{\Lambda}}_2 = \hat{\boldsymbol{\Lambda}}_{20}$. Set $c_1' = c_2' = 0,1$ (initial criterion for constraints).

Step 2:

Calculate the moments of $p(\mathbf{b}_i|\mathbf{y}_i)$ from (4.25) - (4.29).

Step 3:

Calculate the next set of parameter estimates $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Lambda}}_1$, $\hat{\mathbf{U}}_1$, $\hat{\mathbf{D}}_1$, $\hat{\boldsymbol{\Lambda}}_2$, $\hat{\mathbf{U}}_2$ and $\hat{\mathbf{D}}_2$ from (4.7), (4.8), (4.9), (4.10), (4.11), (4.12) and (4.14).

Step 4:

i. If $r_1 > 1$ use $\hat{\boldsymbol{\Lambda}}_1$ as initial value and calculate the next $\hat{\boldsymbol{\Lambda}}_1$ from (4.39), otherwise proceed with Step 5.

ii. Calculate the largest absolute constraint, c_1 .

iii. If $c_1 < c_1'$ set $c_1' = max(c_1, 10^{-6})$ and proceed with Step 5, or else return to Step 4 (i).

Step 5:

i. If $r_2 > 1$ use $\hat{\boldsymbol{\Lambda}}_2$ as initial value and calculate the next $\hat{\boldsymbol{\Lambda}}_2$ from (4.48), otherwise proceed with Step 6.

ii. Calculate the largest absolute constraint, c_2.

iii. If $c_2 < c_2'$ set $c_2' = max(c_2, 10^{-6})$ and proceed with Step 6, or else return to Step 5 (i).

Step 6:

Repeat Step 2 to Step 6 until convergence is met by both the constraints and the parameter estimates.

4.8 Estimation in the confirmatory case

A confirmatory analysis is performed when knowledge about the particular variables is available, or when one wants to test specific hypotheses regarding relationships between the observed and latent variables.

The prior information or the specific hypotheses are in these cases used to pre-specify certain parameters in the model. This means that certain parameters are fixed at pre-specified values, for example a loading parameter is fixed at zero if it is believed that there is no association between the relevant observed and latent variables. Also, since the scale of the latent variables are arbitrary, it should be fixed. This is done either by fixing the variances of the latent variables at unity, or by fixing a loading parameter on each of the latent variables at unity. The first method will result in standardized latent variables while the second method will result in latent variables with the same scale as those observed variables whose loading parameters are fixed at unity.

These parameter specifications should be done in such a way as to ensure that all the free parameters (i.e. the parameters that have to be estimated) are identified, in which case the model will be identified.

For the present model, the scale of the latent variables (factors) will be fixed by fixing a loading parameter on each factor to unity, which means that the factor variances will be left free for estimation. The remaining indeterminacies, namely $r_1^2 - r_1$ on level 1 and $r_2^2 - r_2$ on level 2, will be removed by fixing $r_1 - 1$ and $r_2 - 1$ loading parameters at zero in each column of $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ respectively. The positions of the parameters that are fixed at one and zero are chosen using the guidelines set out earlier in this chapter. If there is only one factor at any level, the only indeterminacy at that level is the scale of the factor that may then be determined by fixing any loading parameter to unity.

Once it has been decided which parameters should be fixed at unity and which should be fixed at zero, the estimation of the remaining (free) parameters can proceed. The

following EM algorithm is proposed for this estimation.

As a starting point, the algorithm needs some initial values to be assigned to all the free parameters. These values for the free parameters and the pre-specified values (zeroes and ones) for the fixed parameters are now used in the E-step to calculate the moments using (4.25) to (4.29). These moments are subsequently used in the M-step to obtain an updated set of parameter estimates. The equations for this step are (4.7) to (4.12) and (4.14), in which the pre-specified values of the fixed parameters are kept unaltered. Alternately computing the E and M steps until the estimates of the free parameters have converged will maximize the likelihood for the data being analyzed.

4.9 Practical applications

To illustrate the estimation procedure presented in this chapter, the same data set that was used in the application section of Chapter 2 is again considered. These analyses have been carried to the point where a subject-matter expert acquainted with the content of the tests could draw inferences about the empirical meaning of the results, but it is not considered part of the scope of this work to engage in interpretation.

The two-level model defined in Section 4.3 will be fitted to the data. The additional information that the sample of N=5 635 students is drawn from M=139 schools will be incorporated. Two models, one with only one factor at each level and one with two factors at each level will be considered. For these two models, the parameters will be estimated in both the exploratory and confirmatory case. In the first model, the six tests described in the one-factor example in Chapter 2 will be used to fit a two-level factor analysis model with one factor on each level. In the second model, the twelve tests used in the two-factor example in Chapter 2 will be used to fit a two-level model with two factors on each of the levels.

Example 4.6.1: One factor at each level (Exploratory analysis)

Let the 6×1 vector $\mathbf{y}$ contain the six variables for this example. The observed vector for the j-th student in the i-th school is denoted by $\mathbf{y}_{ij}$ where $i = 1, 2, \ldots, 139$ and $j = 1, 2, \ldots, n_i$ while $\mathbf{y}_i$ denotes the $6n_i$ observations made on the i-th school. The n_i's for this data set take on values that range from 17 to 60 students, or level-one units, within the 139 schools, or level-two units. It is assumed that the model defined by (4.2) holds for each school or, equivalently, that for the i-th school the covariance matrix of $\mathbf{y}_i$ has the structure

$$\text{Cov}(\mathbf{y}_i, \mathbf{y}_i') = \mathbf{I}_{n_i} \otimes \mathbf{V}_1 + \mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes \mathbf{V}_2$$

where

$$\mathbf{V}_1 = \mathbf{\Lambda}_1\mathbf{\Lambda}_1' + \mathbf{D}_1$$

and

$$\mathbf{V}_2 = \mathbf{\Lambda}_2\mathbf{\Lambda}_2' + \mathbf{D}_2.$$

The unknown parameters that have to be estimated are the 6×1 mean vector $\boldsymbol{\mu}$, the two 6×1 vectors of factor loadings $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$, and the two 6×6 diagonal matrices $\mathbf{D}_1$ and $\mathbf{D}_2$ containing the variance parameters on the different levels.

A computer program EMBIFAC - written in the FORTRAN language - was used to obtain estimates of the unknown parameters, applying the EM algorithm to the bilevel factor analysis model. In this case, the iterative procedure was terminated when the parameter estimates differed from one iteration to the next by less than 0,001. These parameter estimates were merely obtained so that they could be used as starting values in the Fisher scoring algorithm in the next chapter.

The convergence criterion was satisfied after 15 iterations. The parameter values used as starting values for this iterative procedure are provided in Table 4.1, and the estimates obtained after convergence are given in Table 4.2.

TABLE 4.1
Initial parameter values

$\hat{\boldsymbol{\mu}}$	$\hat{\boldsymbol{\Lambda}}_1$	$\hat{\mathbf{D}}_1$	$\hat{\boldsymbol{\Lambda}}_2$	$\hat{\mathbf{D}}_2$
2,740	0,500	0,500	0,500	0,100
2,257	0,500	0,500	0,500	0,100
2,549	0,500	0,500	0,500	0,100
2,593	0,500	0,500	0,500	0,100
2,636	0,500	0,500	0,500	0,100
2,607	0,500	0,500	0,500	0,100

As initial values for the parameters in $\boldsymbol{\mu}$, mean values were calculated from the sampled observations: This vector was calculated as

$$\hat{\boldsymbol{\mu}} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{y}_{ij}\right).$$

TABLE 4.2

Parameter estimates: after 15 EM iterations

$\hat{\boldsymbol{\mu}}$	$\hat{\boldsymbol{\Lambda}}_1$	$\hat{\mathbf{D}}_1$	$\hat{\boldsymbol{\Lambda}}_2$	$\hat{\mathbf{D}}_2$
2,744	0,849	0,413	0,369	0,022
2,257	0,736	0,414	0,315	0,025
2,553	0,835	0,298	0,354	0,011
2,595	0,773	0,440	0,357	0,022
2,637	0,647	0,267	0,270	0,031
2,608	0,683	0,325	0,250	0,014

Since the change from one iteration to the next is fairly small, it indicates that the estimates may have taken on values that are close to the true solution, and that these will be good starting values for a Fisher scoring procedure.

Example 4.6.2: One factor at each level (Confirmatory analysis)

In this case, the variance of the factor at the first and second level is not fixed at unity, but is left free for estimation. The scale of the factor is now determined by fixing the first variable's factor loading at unity (at both levels).

These settings do not change the model, except that $\mathbf{V}_1$ and $\mathbf{V}_2$ are now

$$\mathbf{V}_1 = \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1 \boldsymbol{\Lambda}_1' + \mathbf{D}_1$$

and

$$\mathbf{V}_2 = \boldsymbol{\Lambda}_2 \boldsymbol{\Phi}_2 \boldsymbol{\Lambda}_2' + \mathbf{D}_2.$$

The scalars $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$ represent the variances of the factors at the two levels. $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ now have the form

$$\mathbf{\Lambda}_i' = (1,0\ ;\ \lambda_{2i}\ ;\ \lambda_{3i}\ ;\ \lambda_{4i}\ ;\ \lambda_{5i}\ ;\ \lambda_{6i})$$

for i=1 and i=2.

The program EMBIFAC was used to obtain the estimates of the unknown parameters. The iterative procedure was terminated when the parameter estimates differed from one iteration to the next by less than 0,001. In the present application this criterion was satisfied after 24 iterations.

The parameter values used as starting values for the iterative procedure are provided in Table 4.3, and the estimates obtained after convergence are given in Table 4.4.

TABLE 4.3
Initial parameter values

$\hat{\mu}$	$\hat{\Lambda}_1$	$\hat{D}_1$	$\hat{U}_1$	$\hat{\Lambda}_2$	$\hat{D}_2$	$\hat{U}_2$
2,74	1,00	0,30	0,80	1,00	0,01	0,40
2,26	0,50	0,30		0,50	0,01	
2,55	0,50	0,30		0,50	0,01	
2,59	0,50	0,30		0,50	0,01	
2,64	0,50	0,30		0,50	0,01	
2,61	0,50	0,30		0,50	0,01	

As initial values for the parameters in $\boldsymbol{\mu}$, mean values were calculated from the sampled observations.

TABLE 4.4

Parameter estimates: after 24 EM iterations

$\hat{\mu}$	$\hat{\Lambda}_1$	$\hat{D}_1$	$\hat{\Phi}_1$	$\hat{\Lambda}_2$	$\hat{D}_2$	$\hat{\Phi}_2$
2,744	1,000	0,393	0,803	1,000	0,001	0,180
2,258	0,866	0,400		0,635	0,039	
2,553	0,988	0,272		0,718	0,029	
2,596	0,909	0,426		0,707	0,042	
2,638	0,760	0,257		0,552	0,039	
2,609	0,803	0,315		0,483	0,024	

Example 4.6.3: Two factors at each level (Exploratory analysis)

Let the 12×1 vector $\mathbf{y}$ contain the twelve variables for this example. It is assumed that the model in (4.2) will also provide an adequate description of the data, and that the covariance structure has the same form as in the previous example. The dimensionality in this application, however, is larger.

The unknown parameters that have to be estimated are the 12×1 mean vector $\boldsymbol{\mu}$, the two 12×2 matrices of factor loadings $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$, and the two 12×12 diagonal matrices $\mathbf{D}_1$ and $\mathbf{D}_2$ containing the variance parameters on the different levels. Since the number of factors is more than one on both levels, the parameters for the factor loadings are estimated subject to the usual constraints. In this case there is only one constraint at each level that is imposed on the factor loading parameters for that level. On level one, the off-diagonal element of $\boldsymbol{\Lambda}_1'\mathbf{D}_1^{-1}\boldsymbol{\Lambda}_1$ is constrained to zero, while on level two, the off-diagonal element of $\boldsymbol{\Lambda}_2'\mathbf{D}_2^{-1}\boldsymbol{\Lambda}_2$ is constrained to zero.

The computer program EMBIFAC was used to estimate the unknown parameters, applying the adapted EM algorithm that handles constraints (cf. Section 4.5). As in the previous example, the iterative procedure was terminated when the parameter esti-

mates changed by less than 0,001 from one iteration to the next. The solution obtained here was subsequently used as initial values in the Fisher scoring algorithm in the next chapter.

The initial parameter values used for this EM application are given in Table 4.5. The algorithm converged after 25 iterations and the solution obtained is given in Table 4.6.

TABLE 4.5

Initial parameter values

$\hat{\boldsymbol{\mu}}$	$\hat{\boldsymbol{\Lambda}}_1$		$\hat{\mathbf{D}}_1$	$\hat{\boldsymbol{\Lambda}}_2$		$\hat{\mathbf{D}}_2$
2,740	0,400	0,400	1,000	0,200	0,200	0,200
2,257	0,400	0,400	1,000	0,200	0,200	0,200
2,549	0,400	0,400	1,000	0,200	0,200	0,200
2,593	0,400	0,400	1,000	0,200	0,200	0,200
2,636	0,400	0,400	1,000	0,200	0,200	0,200
2,607	0,400	0,400	1,000	0,200	0,200	0,200
19,623	1,500	−1,500	5,000	1,000	−1,000	1,000
13,148	1,500	−1,500	5,000	1,000	−1,000	1,000
24,352	1,500	−1,500	5,000	1,000	−1,000	1,000
10,744	1,500	−1,500	5,000	1,000	−1,000	1,000
12,670	1,500	−1,500	5,000	1,000	−1,000	1,000
20,087	1,500	−1,500	5,000	1,000	−1,000	1,000

The initial parameter values for the elements of $\hat{\boldsymbol{\mu}}$ were also calculated from the sampled observations, as described in the previous example.

TABLE 4.6

Parameter estimates: after 25 EM iterations

$\hat{\mu}$	$\hat{\Lambda}_1$		$\hat{D}_1$	$\hat{\Lambda}_2$		$\hat{D}_2$
2,745	0,124	0,897	0,390	0,135	0,338	0,022
2,258	0,123	0,776	0,399	0,152	0,276	0,024
2,554	0,107	0,883	0,273	0,181	0,312	0,005
2,596	0,070	0,814	0,424	0,135	0,316	0,024
2,638	0,085	0,682	0,256	0,001	0,278	0,020
2,609	0,100	0,720	0,314	0,078	0,221	0,014
19,610	1,377	−0,617	5,590	0,515	−0,916	0,461
13,127	2,263	−0,861	11,962	0,607	−1,125	1,081
24,348	2,441	−1,230	7,462	1,717	−2,519	1,656
10,740	2,461	−1,032	7,619	1,534	−1,970	0,375
12,660	2,577	−0,887	7,789	1,476	−1,929	0,531
20,089	1,699	−0,833	12,045	0,526	−1,835	1,426

The largest absolute constraints after 25 iterations are 0,22E-13 for $\hat{\Lambda}_1$ and 0,27E-06 for $\hat{\Lambda}_2$.

Example 4.6.4: Two factors at each level (Confirmatory analysis)

In this application, Φ_1 and Φ_2 are 2×2 covariance matrices left free for estimation. The four constraints that are imposed on the parameters in Λ_1 and Λ_2 are indicated in the table below. The two parameters that are fixed at unity are used to determine the scale of the two factors.

TABLE 4.7

Free and fixed parameters

$\hat{\mathbf{\Lambda}}_i$	
1, 0	0, 0
λ_{i21}	λ_{i22}
λ_{i31}	λ_{i32}
λ_{i41}	λ_{i42}
λ_{i51}	λ_{i52}
λ_{i61}	λ_{i62}
0, 0	1, 0
λ_{i81}	λ_{i82}
λ_{i91}	λ_{i92}
$\lambda_{i,10,1}$	$\lambda_{i,10,2}$
$\lambda_{i,11,1}$	$\lambda_{i,11,2}$
$\lambda_{i,12,1}$	$\lambda_{i,12,2}$

In this table, λ_{ijk} represents the (j,k)-th parameter in $\mathbf{\Lambda}_i$, i=1 or 2. The ones and zeros in their specific positions represent the values at which those parameters are fixed.

The computer program EMBIFAC was used to estimate the unknown parameters. As in the previous examples, the tolerance limit was set at 0,001 to indicate convergence.

The parameter starting values for the iteration procedure are given in Table 4.8. Convergence was reached after 32 iterations and the estimates at that point are given in Table 4.9.

TABLE 4.8

Initial parameter values

$\hat{\boldsymbol{\mu}}$	$\hat{\boldsymbol{\Lambda}}_1$		$\hat{\mathbf{D}}_1$	$\hat{\boldsymbol{\Lambda}}_2$		$\hat{\mathbf{D}}_2$
2,740	0,400	0,400	1,000	0,200	0,200	0,200
2,257	0,400	0,400	1,000	0,200	0,200	0,200
2,549	0,400	0,400	1,000	0,200	0,200	0,200
2,593	0,400	0,400	1,000	0,200	0,200	0,200
2,636	0,400	0,400	1,000	0,200	0,200	0,200
2,607	0,400	0,400	1,000	0,200	0,200	0,200
19,623	1,500	−1,500	5,000	1,000	−1,000	1,000
13,148	1,500	−1,500	5,000	1,000	−1,000	1,000
24,352	1,500	−1,500	5,000	1,000	−1,000	1,000
10,744	1,500	−1,500	5,000	1,000	−1,000	1,000
12,670	1,500	−1,500	5,000	1,000	−1,000	1,000
20,087	1,500	−1,500	5,000	1,000	−1,000	1,000

$\hat{\mathbf{U}}_1$		$\hat{\mathbf{U}}_2$	
0,9	0,0	0,5	0,0
−0,5	1,5	−0,5	1,0

The initial parameter values for the elements of $\hat{\boldsymbol{\mu}}$ were also calculated from the sampled observations.

TABLE 4.9

Parameter estimates: after 32 EM iterations

$\hat{\mu}$	$\hat{\Lambda}_1$		$\hat{D}_1$	$\hat{\Lambda}_2$		$\hat{D}_2$
2,745	1,000	0,000	0,398	1,000	0,000	0,023
2,258	0,877	0,006	0,410	0,941	0,021	0,025
2,554	0,985	−0,013	0,285	1,184	0,008	0,007
2,596	0,889	−0,014	0,435	0,994	0,005	0,026
2,638	0,766	−0,002	0,262	0,602	−0,138	0,021
2,609	0,820	−0,000	0,343	0,707	0,021	0,016
19,610	0,000	1,000	5,919	0,000	1,000	0,466
13,127	0,118	1,761	12,022	0,011	1,288	1,088
24,348	−0,111	1,890	7,903	0,757	3,883	1,663
10,740	0,051	1,932	8,263	1,091	2,712	0,384
12,660	0,177	1,955	8,528	0,905	2,749	0,581
20,089	−0,009	1,329	12,884	−1,516	1,552	1,529

$\hat{\Phi}_1$		$\hat{\Phi}_2$	
0,815		0,156	
−0,537	2,554	−0,248	1,203

4.10 Summary

In this chapter, multilevel factor analysis models are introduced and it is shown how they fit into the general framework of Chapter 3. A two-level factor analysis model, assuming common factor structures at both levels, is then discussed. The parameters in this model are examined, and the necessity of imposing constraints in exploratory and confirmatory analysis is investigated.

The MML method that was discussed in general in Chapter 3 is applied to the model,

and the necessary equations for the EM algorithm are derived. This is done for unconstrained and constrained estimation. For both these situations, the basic steps that should be followed in the EM algorithm are given. Also, the basic steps used in the computer program EMBIFAC - written to apply this estimation procedure in practice - are provided.

Finally, four practical examples are given. In the first, only one factor is extracted at each level in an exploratory analysis - this therefore demonstrates the procedure in the case of unconstrained estimation. The second example is the confirmatory analogue of the first example, demonstrating how parameters may be fixed at one and zero. The third example demonstrates non-linear constrained estimation in an exploratory analysis, since two factors are extracted at each level, and the fourth example is the confirmatory analogue, specifying zeroes and ones.

CHAPTER 5

BILEVEL FACTOR ANALYSIS MODELS AND NORMAL MAXIMUM LIKELIHOOD

5.1 Introduction

In this chapter attention will be given to a specific estimation procedure and its application to the bilevel factor analysis model defined in the previous chapter. The estimation procedure that will be discussed is the so-called Fisher scoring method and its use in obtaining the maximum likelihood estimates of parameters under normality assumptions.

The method of maximum likelihood requires the maximisation of the likelihood function with respect to the parameters in the model. This likelihood function will be derived in Section 5.2. Thereafter, in Sections 5.3 and 5.4, the gradient vector and the expected Hessian matrix will be obtained, which are the necessary components in the Fisher scoring method in order to estimate the parameters and standard errors of these estimates. In Section 5.5 the details of the estimation procedure is presented; this is done for exploratory as well as for confirmatory models. In Section 5.6 the goodness of fit of the model criterion and hypothesis testing is discussed. The final section contains practical applications.

5.2 The likelihood function

Suppose a random sample is drawn from a hierarchically structured population with two levels. Let M indicate the number of level two units in the sample and suppose there are n_i level one units drawn from the i-th level two unit. All together there are therefore $N = \sum_{i=1}^{M} n_i$ observations in the sample.

Now suppose that each observation is a p-variate random vector. Denote the M level two units by the $pn_i \times 1$ vectors

$$\mathbf{y}_i = \begin{pmatrix} \mathbf{y}_{i1} \\ \mathbf{y}_{i2} \\ \vdots \\ \mathbf{y}_{in_i} \end{pmatrix}; \qquad i = 1, 2, \ldots, M.$$

It is now assumed that these vectors of observations can be described by the two-level factor analysis model defined in Section 4.3. This model states that

$$\mathbf{y}_i = \mathbf{j}_{n_i} \otimes \mathbf{m}_i + (\mathbf{I}_{n_i} \otimes \boldsymbol{\Lambda}_1)\mathbf{d}_{1,i} + \mathbf{e}_{1,i} \tag{5.1}$$

where

$$\mathbf{m}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}_2 \mathbf{d}_{2,i} + \mathbf{e}_{2,i}$$

and it is assumed that $\mathbf{d}_{1,i} \sim N(\mathbf{0}, \boldsymbol{\Phi}_1)$, $\mathbf{e}_{1,i} \sim N(\mathbf{0}, \mathbf{I}_{n_i} \otimes \mathbf{D}_1)$, $\mathbf{d}_{2,i} \sim N(\mathbf{0}, \boldsymbol{\Phi}_2)$ and $\mathbf{e}_{2,i} \sim N(\mathbf{0}, \mathbf{D}_2)$ and that they all are mutually independent.

From the assumptions above, it follows that

$$\mathbf{y}_i \sim N(\mathbf{j}_{n_i} \otimes \boldsymbol{\mu}, \mathbf{W}_i), \quad i = 1, 2, \ldots, M$$

where $\boldsymbol{\mu}$ is a $p \times 1$ vector of population means and $\mathbf{W}_i$ is the $pn_i \times pn_i$ population covariance matrix of $\mathbf{y}_i$ and which is assumed to have the structure indicated by (4.3).

The density function of $\mathbf{y}_i$ can now be written as

$$f(\mathbf{y}_i) = (2\pi)^{\frac{-pn_i}{2}} |\mathbf{W}_i|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left[\mathbf{W}_i^{-1}\mathbf{G}_i\right]\right\} \tag{5.2}$$

where

$$\begin{aligned} \mathbf{G}_i &= (\mathbf{y}_i - \mathbf{j}_{n_i} \otimes \boldsymbol{\mu})(\mathbf{y}_i - \mathbf{j}_{n_i} \otimes \boldsymbol{\mu})' \\ &= \mathbf{r}_i \mathbf{r}_i' \end{aligned}$$

say, where the symbol r is chosen to indicate residuals. The likelihood function of $\mathbf{y}_i$ $(i = 1, 2, \ldots, M)$ is given by

$$L = \prod_{i=1}^{M} f(\mathbf{y}_i)$$

and the natural logarithm of this function follows as

$$\ell nL = -\frac{1}{2}\sum_{i=1}^{M}\left\{pn_i\ell n(2\pi) + \ell n|\mathbf{W}_i| + \text{tr}\left[\mathbf{W}_i^{-1}\mathbf{G}_i\right]\right\}. \tag{5.3}$$

To obtain the maximum likelihood estimators of the parameters in this model, ℓnL has to be maximised with respect to the parameters, or equivalently, minimisation of

$$F(\boldsymbol{\gamma}) = -\ell nL$$

will yield the parameter estimators.

Note that here, as was the case in MML estimation, we shall also take $\boldsymbol{\Phi}_1 = \mathbf{U}_1\mathbf{U}_1'$ and $\boldsymbol{\Phi}_2 = \mathbf{U}_2\mathbf{U}_2'$ where the $\mathbf{U}_i$ matrices are lower triangular.

5.3 The gradient vector

In this section expressions are derived for the gradient vector of the function $F(\gamma)$ which will be minimised by means of the Fisher scoring method.

The unknown parameters in the model to be estimated are the elements of $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}_1$, $\mathbf{U}_1$, $\mathbf{D}_1$, $\boldsymbol{\Lambda}_2$, $\mathbf{U}_2$ and $\mathbf{D}_2$. Let the $q \times 1$ vector containing all these parameters be denoted by γ where

$$\gamma = \begin{pmatrix} \boldsymbol{\mu} \\ \text{vec}(\boldsymbol{\Lambda}_1) \\ \text{vecs}(\mathbf{U}_1) \\ \text{diag}(\mathbf{D}_1) \\ \text{vec}(\boldsymbol{\Lambda}_2) \\ \text{vecs}(\mathbf{U}_2) \\ \text{diag}(\mathbf{D}_2) \end{pmatrix} = \begin{pmatrix} \gamma^{\boldsymbol{\mu}} \\ \gamma^{\boldsymbol{\Lambda}_1} \\ \gamma^{\mathbf{U}_1} \\ \gamma^{\mathbf{D}_1} \\ \gamma^{\boldsymbol{\Lambda}_2} \\ \gamma^{\mathbf{U}_2} \\ \gamma^{\mathbf{D}_2} \end{pmatrix},$$

and $q = p(r_1 + r_2 + 3) + r_1(r_1 + 1)/2 + r_2(r_2 + 1)/2$.

The gradient of a general discrepancy function under normality assumptions has already been defined in Chapter 3 (cf. (3.26)). For the discrepancy function $F(\gamma)$ which has been derived in the previous section, the expression for a typical element of the gradient vector is now given by

$$[\mathbf{g}(\gamma)]_k = -\sum_{i=1}^{M} \left\{ \text{tr}[\mathbf{r}_i' \mathbf{W}_i^{-1} \frac{\partial \mathbf{j}_{n_i} \otimes \boldsymbol{\mu}}{\partial \gamma_k}] + \frac{1}{2} \text{tr}[\mathbf{W}_i^{-1}(\mathbf{G}_i - \mathbf{W}_i)\mathbf{W}_i^{-1} \frac{\partial \mathbf{W}_i}{\partial \gamma_k}] \right\}. \qquad (5.4)$$

The partitioning of γ into natural subsets of parameters provides a useful way to partition $\mathbf{g}(\gamma)$ accordingly. Consequently $\mathbf{g}(\gamma)$ is now written as

$$\mathbf{g}(\boldsymbol{\gamma}) = \begin{pmatrix} \mathbf{g}(\boldsymbol{\gamma}^{\boldsymbol{\mu}}) \\ \mathbf{g}(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}) \\ \mathbf{g}(\boldsymbol{\gamma}^{\mathbf{U}_1}) \\ \mathbf{g}(\boldsymbol{\gamma}^{\mathbf{D}_1}) \\ \mathbf{g}(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2}) \\ \mathbf{g}(\boldsymbol{\gamma}^{\mathbf{U}_2}) \\ \mathbf{g}(\boldsymbol{\gamma}^{\mathbf{D}_2}) \end{pmatrix}.$$

The subvectors of $\mathbf{g}(\boldsymbol{\gamma})$ will be derived using (5.4) by first obtaining expressions for typical elements of the different subvectors and then generalizing these results.

In advance, note that the inverse of the population covariance matrix, $\mathbf{W}_i^{-1}$, plays an important role in the derivations. It is therefore necessary to use a specific form of $\mathbf{W}_i^{-1}$ to obtain the required results. This form of $\mathbf{W}_i^{-1}$ has been derived in Section 4.4 and is given by the expression

$$\begin{aligned} \mathbf{W}_i^{-1} &= \mathbf{I}_{n_i} \otimes \mathbf{V}_1^{-1} - \mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes \mathbf{C}_i \\ &= \begin{pmatrix} \mathbf{S}_i & -\mathbf{C}_i & \cdots & -\mathbf{C}_i \\ -\mathbf{C}_i & \mathbf{S}_i & \cdots & -\mathbf{C}_i \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{C}_i & -\mathbf{C}_i & \cdots & \mathbf{S}_i \end{pmatrix} \end{aligned} \qquad (5.5)$$

where

$$\mathbf{S}_i = \mathbf{V}_1^{-1} - \mathbf{C}_i \qquad (5.6)$$

and

$$\mathbf{C}_i = \mathbf{V}_1^{-1}(\mathbf{V}_2^{-1} + n_i\mathbf{V}_1^{-1})^{-1}\mathbf{V}_1^{-1}. \tag{5.7}$$

In addition, to simplify expressions in later derivations, define

$$\mathbf{S}_{n_i} = \mathbf{V}_1^{-1} - n_i\mathbf{C}_i. \tag{5.8}$$

The following two propositions provide identities which will be extensively used in deriving the gradient vector.

Proposition 5.1

The n_i submatrices, each of the order $p \times p$, on the main diagonal of the $pn_i \times pn_i$ partitioned matrix

$$\mathbf{W}_i^{-1}(\mathbf{G}_i - \mathbf{W}_i)\mathbf{W}_i^{-1}(\mathbf{I}_{n_i} \otimes \boldsymbol{\Upsilon})$$

where $\boldsymbol{\Upsilon}$ is a $p \times p$ matrix, are the matrices

$$(\mathbf{w}_{i1}\mathbf{w}_{i1}' - \mathbf{S}_i)\boldsymbol{\Upsilon}, \quad (\mathbf{w}_{i2}\mathbf{w}_{i2}' - \mathbf{S}_i)\boldsymbol{\Upsilon}, \quad \ldots \quad , \quad (\mathbf{w}_{in_i}\mathbf{w}_{in_i}' - \mathbf{S}_i)\boldsymbol{\Upsilon}$$

where $\mathbf{w}_{ij}$ is the j-th $p \times 1$ subvector of $\mathbf{W}_i^{-1}\mathbf{r}_i$.

Proof

Write

$$\mathbf{W}_i^{-1}(\mathbf{G}_i - \mathbf{W}_i)\mathbf{W}_i^{-1} \;=\; \mathbf{W}_i^{-1}\mathbf{r}_i\mathbf{r}_i'\mathbf{W}_i^{-1} - \mathbf{W}_i^{-1}$$

$$= \mathbf{w}_i\mathbf{w}_i' - \mathbf{W}_i^{-1}$$

where

$$\mathbf{w}_i = \mathbf{W}_i^{-1}\mathbf{r}_i = \begin{pmatrix} \mathbf{w}_{i1} \\ \mathbf{w}_{i2} \\ \vdots \\ \mathbf{w}_{in_i} \end{pmatrix}.$$

Consequently we obtain

$$\begin{aligned} \mathbf{W}_i^{-1}(\mathbf{G}_i - \mathbf{W}_i)\mathbf{W}_i^{-1} &= \begin{pmatrix} \mathbf{w}_{i1}\mathbf{w}_{i1}' & \cdots & \mathbf{w}_{i1}\mathbf{w}_{in_i}' \\ \vdots & \ddots & \vdots \\ \mathbf{w}_{in_i}\mathbf{w}_{i1}' & \cdots & \mathbf{w}_{in_i}\mathbf{w}_{in_i}' \end{pmatrix} - \begin{pmatrix} \mathbf{S}_i & \cdots & -\mathbf{C}_i \\ \vdots & \ddots & \vdots \\ -\mathbf{C}_i & \cdots & \mathbf{S}_i \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{w}_{i1}\mathbf{w}_{i1}' - \mathbf{S}_i & \cdots & \mathbf{w}_{i1}\mathbf{w}_{in_i}' + \mathbf{C}_i \\ \vdots & \ddots & \vdots \\ \mathbf{w}_{in_i}\mathbf{w}_{i1}' + \mathbf{C}_i & \cdots & \mathbf{w}_{in_i}\mathbf{w}_{in_i}' - \mathbf{S}_i \end{pmatrix}. \end{aligned} \tag{5.9}$$

Note that

$$\mathbf{I}_{n_i} \otimes \boldsymbol{\Upsilon} = \begin{pmatrix} \boldsymbol{\Upsilon} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{\Upsilon} \end{pmatrix}$$

and post-multiplication of (5.9) with the above proves the proposition. □

Proposition 5.2

The n_i submatrices, each of the order $p \times p$, on the main diagonal of the $pn_i \times pn_i$ partitioned matrix

$$\mathbf{W}_i^{-1}(\mathbf{G}_i - \mathbf{W}_i)\mathbf{W}_i^{-1}(\mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes \boldsymbol{\Upsilon})$$

where $\boldsymbol{\Upsilon}$ is a $p \times p$ matrix, are the matrices

$$(\sum_{k=1}^{n_i} \mathbf{w}_{i1}\mathbf{w}_{ik}' - \mathbf{S}_{n_i})\boldsymbol{\Upsilon}, \quad (\sum_{k=1}^{n_i} \mathbf{w}_{i2}\mathbf{w}_{ik}' - \mathbf{S}_{n_i})\boldsymbol{\Upsilon}, \quad \ldots \quad , \quad (\sum_{k=1}^{n_i} \mathbf{w}_{in_i}\mathbf{w}_{ik}' - \mathbf{S}_{n_i})\boldsymbol{\Upsilon}$$

where $\mathbf{w}_{ij}$ is the j-th $p \times 1$ subvector of $\mathbf{W}_i^{-1}\mathbf{r}_i$.

Proof

Note that

$$\mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes \boldsymbol{\Upsilon} = \begin{pmatrix} \boldsymbol{\Upsilon} & \cdots & \boldsymbol{\Upsilon} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Upsilon} & \cdots & \boldsymbol{\Upsilon} \end{pmatrix}$$

and post-multiplication of (5.9) with the above proves the proposition. □

The following five propositions provide expressions for the gradient vector of $F(\boldsymbol{\gamma})$.

Proposition 5.3

The gradient of $F(\boldsymbol{\gamma})$ with respect to the elements of $\boldsymbol{\mu}$ is given by the $p \times 1$ vector defined as

$$\mathbf{g}(\boldsymbol{\gamma}^{\boldsymbol{\mu}}) = -\sum_{i=1}^{M}\sum_{j=1}^{n_i}\mathbf{w}_{ij}. \tag{5.10}$$

Proof

A typical element of $\mathbf{g}(\boldsymbol{\gamma}^{\boldsymbol{\mu}})$, say the s-th element, can be written as

$$\begin{aligned}[\mathbf{g}(\boldsymbol{\gamma}^{\boldsymbol{\mu}})]_s &= -\sum_{i=1}^{M}\mathrm{tr}\left[\mathbf{r}_i'\mathbf{W}_i^{-1}\frac{\partial \mathbf{j}_{n_i}\otimes\boldsymbol{\mu}}{\partial\mu_s}\right] \\ &= -\sum_{i=1}^{M}\mathrm{tr}[\mathbf{w}_i'(\mathbf{j}_{n_i}\otimes\mathbf{J}_{s1})] \end{aligned} \tag{5.11}$$

where $\mathbf{w}_i$ has been defined in the proof of Proposition 5.1.

Note that

$$\mathbf{j}_{n_i}\otimes\mathbf{J}_{s1} = \begin{pmatrix}\mathbf{J}_{s1}\\ \mathbf{J}_{s1}\\ \vdots\\ \mathbf{J}_{s1}\end{pmatrix}$$

and if the partitioning of $\mathbf{w}_i$ into its n_i subvectors $\mathbf{w}_{ij}$ $(j = 1, 2, \ldots, n_i)$ is used it follows that

$$[\mathbf{g}(\boldsymbol{\gamma}^{\boldsymbol{\mu}})]_s == -\sum_{i=1}^{M}\mathrm{tr}[\sum_{k=1}^{n_i}\mathbf{w}_{ik}'\mathbf{J}_{s1}].$$

However, if we use the fact that $\mathrm{tr}(\mathbf{a}'\mathbf{J}_{s1}) = a_s$, where a_s is the s-th element of the vector $\mathbf{a}$, then we can write

$$[\mathbf{g}(\boldsymbol{\gamma}^{\boldsymbol{\mu}})]_s = \left[-\sum_{i=1}^{M}\sum_{k=1}^{n_i}\mathbf{w}'_{ik}\right]_s$$

for a typical element of $\mathbf{g}(\boldsymbol{\gamma}^{\boldsymbol{\mu}})$, and consequently the proposition is proved. □

Proposition 5.4

The gradient of $F(\boldsymbol{\gamma})$ with respect to the elements of $\boldsymbol{\Lambda}_1$ is given by the $pr_1 \times 1$ vector defined as

$$\mathbf{g}(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}) = -\sum_{i=1}^{M}\sum_{j=1}^{n_i}\text{vec}\left\{[\mathbf{w}_{ij}\mathbf{w}'_{ij} - \mathbf{S}_i]\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\right\}. \tag{5.12}$$

Proof

First we derive an expression for a typical element of $\mathbf{g}(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1})$, say with respect to $[\boldsymbol{\Lambda}_1]_{k\ell}$. Now assume that the (k,ℓ)-th position in $\boldsymbol{\Lambda}_1$ corresponds to the s-th position in $\text{vec}(\boldsymbol{\Lambda}_1)$. Then it follows that

$$\begin{aligned}
[\mathbf{g}(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1})]_s &= -\frac{1}{2}\sum_{i=1}^{M}\text{tr}\left[\mathbf{W}_i^{-1}(\mathbf{G}_i - \mathbf{W}_i)\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial[\boldsymbol{\Lambda}_1]_{k\ell}}\right] \\
&= -\frac{1}{2}\sum_{i=1}^{M}\text{tr}\left[\mathbf{W}_i^{-1}(\mathbf{G}_i - \mathbf{W}_i)\mathbf{W}_i^{-1}(\mathbf{I}_{n_i}\otimes(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\mathbf{J}_{\ell k} + \mathbf{J}_{k\ell}\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1'))\right]
\end{aligned}$$

and since $\text{tr}[\mathbf{AB}] = \text{tr}[\mathbf{A}'\mathbf{B}']$, it follows that

$$[\mathbf{g}(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1})]_s = -\sum_{i=1}^{M}\text{tr}[\mathbf{W}_i^{-1}(\mathbf{G}_i - \mathbf{W}_i)\mathbf{W}_i^{-1}(\mathbf{I}_{n_i}\otimes\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\mathbf{J}_{\ell k})].$$

Using Proposition 5.1, the expression for the gradient with respect to the s-th element

in vec($\mathbf{\Lambda}_1$) becomes

$$[\mathbf{g}(\gamma^{\mathbf{\Lambda}_1})]_s = -\sum_{i=1}^{M} \text{tr}\left[\sum_{j=1}^{n_i}[\mathbf{w}_{ij}\mathbf{w}'_{ij} - \mathbf{S}_i]\mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{J}_{\ell k}\right].$$

However, taking the construction of $\mathbf{J}_{\ell k}$ into account, the above expression can equivalently be written as

$$\begin{aligned}[\mathbf{g}(\gamma^{\mathbf{\Lambda}_1})]_s &= -\left[\sum_{i=1}^{M}\sum_{j=1}^{n_i}\left[\mathbf{w}_{ij}\mathbf{w}'_{ij} - \mathbf{S}_i\right]\mathbf{\Lambda}_1\mathbf{\Phi}_1\right]_{k\ell} \\ &= -\left[\text{vec}\left\{\sum_{i=1}^{M}\sum_{j=1}^{n_i}[\mathbf{w}_{ij}\mathbf{w}'_{ij} - \mathbf{S}_i]\mathbf{\Lambda}_1\mathbf{\Phi}_1\right\}\right]_s\end{aligned}$$

and consequently the proposition is proved. □

Proposition 5.5

The gradient of $F(\gamma)$ with respect to the elements of $\mathbf{U}_1$ is given by the $r_1(r_1+1)/2 \times 1$ vector defined as

$$\mathbf{g}(\gamma^{\mathbf{U}_1}) = -\sum_{i=1}^{M}\sum_{j=1}^{n_i}\text{vec}\left\{\mathbf{\Lambda}'_1[\mathbf{w}_{ij}\mathbf{w}'_{ij} - \mathbf{S}_i]\mathbf{\Lambda}_1\mathbf{U}_1\right\}. \tag{5.13}$$

Proof

First we derive an expression for a typical element of $\mathbf{g}(\gamma^{\mathbf{U}_1})$, say with respect to $[\mathbf{U}_1]_{k\ell}$. Now assume that the (k,ℓ)-th position in $\mathbf{U}_1$ corresponds to the s-th position in vecs($\mathbf{U}_1$). Then it follows that

$$
\begin{aligned}
[\mathbf{g}(\boldsymbol{\gamma}^{\mathbf{U}_1})]_s &= -\frac{1}{2}\sum_{i=1}^{M} \mathrm{tr}\left[\mathbf{W}_i^{-1}(\mathbf{G}_i - \mathbf{W}_i)\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{U}_1]_{k\ell}}\right] \\
&= -\frac{1}{2}\sum_{i=1}^{M} \mathrm{tr}\left[\mathbf{W}_i^{-1}(\mathbf{G}_i - \mathbf{W}_i)\mathbf{W}_i^{-1}(\mathbf{I}_{n_i} \otimes (\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_1' + \boldsymbol{\Lambda}_1\mathbf{J}_{k\ell}\mathbf{U}_1'\boldsymbol{\Lambda}_1'))\right].
\end{aligned}
$$

Using Proposition 5.1, the expression for the gradient now becomes

$$
\begin{aligned}
[\mathbf{g}(\boldsymbol{\gamma}^{\mathbf{U}_1})]_s &= -\frac{1}{2}\sum_{i=1}^{M} \mathrm{tr}\left[\sum_{j=1}^{n_i}[\mathbf{w}_{ij}\mathbf{w}_{ij}' - \mathbf{S}_i](\boldsymbol{\Lambda}_1\mathbf{J}_{k\ell}\mathbf{U}_1'\boldsymbol{\Lambda}_1' + \boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_1')\right] \\
&= -\sum_{i=1}^{M} \mathrm{tr}\left[\sum_{j=1}^{n_i}\left(\boldsymbol{\Lambda}_1'(\mathbf{w}_{ij}\mathbf{w}_{ij}' - \mathbf{S}_i)\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{\ell k}\right)\right]
\end{aligned}
$$

However, taking the construction of $\mathbf{J}_{\ell k}$ into account, the above expression can equivalently be written as

$$
\begin{aligned}
[\mathbf{g}(\boldsymbol{\gamma}^{\mathbf{U}_1})]_s &= -\left[\sum_{i=1}^{M}\sum_{j=1}^{n_i}\boldsymbol{\Lambda}_1'\left[\mathbf{w}_{ij}\mathbf{w}_{ij}' - \mathbf{S}_i\right]\boldsymbol{\Lambda}_1\mathbf{U}_1\right]_{k\ell} \\
&= -\left[\mathrm{vec}\left\{\sum_{i=1}^{M}\sum_{j=1}^{n_i}\boldsymbol{\Lambda}_1'[\mathbf{w}_{ij}\mathbf{w}_{ij}' - \mathbf{S}_i]\boldsymbol{\Lambda}_1\mathbf{U}_1\right\}\right]_s
\end{aligned}
$$

and consequently the proposition is proved. □

Proposition 5.6

The gradient of $F(\boldsymbol{\gamma})$ with respect to the elements of $\mathbf{D}_1$ is given by the $p \times 1$ vector defined by the expression

$$\mathbf{g}(\boldsymbol{\gamma}^{\mathbf{D}_1}) = -\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{n_i} \text{diag}\left\{\mathbf{w}_{ij}\mathbf{w}_{ij}' - \mathbf{S}_i\right\}. \tag{5.14}$$

Proof

Consider a typical element of $\mathbf{D}_1$, say the s-th diagonal element. From (5.4) it follows that

$$[\mathbf{g}(\boldsymbol{\gamma}^{\mathbf{D}_1})]_s = -\frac{1}{2}\sum_{i=1}^{M} \text{tr}\left[\mathbf{W}_i^{-1}(\mathbf{G}_i - \mathbf{W}_i)\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{D}_1]_{ss}}\right]$$

and using the special partitioning of $\mathbf{W}_i$ to obtain the derivative, we have

$$[\mathbf{g}(\boldsymbol{\gamma}^{\mathbf{D}_1})]_s = -\frac{1}{2}\sum_{i=1}^{M} \text{tr}\left[\mathbf{W}_i^{-1}(\mathbf{G}_i - \mathbf{W}_i)\mathbf{W}_i^{-1}(\mathbf{I}_{n_i} \otimes \mathbf{J}_{ss})\right].$$

From Proposition 5.1, it follows that

$$\begin{aligned}
[\mathbf{g}(\boldsymbol{\gamma}^{\mathbf{D}_1})]_s &= -\frac{1}{2}\sum_{i=1}^{M} \text{tr}\left[\sum_{j=1}^{n_i}(\mathbf{w}_{ij}\mathbf{w}_{ij}' - \mathbf{S}_i)\mathbf{J}_{ss}\right] \\
&= -\frac{1}{2}\left[\sum_{i=1}^{M}\sum_{j=1}^{n_i}(\mathbf{w}_{ij}\mathbf{w}_{ij}' - \mathbf{S}_i)\right]_{ss}
\end{aligned}$$

and consequently the proposition is proved. □

Proposition 5.7

The gradient of $F(\boldsymbol{\gamma})$ with respect to the elements of $\boldsymbol{\Lambda}_2$ is given by the $pr_2 \times 1$ vector defined by the expression

$$\mathbf{g}(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2}) = -\sum_{i=1}^{M}\sum_{j'=1}^{n_i} \text{vec}\left\{\left[\sum_{j=1}^{n_i}\mathbf{w}_{ij'}\mathbf{w}_{ij}' - \mathbf{S}_{n_i}\right]\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2\right\}. \tag{5.15}$$

Proof

First consider a typical element of $\mathbf{g}(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2})$, say with respect to $[\boldsymbol{\Lambda}_2]_{k\ell}$, where it is assumed that the (k,ℓ)-th position in $\boldsymbol{\Lambda}_2$ corresponds to the s-th position in $\text{vec}(\boldsymbol{\Lambda}_2)$. It then follows from (5.4) that

$$\begin{aligned}
[\mathbf{g}(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2})]_s &= -\frac{1}{2}\sum_{i=1}^{M}\text{tr}\left[\mathbf{W}_i^{-1}(\mathbf{G}_i - \mathbf{W}_i)\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\boldsymbol{\Lambda}_2]_{k\ell}}\right] \\
&= -\frac{1}{2}\sum_{i=1}^{M}\text{tr}\left[\mathbf{W}_i^{-1}(\mathbf{G}_i - \mathbf{W}_i)\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes (\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2\mathbf{J}_{\ell k} + \mathbf{J}_{k\ell}\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2')\right)\right]
\end{aligned}$$

and again using $\text{tr}[\mathbf{AB}] = \text{tr}[\mathbf{A}'\mathbf{B}']$, it follows that

$$[\mathbf{g}(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2})]_s = -\sum_{i=1}^{M}\text{tr}\left[\mathbf{W}_i^{-1}(\mathbf{G}_i - \mathbf{W}_i)\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes \boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2\mathbf{J}_{\ell k}\right)\right].$$

The use of Proposition 5.2 yields

$$\begin{aligned}
[\mathbf{g}(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2})]_s &= -\sum_{i=1}^{M}\text{tr}\left[\sum_{j'=1}^{n_i}\left(\sum_{j=1}^{n_i}\mathbf{w}_{ij'}\mathbf{w}_{ij}' - \mathbf{S}_{n_i}\right)\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2\mathbf{J}_{\ell k}\right] \\
&= -\left[\sum_{i=1}^{M}\sum_{j'=1}^{n_i}\left(\sum_{j=1}^{n_i}\mathbf{w}_{ij'}\mathbf{w}_{ij}' - \mathbf{S}_{n_i}\right)\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2\right]_{k\ell} \\
&= -\left[\text{vec}\sum_{i=1}^{M}\sum_{j'=1}^{n_i}\left(\sum_{j=1}^{n_i}\mathbf{w}_{ij'}\mathbf{w}_{ij}' - \mathbf{S}_{n_i}\right)\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2\right]_s
\end{aligned}$$

and therefore the proposition is proved. □

Proposition 5.8

The gradient of $F(\boldsymbol{\gamma})$ with respect to the elements of $\mathbf{U}_2$ is given by the $r_2(r_2+1)/2 \times 1$ vector defined by the expression

$$\mathbf{g}(\boldsymbol{\gamma}^{\mathbf{U}_2}) = -\sum_{i=1}^{M}\sum_{j'=1}^{n_i} \text{vec}\left\{\boldsymbol{\Lambda}_2'\left[\sum_{j=1}^{n_i}\mathbf{w}_{ij'}\mathbf{w}_{ij}' - \mathbf{S}_{n_i}\right]\boldsymbol{\Lambda}_2\mathbf{U}_2\right\}. \tag{5.16}$$

Proof

First consider a typical element of $\mathbf{g}(\boldsymbol{\gamma}^{\mathbf{U}_2})$, say with respect to $[\mathbf{U}_2]_{k\ell}$, where it is assumed that the (k,ℓ)-th position in $\mathbf{U}_2$ corresponds to the s-th position in vecs$(\mathbf{U}_2)$. It then follows from (5.4) that

$$\begin{aligned}
[\mathbf{g}(\boldsymbol{\gamma}^{\mathbf{U}_2})]_s &= -\frac{1}{2}\sum_{i=1}^{M}\text{tr}\left[\mathbf{W}_i^{-1}(\mathbf{G}_i - \mathbf{W}_i)\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{U}_2]_{k\ell}}\right] \\
&= -\frac{1}{2}\sum_{i=1}^{M}\text{tr}\left[\mathbf{W}_i^{-1}(\mathbf{G}_i - \mathbf{W}_i)\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}'\otimes(\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_2' + \boldsymbol{\Lambda}_2\mathbf{J}_{k\ell}\mathbf{U}_2'\boldsymbol{\Lambda}_2')\right)\right].
\end{aligned}$$

The use of Proposition 5.2 yields

$$\begin{aligned}
[\mathbf{g}(\boldsymbol{\gamma}^{\mathbf{U}_2})]_s &= -\frac{1}{2}\sum_{i=1}^{M}\text{tr}\left[\sum_{j'=1}^{n_i}\left(\sum_{j=1}^{n_i}\mathbf{w}_{ij'}\mathbf{w}_{ij}' - \mathbf{S}_{n_i}\right)(\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_2' + \boldsymbol{\Lambda}_2\mathbf{J}_{k\ell}\mathbf{U}_2'\boldsymbol{\Lambda}_2')\right] \\
&= -\sum_{i=1}^{M}\text{tr}\left[\boldsymbol{\Lambda}_2'\sum_{j'=1}^{n_i}\left(\sum_{j=1}^{n_i}\mathbf{w}_{ij'}\mathbf{w}_{ij}' - \mathbf{S}_{n_i}\right)\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{\ell k}\right] \\
&= -\left[\sum_{i=1}^{M}\sum_{j'=1}^{n_i}\boldsymbol{\Lambda}_2'\left(\sum_{j=1}^{n_i}\mathbf{w}_{ij'}\mathbf{w}_{ij}' - \mathbf{S}_{n_i}\right)\boldsymbol{\Lambda}_2\mathbf{U}_2\right]_{k\ell} \\
&= -\left[\text{vec}\sum_{i=1}^{M}\sum_{j'=1}^{n_i}\boldsymbol{\Lambda}_2'\left(\sum_{j=1}^{n_i}\mathbf{w}_{ij'}\mathbf{w}_{ij}' - \mathbf{S}_{n_i}\right)\boldsymbol{\Lambda}_2\mathbf{U}_2\right]_s
\end{aligned}$$

and therefore the proposition is proved. □

Proposition 5.9

The gradient of $F(\boldsymbol{\gamma})$ with respect to the elements of $\mathbf{D}_2$ is given by the $p \times 1$ vector defined as

$$\mathbf{g}(\boldsymbol{\gamma}^{\mathbf{D}_2}) = -\frac{1}{2}\sum_{i=1}^{M}\sum_{k=1}^{n_i} \text{diag}\left\{\sum_{j=1}^{n_i} \mathbf{w}_{ik}\mathbf{w}_{ij}' - \mathbf{S}_{n_i}\right\}. \tag{5.17}$$

Proof

An expression for the gradient with respect to the s-th diagonal element of $\mathbf{D}_2$ follows from (5.4) as

$$\begin{aligned}
[\mathbf{g}(\boldsymbol{\gamma}^{\mathbf{D}_2})]_s &= -\frac{1}{2}\sum_{i=1}^{M} \text{tr}\left[\mathbf{W}_i^{-1}(\mathbf{G}_i - \mathbf{W}_i)\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{D}_2]_{ss}}\right] \\
&= -\frac{1}{2}\sum_{i=1}^{M} \text{tr}\left[\mathbf{W}_i^{-1}(\mathbf{G}_i - \mathbf{W}_i)\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes \mathbf{J}_{ss}\right)\right].
\end{aligned}$$

From Proposition 5.2, it follows that

$$\begin{aligned}
[\mathbf{g}(\boldsymbol{\gamma}^{\mathbf{D}_2})]_s &= -\frac{1}{2}\sum_{i=1}^{M} \text{tr}\left[\sum_{k=1}^{n_i}\left(\sum_{j=1}^{n_i}\mathbf{w}_{ik}\mathbf{w}_{ij}' - \mathbf{S}_{n_i}\right)\mathbf{J}_{ss}\right] \\
&= -\frac{1}{2}\left[\sum_{i=1}^{M}\sum_{k=1}^{n_i}\left(\sum_{j=1}^{n_i}\mathbf{w}_{ik}\mathbf{w}_{ij}' - \mathbf{S}_{n_i}\right)\right]
\end{aligned}$$

and consequently the proposition is proved. □

By combining the results of the previous five propositions, a general expression for the gradient of $F(\boldsymbol{\gamma})$ with respect to the full parameter vector $\boldsymbol{\gamma}$ can now be given:

$$\mathbf{g}(\boldsymbol{\gamma}) = -\sum_{i=1}^{M}\sum_{j=1}^{n_i} \begin{pmatrix} \mathbf{w}_{ij} \\ \text{vec}\{(\mathbf{w}_{ij}\mathbf{w}_{ij}' - \mathbf{S}_i)\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\} \\ \text{vec}\{\boldsymbol{\Lambda}_1'(\mathbf{w}_{ij}\mathbf{w}_{ij}' - \mathbf{S}_i)\boldsymbol{\Lambda}_1\mathbf{U}_1\} \\ \frac{1}{2}\text{diag}\{\mathbf{w}_{ij}\mathbf{w}_{ij}' - \mathbf{S}_i\} \\ \text{vec}\left\{(\sum_{k=1}^{n_i}\mathbf{w}_{ik}\mathbf{w}_{ij}' - \mathbf{S}_{n_i})\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2\right\} \\ \text{vec}\left\{\boldsymbol{\Lambda}_2'(\sum_{k=1}^{n_i}\mathbf{w}_{ik}\mathbf{w}_{ij}' - \mathbf{S}_{n_i})\boldsymbol{\Lambda}_2\mathbf{U}_2\right\} \\ \frac{1}{2}\text{diag}\left\{(\sum_{k=1}^{n_i}\mathbf{w}_{ik}\mathbf{w}_{ij}' - \mathbf{S}_{n_i})\right\} \end{pmatrix}. \tag{5.18}$$

The vectors $\mathbf{w}_{ij}$ $(i = 1, 2, \ldots, M;\ j = 1, 2, \ldots, n_i)$ play a prominent role in determining the gradient vector since they appear in every subvector of the gradient. An expression which may be used in calculating these vectors in practical applications will now be provided.

Recall from the proof of Proposition 5.1 that $\mathbf{w}_{ij}$ is the j-th $p \times 1$ subvector of the $pn_i \times 1$ partitioned vector $\mathbf{w}_i = \mathbf{W}_i^{-1}\mathbf{r}_i$. Making use of (5.5) and if we write

$$\mathbf{r}_i = \begin{pmatrix} \mathbf{r}_{i1} \\ \mathbf{r}_{i2} \\ \vdots \\ \mathbf{r}_{in_i} \end{pmatrix}$$

where each $\mathbf{r}_{ij}$ is the $p \times 1$ subvector of residuals defined by

$$\mathbf{r}_{ij} = \mathbf{y}_{ij} - \boldsymbol{\mu},$$

it follows that

$$\begin{aligned}
\mathbf{W}_i^{-1}\mathbf{r}_i &= \begin{pmatrix} \mathbf{S}_i & \cdots & -\mathbf{C}_i \\ \vdots & \ddots & \vdots \\ -\mathbf{C}_i & \cdots & \mathbf{S}_i \end{pmatrix} \begin{pmatrix} \mathbf{r}_{i1} \\ \mathbf{r}_{i2} \\ \vdots \\ \mathbf{r}_{in_i} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{V}_1^{-1}\mathbf{r}_{i1} - \sum_{k=1}^{n_i} \mathbf{C}_i\mathbf{r}_{ik} \\ \mathbf{V}_1^{-1}\mathbf{r}_{i2} - \sum_{k=1}^{n_i} \mathbf{C}_i\mathbf{r}_{ik} \\ \vdots \\ \mathbf{V}_1^{-1}\mathbf{r}_{in_i} - \sum_{k=1}^{n_i} \mathbf{C}_i\mathbf{r}_{ik} \end{pmatrix}.
\end{aligned}$$

Consequently $\mathbf{w}_{ij}$ may be computed using

$$\mathbf{w}_{ij} = \mathbf{V}_1^{-1}\mathbf{r}_{ij} - \mathbf{C}_i \sum_{k=1}^{n_i} \mathbf{r}_{ik}.$$

Having obtained expressions for the different subvectors of the gradient $\mathbf{g}(\boldsymbol{\gamma})$ of $F(\boldsymbol{\gamma})$, we shall now derive expressions for the different submatrices of the approximate Hessian matrix, $\mathbf{H}(\boldsymbol{\gamma})$.

5.4 The expected Hessian matrix

In this section expressions are derived to obtain the elements of the expected Hessian matrix of $F(\boldsymbol{\gamma})$. An expression for a typical element of the expected Hessian matrix of

a general discrepancy function under normality assumptions was given in Chapter 3 (cf. expression (3.27)). For the discrepancy function $F(\boldsymbol{\gamma})$ which was derived in Section 5.2, it now follows that a typical element of the expected Hessian matrix is given by

$$[\mathbf{H}(\boldsymbol{\gamma})]_{k\ell} = \sum_{i=1}^{M} \left\{ \text{tr} \left[\frac{\partial(\mathbf{j}_{n_i} \otimes \boldsymbol{\mu})'}{\partial \gamma_k} \mathbf{W}_i^{-1} \frac{\partial \mathbf{j}_{n_i} \otimes \boldsymbol{\mu}}{\partial \gamma_\ell} \right] + \frac{1}{2} \text{tr} \left[\mathbf{W}_i^{-1} \frac{\partial \mathbf{W}_i}{\partial \gamma_k} \mathbf{W}_i^{-1} \frac{\partial \mathbf{W}_i}{\partial \gamma_\ell} \right] \right\}. \quad (5.19)$$

Using the specific partitioning of the parameter vector $\boldsymbol{\gamma}$, it is convenient to write $\mathbf{H}(\boldsymbol{\gamma})$ as

$$\mathbf{H}(\boldsymbol{\gamma}) = \begin{pmatrix} \mathbf{H}_{\boldsymbol{\mu}\boldsymbol{\mu}} & & & & & & \\ 0 & \mathbf{H}_{\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_1} & & & & & \\ 0 & \mathbf{H}_{\mathbf{U}_1\boldsymbol{\Lambda}_1} & \mathbf{H}_{\mathbf{U}_1\mathbf{U}_1} & & & & \\ 0 & \mathbf{H}_{\mathbf{D}_1\boldsymbol{\Lambda}_1} & \mathbf{H}_{\mathbf{D}_1\mathbf{U}_1} & \mathbf{H}_{\mathbf{D}_1\mathbf{D}_1} & & & \\ 0 & \mathbf{H}_{\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_1} & \mathbf{H}_{\boldsymbol{\Lambda}_2\mathbf{U}_1} & \mathbf{H}_{\boldsymbol{\Lambda}_2\mathbf{D}_1} & \mathbf{H}_{\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_2} & & \\ 0 & \mathbf{H}_{\mathbf{U}_2\boldsymbol{\Lambda}_1} & \mathbf{H}_{\mathbf{U}_2\mathbf{U}_1} & \mathbf{H}_{\mathbf{U}_2\mathbf{D}_1} & \mathbf{H}_{\mathbf{U}_2\boldsymbol{\Lambda}_2} & \mathbf{H}_{\mathbf{U}_2\mathbf{U}_2} & \\ 0 & \mathbf{H}_{\mathbf{D}_2\boldsymbol{\Lambda}_1} & \mathbf{H}_{\mathbf{D}_2\mathbf{U}_1} & \mathbf{H}_{\mathbf{D}_2\mathbf{D}_1} & \mathbf{H}_{\mathbf{D}_2\boldsymbol{\Lambda}_2} & \mathbf{H}_{\mathbf{D}_2\mathbf{U}_2} & \mathbf{H}_{\mathbf{D}_2\mathbf{D}_2} \end{pmatrix}$$

where

$$\mathbf{H}_{\boldsymbol{\mu}\boldsymbol{\mu}} = \mathbf{H}\left(\boldsymbol{\gamma}^{\boldsymbol{\mu}}, (\boldsymbol{\gamma}^{\boldsymbol{\mu}})'\right),$$

$$\mathbf{H}_{\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_1} = \mathbf{H}\left(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}, (\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1})'\right),$$

$$\mathbf{H}_{\mathbf{U}_1\boldsymbol{\Lambda}_1} = \mathbf{H}\left(\boldsymbol{\gamma}^{\mathbf{U}_1}, (\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1})'\right),$$

and similarly for the other submatrices of $\mathbf{H}(\boldsymbol{\gamma})$.

Before we proceed to derive expressions for the elements of the expected Hessian matrix, it is necessary to simplify specific matrix expressions to be used in these derivations. Let $\boldsymbol{\Upsilon}$ and $\boldsymbol{\Gamma}$ be two $p \times p$ matrices.

Proposition 5.10

The n_i submatrices, each of the order $p \times p$, on the main diagonal of the $pn_i \times pn_i$ partitioned matrix

$$\mathbf{W}_i^{-1}(\mathbf{I}_{n_i} \otimes \boldsymbol{\Upsilon})\mathbf{W}_i^{-1}(\mathbf{I}_{n_i} \otimes \boldsymbol{\Gamma})$$

are identical and have the form

$$\mathbf{S}_i\boldsymbol{\Upsilon}\mathbf{S}_i\boldsymbol{\Gamma} + (n_i - 1)\mathbf{C}_i\boldsymbol{\Upsilon}\mathbf{C}_i\boldsymbol{\Gamma}.$$

Proof

Note that

$$\mathbf{W}_i^{-1}(\mathbf{I}_{n_i} \otimes \boldsymbol{\Upsilon}) = \begin{pmatrix} \mathbf{S}_i\boldsymbol{\Upsilon} & -\mathbf{C}_i\boldsymbol{\Upsilon} & \cdots & -\mathbf{C}_i\boldsymbol{\Upsilon} \\ -\mathbf{C}_i\boldsymbol{\Upsilon} & \mathbf{S}_i\boldsymbol{\Upsilon} & \cdots & -\mathbf{C}_i\boldsymbol{\Upsilon} \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{C}_i\boldsymbol{\Upsilon} & -\mathbf{C}_i\boldsymbol{\Upsilon} & \cdots & \mathbf{S}_i\boldsymbol{\Upsilon} \end{pmatrix} \tag{5.20}$$

and similarly for $\mathbf{W}_i^{-1}(\mathbf{I}_{n_i} \otimes \boldsymbol{\Gamma})$. Multiplication of these two $pn_i \times pn_i$ matrices will yield the required result. □

Proposition 5.11

The n_i submatrices, each of the order $p \times p$, on the main diagonal of the $pn_i \times pn_i$ partitioned matrix

$$\mathbf{W}_i^{-1}(\mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes \boldsymbol{\Upsilon})\mathbf{W}_i^{-1}(\mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes \boldsymbol{\Gamma})$$

are identical and have the form

$$n_i\mathbf{S}_{n_i}\boldsymbol{\Upsilon}\mathbf{S}_{n_i}\boldsymbol{\Gamma}.$$

Proof

Note that

$$\mathbf{W}_i^{-1}(\mathbf{j}_{n_i}\mathbf{j}'_{n_i} \otimes \boldsymbol{\Upsilon}) = \begin{pmatrix} \mathbf{S}_i\boldsymbol{\Upsilon} - (n_i-1)\mathbf{C}_i\boldsymbol{\Upsilon} & \cdots & \mathbf{S}_i\boldsymbol{\Upsilon} - (n_i-1)\mathbf{C}_i\boldsymbol{\Upsilon} \\ \vdots & \ddots & \vdots \\ \mathbf{S}_i\boldsymbol{\Upsilon} - (n_i-1)\mathbf{C}_i\boldsymbol{\Upsilon} & \cdots & \mathbf{S}_i\boldsymbol{\Upsilon} - (n_i-1)\mathbf{C}_i\boldsymbol{\Upsilon} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{S}_{n_i}\boldsymbol{\Upsilon} & \cdots & \mathbf{S}_{n_i}\boldsymbol{\Upsilon} \\ \vdots & \ddots & \vdots \\ \mathbf{S}_{n_i}\boldsymbol{\Upsilon} & \cdots & \mathbf{S}_{n_i}\boldsymbol{\Upsilon} \end{pmatrix} \tag{5.21}$$

and similarly for $\mathbf{W}_i^{-1}(\mathbf{j}_{n_i}\mathbf{j}'_{n_i} \otimes \boldsymbol{\Gamma})$. Multiplication of these two $pn_i \times pn_i$ matrices will yield the required result. □

Proposition 5.12

The n_i submatrices, each of order $p \times p$, on the main diagonal of the $pn_i \times pn_i$ partitioned matrix

$$\mathbf{W}_i^{-1}(\mathbf{j}_{n_i}\mathbf{j}'_{n_i} \otimes \boldsymbol{\Upsilon})\mathbf{W}_i^{-1}(\mathbf{I}_{n_i} \otimes \boldsymbol{\Gamma})$$

are identical and have the form

$$\mathbf{S}_{n_i}\boldsymbol{\Upsilon}\mathbf{S}_{n_i}\boldsymbol{\Gamma}.$$

Proof

Multiplication of the two $pn_i \times pn_i$ matrices of the forms provided in expressions (5.20) and (5.21) in the previous two propositions, will yield the required result. □

The simplified forms of some important matrix expressions which are given by Propositions 5.10, 5.11 and 5.12 will now be particularly helpful in obtaining expressions for

the elements of the expected Hessian matrix. Use will also be made of the results

$$\text{tr}[\mathbf{AB}] = \text{tr}[\mathbf{A}'\mathbf{B}'],$$

$$\text{tr}\left[\mathbf{A}\mathbf{J}_{ij}\mathbf{B}\mathbf{J}_{rs}\right] = \text{tr}\left[\mathbf{J}_{ij}\mathbf{B}\mathbf{J}_{rs}\mathbf{A}\right] = [\mathbf{A}]_{si}\,[\mathbf{B}]_{jr}$$

(see Browne (1991)), and

$$[\mathbf{A}]_{ij}[\mathbf{B}]_{rs} = [\mathbf{A} \otimes \mathbf{B}]_{ri,sj}$$

(see e.g. Magnus and Neudecker (1988)).

Proposition 5.13

The (k,ℓ)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\boldsymbol{\mu}}, (\boldsymbol{\gamma}^{\boldsymbol{\mu}})')$ is given by the expression

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\boldsymbol{\mu}}\right]_k, \left[\boldsymbol{\gamma}^{\boldsymbol{\mu}}\right]_\ell\right) = \sum_{i=1}^{M} n_i \left[\mathbf{S}_{n_i}\right]_{k\ell}. \tag{5.22}$$

Proof

Equation (5.19) may be used to write

$$\begin{aligned}
\mathbf{H}([\boldsymbol{\gamma}^{\boldsymbol{\mu}}]_k, [\boldsymbol{\gamma}^{\boldsymbol{\mu}}]_\ell) &= \sum_{i=1}^{M} \text{tr}\left[\frac{\partial(\mathbf{j}_{n_i} \otimes \boldsymbol{\mu})'}{\partial \mu_k}\mathbf{W}_i^{-1}\frac{\partial \mathbf{j}_{n_i} \otimes \boldsymbol{\mu}}{\partial \mu_\ell}\right] \\
&= \sum_{i=1}^{M} \text{tr}\left[(\mathbf{j}_{n_i} \otimes \mathbf{J}_{k1})'\mathbf{W}_i^{-1}(\mathbf{j}_{n_i} \otimes \mathbf{J}_{\ell 1})\right]
\end{aligned}$$

and using the special construction of the $\mathbf{J}$-matrices and the partitioning of $\mathbf{W}_i^{-1}$, the above result simplifies to

$$\mathbf{H}([\boldsymbol{\gamma}^{\boldsymbol{\mu}}]_k, [\boldsymbol{\gamma}^{\boldsymbol{\mu}}]_\ell) = \sum_{i=1}^{M} n_i[\mathbf{V}_i^{-1} - n_i\mathbf{C}_i]_{k\ell}$$

which proves the proposition. □

Proposition 5.14

The (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1},(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1})')$ is given by the expression

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}\right]_u,\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}\right]_v\right) = \sum_{i=1}^{M} n_i \left\{\left[(\mathbf{S}_i\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)\otimes(\mathbf{V}_1^{-1}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)-(\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)\otimes(\mathbf{C}_i\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)\right]_{kr,s\ell}\right.$$

$$\left.+\left[\mathbf{S}_i\otimes(\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1'\mathbf{V}_1^{-1}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)-\mathbf{S}_{n_i}\otimes(\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)\right]_{\ell r,sk}\right\}. \quad (5.23)$$

Proof

Consider a typical element when the derivative is obtained with respect to $[\boldsymbol{\Lambda}_1]_{k\ell}$ and $[\boldsymbol{\Lambda}_1]_{rs}$. Assume now that the (k,ℓ)-th and (r,s)-th positions in $\boldsymbol{\Lambda}_1$ correspond to the u-th and v-th positions in $\mathrm{vec}(\boldsymbol{\Lambda}_1)$ respectively. Then, from (5.19), we have

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}\right]_u,\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}\right]_v\right)$$

$$= \frac{1}{2}\sum_{i=1}^{M}\mathrm{tr}\left[\mathbf{W}_i^{-1}\frac{\partial\mathbf{W}_i}{\partial[\boldsymbol{\Lambda}_1]_{k\ell}}\mathbf{W}_i^{-1}\frac{\partial\mathbf{W}_i}{\partial[\boldsymbol{\Lambda}_1]_{rs}}\right]$$

$$= \frac{1}{2}\sum_{i=1}^{M}\mathrm{tr}\left[\mathbf{W}_i^{-1}\left(\mathbf{I}_{n_i}\otimes(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\mathbf{J}_{\ell k}+\mathbf{J}_{k\ell}\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1')\right)\mathbf{W}_i^{-1}\left(\mathbf{I}_{n_i}\otimes(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\mathbf{J}_{sr}+\mathbf{J}_{rs}\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1')\right)\right].$$

Proposition 5.10 may subsequently be used to simplify this expression further. Consequently it follows that

$$\begin{aligned}\mathbf{H}\left(\left[\gamma^{\mathbf{\Lambda}_1}\right]_u,\left[\gamma^{\mathbf{\Lambda}_1}\right]_v\right) &= \frac{1}{2}\sum_{i=1}^{M} n_i \mathrm{tr}\left[\mathbf{S}_i(\mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{J}_{\ell k}+\mathbf{J}_{k\ell}\mathbf{\Phi}_1\mathbf{\Lambda}_1')\mathbf{S}_i(\mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{J}_{sr}+\mathbf{J}_{rs}\mathbf{\Phi}_1\mathbf{\Lambda}_1')\right.\\ &\quad \left.+(n_i-1)\mathbf{C}_i(\mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{J}_{\ell k}+\mathbf{J}_{k\ell}\mathbf{\Phi}_1\mathbf{\Lambda}_1')\mathbf{C}_i(\mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{J}_{sr}+\mathbf{J}_{rs}\mathbf{\Phi}_1\mathbf{\Lambda}_1')\right]\\ &= \sum_{i=1}^{M} n_i \mathrm{tr}\left[\mathbf{S}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{J}_{\ell k}\mathbf{S}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{J}_{sr}+\mathbf{S}_i\mathbf{J}_{k\ell}\mathbf{\Phi}_1\mathbf{\Lambda}_1'\mathbf{S}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{J}_{sr}\right.\\ &\quad \left.+(n_i-1)\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{J}_{\ell k}\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{J}_{sr}+(n_i-1)\mathbf{C}_i\mathbf{J}_{k\ell}\mathbf{\Phi}_1\mathbf{\Lambda}_1'\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{J}_{sr}\right].\end{aligned}$$

To simplify further, use will be made of the result (see Browne (1991)) $\mathrm{tr}\left[\mathbf{A}\mathbf{J}_{ij}\mathbf{B}\mathbf{J}_{rs}\right] = \mathrm{tr}\left[\mathbf{J}_{ij}\mathbf{B}\mathbf{J}_{rs}\mathbf{A}\right] = [\mathbf{A}]_{si}[\mathbf{B}]_{jr}$.

We then have

$$\begin{aligned}\mathbf{H}\left(\left[\gamma^{\mathbf{\Lambda}_1}\right]_u,\left[\gamma^{\mathbf{\Lambda}_1}\right]_v\right) &= \sum_{i=1}^{M} n_i \left\{[\mathbf{S}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{r\ell}[\mathbf{S}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{ks}+[\mathbf{S}_i]_{rk}[\mathbf{\Phi}_1\mathbf{\Lambda}_1'\mathbf{S}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{\ell s}\right.\\ &\quad \left.+(n_i-1)[\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{r\ell}[\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{ks}+(n_i-1)[\mathbf{C}_i]_{rk}[\mathbf{\Phi}_1\mathbf{\Lambda}_1'\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{\ell s}\right\}\end{aligned}$$

in which (5.6) is substituted to obtain

$$\begin{aligned}\mathbf{H}\left([\gamma^{\mathbf{\Lambda}_1}]_u,[\gamma^{\mathbf{\Lambda}_1}]_v\right) &= \sum_{i=1}^{M} n_i \left\{[\mathbf{S}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{r\ell}[(\mathbf{V}_1^{-1}-\mathbf{C}_i)\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{ks}\right.\\ &\quad +[\mathbf{S}_i]_{rk}[\mathbf{\Phi}_1\mathbf{\Lambda}_1'(\mathbf{V}_1^{-1}-\mathbf{C}_i)\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{\ell s}\\ &\quad \left.+(n_i-1)[\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{r\ell}[\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{ks}+(n_i-1)[\mathbf{C}_i]_{rk}[\mathbf{\Phi}_1\mathbf{\Lambda}_1'\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{\ell s}\right\}\\ &= \sum_{i=1}^{M} n_i \left\{[\mathbf{S}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{r\ell}[\mathbf{V}_1^{-1}\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{ks}-[\mathbf{S}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{r\ell}[\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{ks}\right.\\ &\quad +[\mathbf{S}_i]_{rk}[\mathbf{\Phi}_1\mathbf{\Lambda}_1'\mathbf{V}_1^{-1}\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{\ell s}-[\mathbf{S}_i]_{rk}[\mathbf{\Phi}_1\mathbf{\Lambda}_1'\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{\ell s}\\ &\quad \left.+(n_i-1)[\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{r\ell}[\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{ks}+(n_i-1)[\mathbf{C}_i]_{rk}[\mathbf{\Phi}_1\mathbf{\Lambda}_1'\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{\ell s}\right\}\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{M} n_i \Big\{ [\mathbf{S}_i \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1]_{r\ell} [\mathbf{V}_1^{-1} \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1]_{ks} - [(\mathbf{V}_1^{-1} - n_i \mathbf{C}_i) \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1]_{r\ell} [\mathbf{C}_i \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1]_{ks} \\
&\quad + [\mathbf{S}_i]_{rk} [\boldsymbol{\Phi}_1 \boldsymbol{\Lambda}_1' \mathbf{V}_1^{-1} \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1]_{\ell s} - [\mathbf{V}_1^{-1} - n_i \mathbf{C}_i]_{rk} [\boldsymbol{\Phi}_1 \boldsymbol{\Lambda}_1' \mathbf{C}_i \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1]_{\ell s} \Big\} \\
&= \sum_{i=1}^{M} n_i \Big\{ [\mathbf{S}_i \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1]_{r\ell} [\mathbf{V}_1^{-1} \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1]_{ks} - [\mathbf{S}_{n_i} \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1]_{r\ell} [\mathbf{C}_i \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1]_{ks} \\
&\quad + [\mathbf{S}_i]_{rk} [\boldsymbol{\Phi}_1 \boldsymbol{\Lambda}_1' \mathbf{V}_1^{-1} \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1]_{\ell s} - [\mathbf{S}_{n_i}]_{rk} [\boldsymbol{\Phi}_1 \boldsymbol{\Lambda}_1' \mathbf{C}_i \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1]_{\ell s} \Big\}.
\end{aligned}
$$

Finally (see e.g. Magnus and Neudecker (1988)) we use $[\mathbf{A}]_{ij}[\mathbf{B}]_{rs} = [\mathbf{A} \otimes \mathbf{B}]_{ri,sj}$ to write the final result as

$$
\begin{aligned}
\mathbf{H}\left([\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}]_u, [\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}]_v\right) &= \sum_{i=1}^{M} n_i \Big\{ \left[(\mathbf{S}_i \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1) \otimes (\mathbf{V}_1^{-1} \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1) - (\mathbf{S}_{n_i} \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1) \otimes (\mathbf{C}_i \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1) \right]_{kr,s\ell} \\
&\quad + \left[\mathbf{S}_i \otimes (\boldsymbol{\Phi}_1 \boldsymbol{\Lambda}_1' \mathbf{V}_1^{-1} \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1) - \mathbf{S}_{n_i} \otimes (\boldsymbol{\Phi}_1 \boldsymbol{\Lambda}_1' \mathbf{C}_i \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1) \right]_{\ell r,sk} \Big\}
\end{aligned}
$$

which proves the proposition. □

Proposition 5.15

The (u, v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{U}_1}, (\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1})')$ is given by the expression

$$
\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{U}_1}\right]_u, \left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}\right]_v\right) \tag{5.24}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{M} n_i \Big\{ \left[(\mathbf{S}_i \boldsymbol{\Lambda}_1 \mathbf{U}_1) \otimes (\boldsymbol{\Lambda}_1' \mathbf{V}_1^{-1} \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1) - (\mathbf{S}_{n_i} \boldsymbol{\Lambda}_1 \mathbf{U}_1) \otimes (\boldsymbol{\Lambda}_1' \mathbf{C}_i \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1) \right]_{kr,s\ell} \\
&\quad + \left[(\mathbf{S}_i \boldsymbol{\Lambda}_1) \otimes (\mathbf{U}_1' \boldsymbol{\Lambda}_1' \mathbf{V}_1^{-1} \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1) - (\mathbf{S}_{n_i} \boldsymbol{\Lambda}_1) \otimes (\mathbf{U}_1' \boldsymbol{\Lambda}_1' \mathbf{C}_i \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1) \right]_{\ell r,sk} \Big\}.
\end{aligned} \tag{5.25}
$$

Proof

Consider a typical element when the derivative is obtained with respect to $[\mathbf{U}_1]_{k\ell}$ and $[\mathbf{\Lambda}_1]_{rs}$. Assume now that the (k,ℓ)-th position in $\mathbf{U}_1$ and the (r,s)-th position in $\mathbf{\Lambda}_1$ correspond to the u-th position in vecs$(\mathbf{U}_1)$ and the v-th position in vec$(\mathbf{\Lambda}_1)$ respectively. Then, from (5.19), we have

$$\begin{aligned}
&\mathbf{H}\left(\left[\gamma^{\mathbf{U}_1}\right]_u, \left[\gamma^{\mathbf{\Lambda}_1}\right]_v\right) \\
&= \frac{1}{2}\sum_{i=1}^{M}\text{tr}\left[\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{U}_1]_{k\ell}}\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{\Lambda}_1]_{rs}}\right] \\
&= \frac{1}{2}\sum_{i=1}^{M}\text{tr}\left[\mathbf{W}_i^{-1}\left(\mathbf{I}_{n_i}\otimes\left(\mathbf{\Lambda}_1\mathbf{U}_1\mathbf{J}_{\ell k}\mathbf{\Lambda}_1' + \mathbf{\Lambda}_1\mathbf{J}_{k\ell}\mathbf{U}_1'\mathbf{\Lambda}_1'\right)\right)\mathbf{W}_i^{-1}\left(\mathbf{I}_{n_i}\otimes\left(\mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{J}_{sr} + \mathbf{J}_{rs}\mathbf{\Phi}_1\mathbf{\Lambda}_1'\right)\right)\right].
\end{aligned}$$

Proposition 5.10 may subsequently be used to simplify this expression further. Consequently it follows that

$$\begin{aligned}
\mathbf{H}\left(\left[\gamma^{\mathbf{U}_1}\right]_u, \left[\gamma^{\mathbf{\Lambda}_1}\right]_v\right) &= \frac{1}{2}\sum_{i=1}^{M} n_i \text{tr}\left[\mathbf{S}_i(\mathbf{\Lambda}_1\mathbf{U}_1\mathbf{J}_{\ell k}\mathbf{\Lambda}_1' + \mathbf{\Lambda}_1\mathbf{J}_{k\ell}\mathbf{U}_1'\mathbf{\Lambda}_1')\mathbf{S}_i(\mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{J}_{sr} + \mathbf{J}_{rs}\mathbf{\Phi}_1\mathbf{\Lambda}_1')\right. \\
&\quad \left. +(n_i-1)\mathbf{C}_i(\mathbf{\Lambda}_1\mathbf{U}_1\mathbf{J}_{\ell k}\mathbf{\Lambda}_1' + \mathbf{\Lambda}_1\mathbf{J}_{k\ell}\mathbf{U}_1'\mathbf{\Lambda}_1')\mathbf{C}_i(\mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{J}_{sr} + \mathbf{J}_{rs}\mathbf{\Phi}_1\mathbf{\Lambda}_1')\right] \\
&= \sum_{i=1}^{M} n_i \text{tr}\left[\mathbf{S}_i\mathbf{\Lambda}_1\mathbf{U}_1\mathbf{J}_{\ell k}\mathbf{\Lambda}_1'\mathbf{S}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{J}_{sr} + \mathbf{S}_i\mathbf{\Lambda}_1\mathbf{J}_{k\ell}\mathbf{U}_1'\mathbf{\Lambda}_1'\mathbf{S}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{J}_{sr}\right. \\
&\quad +(n_i-1)\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{U}_1\mathbf{J}_{\ell k}\mathbf{\Lambda}_1'\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{J}_{sr} \\
&\quad \left. +(n_i-1)\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{J}_{k\ell}\mathbf{U}_1'\mathbf{\Lambda}_1'\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{J}_{sr}\right].
\end{aligned}$$

To simplify further, use will again be made of the result $\text{tr}\left[\mathbf{A}\mathbf{J}_{ij}\mathbf{B}\mathbf{J}_{rs}\right] = \text{tr}\left[\mathbf{J}_{ij}\mathbf{B}\mathbf{J}_{rs}\mathbf{A}\right] = [\mathbf{A}]_{si}[\mathbf{B}]_{jr}$.

We then have

$$\begin{aligned}\mathbf{H}\left(\left[\gamma^{\mathbf{U}_1}\right]_u,\left[\gamma^{\mathbf{\Lambda}_1}\right]_v\right) &= \sum_{i=1}^{M} n_i \{[\mathbf{S}_i\mathbf{\Lambda}_1\mathbf{U}_1]_{r\ell}[\mathbf{\Lambda}_1'\mathbf{S}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{ks} + [\mathbf{S}_i\mathbf{\Lambda}_1]_{rk}[\mathbf{U}_1'\mathbf{\Lambda}_1'\mathbf{S}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{\ell s} \\ &\quad +(n_i-1)[\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{U}_1]_{r\ell}[\mathbf{\Lambda}_1'\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{ks} \\ &\quad +(n_i-1)[\mathbf{C}_i\mathbf{\Lambda}_1]_{rk}[\mathbf{U}_1'\mathbf{\Lambda}_1'\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{\ell s}\}\end{aligned}$$

in which (5.6) is substituted to obtain

$$\begin{aligned}\mathbf{H}\left([\gamma^{\mathbf{U}_1}]_u,[\gamma^{\mathbf{\Lambda}_1}]_v\right) &= \sum_{i=1}^{M} n_i \left\{[\mathbf{S}_i\mathbf{\Lambda}_1\mathbf{U}_1]_{r\ell}[\mathbf{\Lambda}_1'(\mathbf{V}_1^{-1}-\mathbf{C}_i)\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{ks}\right. \\ &\quad +[\mathbf{S}_i\mathbf{\Lambda}_1]_{rk}[\mathbf{U}_1'\mathbf{\Lambda}_1'(\mathbf{V}_1^{-1}-\mathbf{C}_i)\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{\ell s} \\ &\quad +(n_i-1)[\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{U}_1]_{r\ell}[\mathbf{\Lambda}_1'\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{ks} \\ &\quad \left.+(n_i-1)[\mathbf{C}_i\mathbf{\Lambda}_1]_{rk}[\mathbf{U}_1'\mathbf{\Lambda}_1'\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{\ell s}\right\} \\ &= \sum_{i=1}^{M} n_i \left\{[\mathbf{S}_i\mathbf{\Lambda}_1\mathbf{U}_1]_{r\ell}[\mathbf{\Lambda}_1'\mathbf{V}_1^{-1}\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{ks} - [\mathbf{S}_i\mathbf{\Lambda}_1\mathbf{U}_1]_{r\ell}[\mathbf{\Lambda}_1'\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{ks}\right. \\ &\quad + [\mathbf{S}_i\mathbf{\Lambda}_1]_{rk}[\mathbf{U}_1'\mathbf{\Lambda}_1'\mathbf{V}_1^{-1}\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{\ell s} - [\mathbf{S}_i\mathbf{\Lambda}_1]_{rk}[\mathbf{U}_1'\mathbf{\Lambda}_1'\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{\ell s} \\ &\quad \left.+ (n_i-1)[\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{U}_1]_{r\ell}[\mathbf{\Lambda}_1'\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{ks} + (n_i-1)[\mathbf{C}_i\mathbf{\Lambda}_1]_{rk}[\mathbf{U}_1'\mathbf{\Lambda}_1'\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{\ell s}\right\} \\ &= \sum_{i=1}^{M} n_i \left\{[\mathbf{S}_i\mathbf{\Lambda}_1\mathbf{U}_1]_{r\ell}[\mathbf{\Lambda}_1'\mathbf{V}_1^{-1}\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{ks} - [\mathbf{S}_{n_i}\mathbf{\Lambda}_1\mathbf{U}_1]_{r\ell}[\mathbf{\Lambda}_1'\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{ks}\right. \\ &\quad \left.+ [\mathbf{S}_i\mathbf{\Lambda}_1]_{rk}[\mathbf{U}_1'\mathbf{\Lambda}_1'\mathbf{V}_1^{-1}\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{\ell s} - [\mathbf{S}_{n_i}\mathbf{\Lambda}_1]_{rk}[\mathbf{U}_1'\mathbf{\Lambda}_1'\mathbf{C}_i\mathbf{\Lambda}_1\mathbf{\Phi}_1]_{\ell s}\right\}.\end{aligned}$$

Finally (see e.g. Magnus and Neudecker (1988)) we use $[\mathbf{A}]_{ij}[\mathbf{B}]_{rs} = [\mathbf{A}\otimes\mathbf{B}]_{ri,sj}$ to write the final result as

$$\mathbf{H}\left([\boldsymbol{\gamma}^{\mathbf{U}_1}]_u, [\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}]_v\right) = \sum_{i=1}^{M} n_i \left\{ \left[(\mathbf{S}_i\boldsymbol{\Lambda}_1\mathbf{U}_1) \otimes (\boldsymbol{\Lambda}_1'\mathbf{V}_1^{-1}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1) - (\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\mathbf{U}_1) \otimes (\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)\right]_{kr,s\ell} + \left[(\mathbf{S}_i\boldsymbol{\Lambda}_1) \otimes (\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{V}_1^{-1}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1) - (\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1) \otimes (\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)\right]_{\ell r,sk} \right\}$$

which proves the proposition. □

Proposition 5.16

The (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{U}_1}, (\boldsymbol{\gamma}^{\mathbf{U}_1})')$ is given by the expression

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{U}_1}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{U}_1}\right]_v\right)$$

$$= \sum_{i=1}^{M} n_i \left\{ \left[(\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i\boldsymbol{\Lambda}_1\mathbf{U}_1) \otimes (\boldsymbol{\Lambda}_1'\mathbf{V}_1^{-1}\boldsymbol{\Lambda}_1) - (\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\mathbf{U}_1) \otimes (\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1)\right]_{ks,r\ell} + \left[(\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i\boldsymbol{\Lambda}_1) \otimes (\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{V}_1^{-1}\boldsymbol{\Lambda}_1) - (\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1) \otimes (\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1)\right]_{\ell s,rk} \right\}. \quad (5.26)$$

Proof

Consider a typical element when the derivative is obtained with respect to $[\mathbf{U}_1]_{k\ell}$ and $[\mathbf{U}_1]_{rs}$. Assume now that the (k,ℓ)-th and the (r,s)-th position in $\mathbf{U}_1$ correspond to the u-th and the v-th position in vecs$(\mathbf{U}_1)$ respectively. Then, from (5.19), we have

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{U}_1}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{U}_1}\right]_v\right)$$

$$= \frac{1}{2}\sum_{i=1}^{M} \mathrm{tr}\left[\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{U}_1]_{k\ell}}\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{U}_1]_{rs}}\right]$$

$$= \frac{1}{2}\sum_{i=1}^{M} \text{tr}\left[\mathbf{W}_i^{-1}\left(\mathbf{I}_{n_i} \otimes \left(\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_1' + \boldsymbol{\Lambda}_1\mathbf{J}_{k\ell}\mathbf{U}_1'\boldsymbol{\Lambda}_1'\right)\right)\mathbf{W}_i^{-1}\left(\mathbf{I}_{n_i} \otimes \left(\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{sr}\boldsymbol{\Lambda}_1' + \boldsymbol{\Lambda}_1\mathbf{J}_{rs}\mathbf{U}_1'\boldsymbol{\Lambda}_1'\right)\right)\right]$$

Proposition 5.10 may subsequently be used to simplify this expression further. Consequently it follows that

$$\begin{aligned}
&\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{U}_1}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{U}_1}\right]_v\right) \\
&= \frac{1}{2}\sum_{i=1}^{M} n_i \text{tr}\left[\mathbf{S}_i(\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_1' + \boldsymbol{\Lambda}_1\mathbf{J}_{k\ell}\mathbf{U}_1'\boldsymbol{\Lambda}_1')\mathbf{S}_i(\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{sr}\boldsymbol{\Lambda}_1' + \boldsymbol{\Lambda}_1\mathbf{J}_{rs}\mathbf{U}_1'\boldsymbol{\Lambda}_1')\right. \\
&\quad \left.+(n_i-1)\mathbf{C}_i(\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_1' + \boldsymbol{\Lambda}_1\mathbf{J}_{k\ell}\mathbf{U}_1'\boldsymbol{\Lambda}_1')\mathbf{C}_i(\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{sr}\boldsymbol{\Lambda}_1' + \boldsymbol{\Lambda}_1\mathbf{J}_{rs}\mathbf{U}_1'\boldsymbol{\Lambda}_1')\right] \\
&= \sum_{i=1}^{M} n_i \text{tr}\left[\mathbf{S}_i\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_1'\mathbf{S}_i\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{sr}\boldsymbol{\Lambda}_1' + \mathbf{S}_i\boldsymbol{\Lambda}_1\mathbf{J}_{k\ell}\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{sr}\boldsymbol{\Lambda}_1'\right. \\
&\quad \left.+(n_i-1)\mathbf{C}_i\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{sr}\boldsymbol{\Lambda}_1' + (n_i-1)\mathbf{C}_i\boldsymbol{\Lambda}_1\mathbf{J}_{k\ell}\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{sr}\boldsymbol{\Lambda}_1'\right].
\end{aligned}$$

To simplify further, use will again be made of the result $\text{tr}\left[\mathbf{A}\mathbf{J}_{ij}\mathbf{B}\mathbf{J}_{rs}\right] = \text{tr}\left[\mathbf{J}_{ij}\mathbf{B}\mathbf{J}_{rs}\mathbf{A}\right] = \left[\mathbf{A}\right]_{si}\left[\mathbf{B}\right]_{jr}$.

We then have

$$\begin{aligned}
&\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{U}_1}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{U}_1}\right]_v\right) \\
&= \sum_{i=1}^{M} n_i \left\{[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i\boldsymbol{\Lambda}_1\mathbf{U}_1]_{s\ell}[\boldsymbol{\Lambda}_1'\mathbf{S}_i\boldsymbol{\Lambda}_1]_{kr} + [\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i\boldsymbol{\Lambda}_1]_{sk}[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i\boldsymbol{\Lambda}_1]_{\ell r}\right. \\
&\quad \left.+(n_i-1)[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1\mathbf{U}_1]_{s\ell}[\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1]_{kr} + (n_i-1)[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1]_{sk}[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1]_{\ell r}\right\}
\end{aligned}$$

in which (5.6) is substituted to obtain

$$\mathbf{H}\left([\boldsymbol{\gamma}^{\mathbf{U}_1}]_u, [\boldsymbol{\gamma}^{\mathbf{U}_1}]_v\right)$$

$$\begin{aligned}
&= \sum_{i=1}^{M} n_i \left\{ [\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i\boldsymbol{\Lambda}_1\mathbf{U}_1]_{s\ell}[\boldsymbol{\Lambda}_1'(\mathbf{V}_1^{-1} - \mathbf{C}_i)\boldsymbol{\Lambda}_1]_{kr} \right. \\
&\quad + [\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i\boldsymbol{\Lambda}_1]_{sk}[\mathbf{U}_1'\boldsymbol{\Lambda}_1'(\mathbf{V}_1^{-1} - \mathbf{C}_i)\boldsymbol{\Lambda}_1]_{\ell r} \\
&\quad \left. + (n_i - 1)[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1\mathbf{U}_1]_{s\ell}[\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1]_{kr} + (n_i - 1)[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1]_{sk}[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1]_{\ell r} \right\}
\end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^{M} n_i \left\{ [\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i\boldsymbol{\Lambda}_1\mathbf{U}_1]_{s\ell}[\boldsymbol{\Lambda}_1'\mathbf{V}_1^{-1}\boldsymbol{\Lambda}_1]_{kr} - [\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i\boldsymbol{\Lambda}_1\mathbf{U}_1]_{s\ell}[\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1]_{kr} \right. \\
&\quad + [\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i\boldsymbol{\Lambda}_1]_{sk}[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{V}_1^{-1}\boldsymbol{\Lambda}_1]_{\ell r} - [\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i\boldsymbol{\Lambda}_1]_{sk}[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1]_{\ell r} \\
&\quad \left. + (n_i - 1)[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1\mathbf{U}_1]_{s\ell}[\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1]_{kr} + (n_i - 1)[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1]_{sk}[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1]_{\ell r} \right\}
\end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^{M} n_i \left\{ [\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i\boldsymbol{\Lambda}_1\mathbf{U}_1]_{s\ell}[\boldsymbol{\Lambda}_1'\mathbf{V}_1^{-1}\boldsymbol{\Lambda}_1]_{kr} - [\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\mathbf{U}_1]_{s\ell}[\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1]_{kr} \right. \\
&\quad \left. + [\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i\boldsymbol{\Lambda}_1]_{sk}[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{V}_1^{-1}\boldsymbol{\Lambda}_1]_{\ell r} - [\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1]_{sk}[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1]_{\ell r} \right\}.
\end{aligned}$$

Finally (see e.g. Magnus and Neudecker (1988)) we use $[\mathbf{A}]_{ij}[\mathbf{B}]_{rs} = [\mathbf{A} \otimes \mathbf{B}]_{ri,sj}$ to write the final result as

$$\begin{aligned}
\mathbf{H}\left([\boldsymbol{\gamma}^{\mathbf{U}_1}]_u, [\boldsymbol{\gamma}^{\mathbf{U}_1}]_v\right) &= \sum_{i=1}^{M} n_i \left\{ \left[(\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i\boldsymbol{\Lambda}_1\mathbf{U}_1) \otimes (\boldsymbol{\Lambda}_1'\mathbf{V}_1^{-1}\boldsymbol{\Lambda}_1) - (\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\mathbf{U}_1) \otimes (\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1)\right]_{ks,r\ell} \right. \\
&\quad \left. + \left[(\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i\boldsymbol{\Lambda}_1) \otimes (\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{V}_1^{-1}\boldsymbol{\Lambda}_1) - (\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1) \otimes (\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{C}_i\boldsymbol{\Lambda}_1)\right]_{\ell s,rk} \right\}
\end{aligned}$$

which proves the proposition. □

Proposition 5.17

The (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{D}_1},(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1})')$ is given by

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{D}_1}\right]_u,\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}\right]_v\right)=\sum_{i=1}^{M}n_i\left\{\left[\mathbf{V}_i^{-1}\otimes(\mathbf{S}_i\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)-\mathbf{C}_i\otimes(\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)\right]_{ur,su}\right\}. \qquad (5.27)$$

Proof

The (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{D}_1},(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1})')$ is obtained as the expected second order partial derivative of $F(\boldsymbol{\gamma})$ with respect to $[\mathbf{D}_1]_{uu}$ and $[\boldsymbol{\Lambda}_1]_{rs}$. Use is made of (5.19) to write

$$\begin{aligned}\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{D}_1}\right]_u,\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}\right]_v\right) &= \frac{1}{2}\sum_{i=1}^{M}\mathrm{tr}\left[\mathbf{W}_i^{-1}\frac{\partial\mathbf{W}_i}{\partial[\mathbf{D}_1]_{uu}}\mathbf{W}_i^{-1}\frac{\partial\mathbf{W}_i}{\partial[\boldsymbol{\Lambda}_1]_{rs}}\right]\\ &= \frac{1}{2}\sum_{i=1}^{M}\mathrm{tr}\left[\mathbf{W}_i^{-1}(\mathbf{I}_{n_i}\otimes\mathbf{J}_{uu})\mathbf{W}_i^{-1}\left(\mathbf{I}_{n_i}\otimes(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\mathbf{J}_{sr}+\mathbf{J}_{rs}\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1')\right)\right].\end{aligned}$$

This expression may further be simplified using Proposition 5.10 and hence may be written as

$$\begin{aligned}\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{D}_1}\right]_u,\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}\right]_v\right) &= \frac{1}{2}\sum_{i=1}^{M}n_i\mathrm{tr}[\mathbf{S}_i\mathbf{J}_{uu}\mathbf{S}_i(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\mathbf{J}_{sr}+\mathbf{J}_{rs}\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1')\\ &\quad +(n_i-1)\mathbf{C}_i\mathbf{J}_{uu}\mathbf{C}_i(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\mathbf{J}_{sr}+\mathbf{J}_{rs}\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1')]\\ &= \sum_{i=1}^{M}n_i\mathrm{tr}\left[\mathbf{S}_i\mathbf{J}_{uu}\mathbf{S}_i\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\mathbf{J}_{sr}+(n_i-1)\mathbf{C}_i\mathbf{J}_{uu}\mathbf{C}_i\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\mathbf{J}_{sr}\right]\end{aligned}$$

Using the result $\mathrm{tr}[\mathbf{A}\mathbf{J}_{ij}\mathbf{B}\mathbf{J}_{rs}]=[\mathbf{A}]_{si}[\mathbf{B}]_{jr}$ and (5.6), it follows that

$$\begin{aligned}
\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{D}_1}\right]_u, \left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}\right]_v\right) &= \sum_{i=1}^{M} n_i \left\{ \left[\mathbf{V}_1^{-1} - \mathbf{C}_i\right]_{ru} \left[\mathbf{S}_i \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1\right]_{us} + (n_i - 1) \left[\mathbf{C}_i\right]_{ru} \left[\mathbf{C}_i \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1\right]_{us} \right\} \\
&= \sum_{i=1}^{M} n_i \left\{ \left[\mathbf{V}_1^{-1}\right]_{ru} \left[\mathbf{S}_i \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1\right]_{us} - \left[\mathbf{C}_i\right]_{ru} \left[(\mathbf{V}_1^{-1} - n_i \mathbf{C}_i) \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1\right]_{us} \right\} \\
&= \sum_{i=1}^{M} n_i \left\{ \left[\mathbf{V}_1^{-1} \otimes (\mathbf{S}_i \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1) - \mathbf{C}_i \otimes (\mathbf{S}_{n_i} \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_1)\right]_{ur,su} \right\}
\end{aligned}$$

where we used $[\mathbf{A}]_{ij}[\mathbf{B}]_{rs} = [\mathbf{A} \otimes \mathbf{B}]_{ri,sj}$ to obtain the final expression, which proves the proposition. □

Proposition 5.18

The (u, v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{D}_1}, (\boldsymbol{\gamma}^{\mathbf{U}_1})')$ is given by

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{D}_1}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{U}_1}\right]_v\right) = \sum_{i=1}^{M} n_i \left\{ \left[(\mathbf{U}_1' \boldsymbol{\Lambda}_1' \mathbf{S}_i) \otimes (\mathbf{V}_1^{-1} \boldsymbol{\Lambda}_1) - (\mathbf{U}_1' \boldsymbol{\Lambda}_1' \mathbf{S}_{n_i}) \otimes (\mathbf{C}_i \boldsymbol{\Lambda}_1)\right]_{us,ru} \right\}. \tag{5.28}$$

Proof

Suppose the (r, s)-th element of $\mathbf{U}_1$ corresponds to the v-th element of $\text{vecs}(\mathbf{U}_1)$. The (u, v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{D}_1}, (\boldsymbol{\gamma}^{\mathbf{U}_1})')$ is then obtained as the expected second order partial derivative of $F(\boldsymbol{\gamma})$ with respect to $[\mathbf{D}_1]_{uu}$ and $[\mathbf{U}_1]_{rs}$. Use is made of (5.19) to write

$$\begin{aligned}
\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{D}_1}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{U}_1}\right]_v\right) &= \frac{1}{2} \sum_{i=1}^{M} \text{tr} \left[\mathbf{W}_i^{-1} \frac{\partial \mathbf{W}_i}{\partial [\mathbf{D}_1]_{uu}} \mathbf{W}_i^{-1} \frac{\partial \mathbf{W}_i}{\partial [\mathbf{U}_1]_{rs}} \right] \\
&= \frac{1}{2} \sum_{i=1}^{M} \text{tr} \left[\mathbf{W}_i^{-1} (\mathbf{I}_{n_i} \otimes \mathbf{J}_{uu}) \mathbf{W}_i^{-1} \left(\mathbf{I}_{n_i} \otimes (\boldsymbol{\Lambda}_1 \mathbf{J}_{rs} \mathbf{U}_1' \boldsymbol{\Lambda}_1' + \boldsymbol{\Lambda}_1 \mathbf{U}_1 \mathbf{J}_{sr} \boldsymbol{\Lambda}_1')\right) \right].
\end{aligned}$$

This expression may further be simplified using Proposition 5.10 and hence may be written as

$$\begin{aligned}\mathbf{H}\left(\left[\gamma^{\mathbf{D}_1}\right]_u,\left[\gamma^{\mathbf{U}_1}\right]_v\right) &= \frac{1}{2}\sum_{i=1}^{M} n_i \mathrm{tr}[\mathbf{S}_i\mathbf{J}_{uu}\mathbf{S}_i(\boldsymbol{\Lambda}_1\mathbf{J}_{rs}\mathbf{U}_1'\boldsymbol{\Lambda}_1' + \boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{sr}\boldsymbol{\Lambda}_1') \\ &\quad +(n_i-1)\mathbf{C}_i\mathbf{J}_{uu}\mathbf{C}_i(\boldsymbol{\Lambda}_1\mathbf{J}_{rs}\mathbf{U}_1'\boldsymbol{\Lambda}_1' + \boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{sr}\boldsymbol{\Lambda}_1')] \\ &= \sum_{i=1}^{M} n_i \mathrm{tr}\left[\mathbf{S}_i\mathbf{J}_{uu}\mathbf{S}_i\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{sr}\boldsymbol{\Lambda}_1' + (n_i-1)\mathbf{C}_i\mathbf{J}_{uu}\mathbf{C}_i\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{sr}\boldsymbol{\Lambda}_1'\right]\end{aligned}$$

Using the result $\mathrm{tr}[\mathbf{A}\mathbf{J}_{ij}\mathbf{B}\mathbf{J}_{rs}] = [\mathbf{A}]_{si}[\mathbf{B}]_{jr}$ and (5.6), it follows that

$$\begin{aligned}&\mathbf{H}\left(\left[\gamma^{\mathbf{D}_1}\right]_u,\left[\gamma^{\mathbf{U}_1}\right]_v\right) \\ &= \sum_{i=1}^{M} n_i\left\{[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i]_{su}[\mathbf{S}_i\boldsymbol{\Lambda}_1]_{ur} + (n_i-1)[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{C}_i]_{su}[\mathbf{C}_i\boldsymbol{\Lambda}_1]_{ur}\right\} \\ &= \sum_{i=1}^{M} n_i\left\{[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i]_{su}\left[\mathbf{V}_1^{-1}\boldsymbol{\Lambda}_1\right]_{ur} - [\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i]_{su}[\mathbf{C}_i\boldsymbol{\Lambda}_1]_{ur} + (n_i-1)[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{C}_i]_{su}[\mathbf{C}_i\boldsymbol{\Lambda}_1]_{ur}\right\} \\ &= \sum_{i=1}^{M} n_i\left\{[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i]_{su}\left[\mathbf{V}_1^{-1}\boldsymbol{\Lambda}_1\right]_{ur} - \left[\mathbf{U}_1'\boldsymbol{\Lambda}_1'(\mathbf{V}_1^{-1}-\mathbf{C}_i)\right]_{su}[\mathbf{C}_i\boldsymbol{\Lambda}_1]_{ur} + (n_i-1)[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{C}_i]_{su}[\mathbf{C}_i\boldsymbol{\Lambda}_1]_{ur}\right\} \\ &= \sum_{i=1}^{M} n_i\left\{[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i]_{su}\left[\mathbf{V}_1^{-1}\boldsymbol{\Lambda}_1\right]_{ur} - [\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i}]_{su}[\mathbf{C}_i\boldsymbol{\Lambda}_1]_{ur}\right\} \\ &= \sum_{i=1}^{M} n_i\left\{\left[(\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_i)\otimes(\mathbf{V}_1^{-1}\boldsymbol{\Lambda}_1) - (\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i})\otimes(\mathbf{C}_i\boldsymbol{\Lambda}_1)\right]_{us,ru}\right\}\end{aligned}$$

where we used $[\mathbf{A}]_{ij}[\mathbf{B}]_{rs} = [\mathbf{A}\otimes\mathbf{B}]_{ri,sj}$ to obtain the final expression, which proves the proposition. □

Proposition 5.19

The (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{D}_1},(\boldsymbol{\gamma}^{\mathbf{D}_1})')$ is given by

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{D}_1}\right]_u,\left[\boldsymbol{\gamma}^{\mathbf{D}_1}\right]_v\right)=\frac{1}{2}\sum_{i=1}^{M}n_i\left\{[\mathbf{S}_i]_{uv}^2+(n_i-1)\,[\mathbf{C}_i]_{uv}^2\right\}. \tag{5.29}$$

Proof

It follows from (5.19) that the (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{D}_1},(\boldsymbol{\gamma}^{\mathbf{D}_1})')$ can be written and simplified as

$$\begin{aligned}\mathbf{H}\left([\boldsymbol{\gamma}^{\mathbf{D}_1}]_u,[\boldsymbol{\gamma}^{\mathbf{D}_1}]_v\right) &= \frac{1}{2}\sum_{i=1}^{M}\operatorname{tr}\left[\mathbf{W}_i^{-1}\frac{\partial\mathbf{W}_i}{\partial[\mathbf{D}_1]_{uu}}\mathbf{W}_i^{-1}\frac{\partial\mathbf{W}_i}{\partial[\mathbf{D}_1]_{vv}}\right]\\ &= \frac{1}{2}\sum_{i=1}^{M}\operatorname{tr}\left[\mathbf{W}_i^{-1}\left(\mathbf{I}_{n_i}\otimes\mathbf{J}_{uu}\right)\mathbf{W}_i^{-1}\left(\mathbf{I}_{n_i}\otimes\mathbf{J}_{vv}\right)\right]\end{aligned}$$

and, using Proposition 5.10, it further follows that

$$\begin{aligned}\mathbf{H}\left([\boldsymbol{\gamma}^{\mathbf{D}_1}]_u,[\boldsymbol{\gamma}^{\mathbf{D}_1}]_v\right) &= \frac{1}{2}\sum_{i=1}^{M}n_i\operatorname{tr}\left[\mathbf{S}_i\mathbf{J}_{uu}\mathbf{S}_i\mathbf{J}_{vv}+(n_i-1)\mathbf{C}_i\mathbf{J}_{uu}\mathbf{C}_i\mathbf{J}_{vv}\right]\\ &= \frac{1}{2}\sum_{i=1}^{M}n_i\left\{[\mathbf{S}_i]_{vu}\,[\mathbf{S}_i]_{uv}+(n_i-1)[\mathbf{C}_i]_{vu}[\mathbf{C}_i]_{uv}\right\}\end{aligned}$$

which proves the proposition. □

Proposition 5.20

The (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2},(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1})')$ is given by

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2}\right]_u,\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}\right]_v\right)=\sum_{i=1}^{M}n_i\left\{\left[(\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2)\otimes(\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)\right]_{kr,s\ell}+\left[\mathbf{S}_{n_i}\otimes(\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)\right]_{\ell r,sk}\right\}. \tag{5.30}$$

Proof

Suppose the (k,ℓ)-th element of $\boldsymbol{\Lambda}_2$ corresponds to the u-th element of $\mathrm{vec}(\boldsymbol{\Lambda}_2)$ and the (r,s)-th element of $\boldsymbol{\Lambda}_1$ corresponds to the v-th element of $\mathrm{vec}(\boldsymbol{\Lambda}_1)$. Then, using (5.19), the (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2},(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1})')$ can be written as

$$\begin{aligned}
&\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2}\right]_u,\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}\right]_v\right)\\
&=\frac{1}{2}\sum_{i=1}^{M}\mathrm{tr}\left[\mathbf{W}_i^{-1}\frac{\partial\mathbf{W}_i}{\partial[\boldsymbol{\Lambda}_2]_{k\ell}}\mathbf{W}_i^{-1}\frac{\partial\mathbf{W}_i}{\partial[\boldsymbol{\Lambda}_1]_{rs}}\right]\\
&=\frac{1}{2}\sum_{i=1}^{M}\mathrm{tr}\left[\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}'\otimes(\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2\mathbf{J}_{\ell k}+\mathbf{J}_{k\ell}\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2')\right)\mathbf{W}_i^{-1}\left(\mathbf{I}_{n_i}\otimes(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\mathbf{J}_{sr}+\mathbf{J}_{rs}\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1')\right)\right].
\end{aligned}$$

Using Proposition 5.12, it follows that

$$\begin{aligned}
\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2}\right]_u,\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}\right]_v\right)&=\frac{1}{2}\sum_{i=1}^{M}n_i\mathrm{tr}\left[\mathbf{S}_{n_i}(\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2\mathbf{J}_{\ell k}+\mathbf{J}_{k\ell}\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2')\mathbf{S}_{n_i}(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\mathbf{J}_{sr}+\mathbf{J}_{rs}\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1')\right]\\
&=\frac{1}{2}\sum_{i=1}^{M}n_i\mathrm{tr}\left[\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2\mathbf{J}_{\ell k}\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\mathbf{J}_{sr}+\mathbf{S}_{n_i}\mathbf{J}_{k\ell}\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\mathbf{J}_{sr}\right.\\
&\quad\left.+\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2\mathbf{J}_{\ell k}\mathbf{S}_{n_i}\mathbf{J}_{sr}\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1'+\mathbf{S}_{n_i}\mathbf{J}_{k\ell}\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\mathbf{J}_{rs}\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1'\right]
\end{aligned}$$

$$= \sum_{i=1}^{M} n_i \text{tr}\left[\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2\mathbf{J}_{\ell k}\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\mathbf{J}_{sr} + \mathbf{S}_{n_i}\mathbf{J}_{k\ell}\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\mathbf{J}_{sr}\right]$$

$$= \sum_{i=1}^{M} n_i \left\{[\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2]_{r\ell}[\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1]_{ks} + [\mathbf{S}_{n_i}]_{rk}[\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1]_{\ell s}\right\}$$

$$= \sum_{i=1}^{M} n_i \left\{[(\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2)\otimes(\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)]_{kr,s\ell} + [\mathbf{S}_{n_i}\otimes(\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)]_{\ell r,sk}\right\}$$

which proves the proposition. □

Proposition 5.21

The (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2},(\boldsymbol{\gamma}^{\mathbf{U}_1})')$ is given by

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2}\right]_u,\left[\boldsymbol{\gamma}^{\mathbf{U}_1}\right]_v\right) = \sum_{i=1}^{M} n_i \left\{[(\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2)\otimes(\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1)]_{ks,r\ell} + [(\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i})\otimes(\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1)]_{\ell s,rk}\right. \tag{5.31}$$

Proof

Suppose the (k,ℓ)-th element of $\boldsymbol{\Lambda}_2$ corresponds to the u-th element of $\text{vec}(\boldsymbol{\Lambda}_2)$ and the (r,s)-th element of $\mathbf{U}_1$ corresponds to the v-th element of $\text{vecs}(\mathbf{U}_1)$. Then, using (5.19), the (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2},(\boldsymbol{\gamma}^{\mathbf{U}_1})')$ can be written as

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2}\right]_u,\left[\boldsymbol{\gamma}^{\mathbf{U}_1}\right]_v\right)$$

$$= \frac{1}{2}\sum_{i=1}^{M} \text{tr}\left[\mathbf{W}_i^{-1}\frac{\partial\mathbf{W}_i}{\partial[\boldsymbol{\Lambda}_2]_{k\ell}}\mathbf{W}_i^{-1}\frac{\partial\mathbf{W}_i}{\partial[\mathbf{U}_1]_{rs}}\right]$$

$$= \frac{1}{2}\sum_{i=1}^{M} \text{tr}\left[\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}'\otimes(\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2\mathbf{J}_{\ell k} + \mathbf{J}_{k\ell}\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2')\right)\mathbf{W}_i^{-1}(\mathbf{I}_{n_i}\otimes(\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{sr}\boldsymbol{\Lambda}_1' + \boldsymbol{\Lambda}_1\mathbf{J}_{rs}\mathbf{U}_1'\boldsymbol{\Lambda}_1'))\right].$$

Using Proposition 5.12, it follows that

$$\begin{aligned}
&\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{U}_1}\right]_v\right) \\
&= \sum_{i=1}^{M} n_i \mathrm{tr}\left[\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2\mathbf{J}_{\ell k}\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{sr}\boldsymbol{\Lambda}_1' + \mathbf{S}_{n_i}\mathbf{J}_{k\ell}\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{sr}\boldsymbol{\Lambda}_1'\right] \\
&= \sum_{i=1}^{M} n_i \left\{[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2]_{s\ell}[\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1]_{kr} + [\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i}]_{sk}[\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1]_{\ell r}\right\} \\
&= \sum_{i=1}^{M} n_i \left\{[(\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2) \otimes (\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1)]_{ks,r\ell} + [(\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i}) \otimes (\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1)]_{\ell s,rk}\right\}
\end{aligned}$$

which proves the proposition. □

Proposition 5.22

The (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2}, (\boldsymbol{\gamma}^{\mathbf{D}_1})')$ is given by

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{D}_1}\right]_v\right) = \sum_{i=1}^{M} n_i \left[(\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2) \otimes \mathbf{S}_{n_i}\right]_{kv,v\ell}. \tag{5.32}$$

Proof

Let us consider the submatrix $\mathbf{H}(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2}, (\boldsymbol{\gamma}^{\mathbf{D}_1})')$ of $\mathbf{H}(\boldsymbol{\gamma})$. Suppose the (k,ℓ)-th element of $\boldsymbol{\Lambda}_2$ corresponds to the u-th element of $\mathrm{vec}(\boldsymbol{\Lambda}_2)$. The (u,v)-th element of the submatrix can now be written as (see (5.19))

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{D}_1}\right]_v\right) = \frac{1}{2}\sum_{i=1}^{M} \mathrm{tr}\left[\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\boldsymbol{\Lambda}_2]_{k\ell}}\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{D}_1]_{vv}}\right]$$

$$= \frac{1}{2}\sum_{i=1}^{M} \text{tr}\left[\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes (\mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{J}_{\ell k} + \mathbf{J}_{k\ell}\mathbf{\Phi}_2\mathbf{\Lambda}_2')\right)\mathbf{W}_i^{-1}(\mathbf{I}_{n_i} \otimes \mathbf{J}_{vv})\right]$$

and using Proposition 5.12, this simplifies to

$$\begin{aligned}
\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{\Lambda}_2}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{D}_1}\right]_v\right) &= \frac{1}{2}\sum_{i=1}^{M} n_i \text{tr}\left[\mathbf{S}_{n_i}(\mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{J}_{\ell k} + \mathbf{J}_{k\ell}\mathbf{\Phi}_2\mathbf{\Lambda}_2')\mathbf{S}_{n_i}\mathbf{J}_{vv}\right] \\
&= \sum_{i=1}^{M} n_i \text{tr}\left[\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{J}_{\ell k}\mathbf{S}_{n_i}\mathbf{J}_{vv}\right] \\
&= \sum_{i=1}^{M} n_i \left[\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2\right]_{v\ell}\left[\mathbf{S}_{n_i}\right]_{kv} \\
&= \sum_{i=1}^{M} n_i \left[(\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2) \otimes \mathbf{S}_{n_i}\right]_{kv,v\ell}
\end{aligned}$$

which proves the proposition. □

Proposition 5.23

The (u, v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{\Lambda}_2}, (\boldsymbol{\gamma}^{\mathbf{\Lambda}_2})')$ is given by

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{\Lambda}_2}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{\Lambda}_2}\right]_v\right) = \sum_{i=1}^{M} n_i^2 \left\{\left[(\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2) \otimes (\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2)\right]_{kr,s\ell} + \left[\mathbf{S}_{n_i} \otimes (\mathbf{\Phi}_2\mathbf{\Lambda}_2'\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2)\right]_{\ell r,sk}\right\}. \tag{5.33}$$

Proof

Consider the (u, v)-th element of the submatrix $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{\Lambda}_2}, (\boldsymbol{\gamma}^{\mathbf{\Lambda}_2})')$ where the u-th and v-th elements of $\text{vec}(\mathbf{\Lambda}_2)$ correspond to the (k, ℓ)-th and (r, s)-th elements of $\mathbf{\Lambda}_2$ respectively.

Consequently (5.19) can be used to write

$$\begin{aligned}
&\mathbf{H}\left(\left[\gamma^{\mathbf{\Lambda}_2}\right]_u, \left[\gamma^{\mathbf{\Lambda}_2}\right]_v\right) \\
&= \frac{1}{2}\sum_{i=1}^{M} \mathrm{tr}\left[\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{\Lambda}_2]_{k\ell}}\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{\Lambda}_2]_{rs}}\right] \\
&= \frac{1}{2}\sum_{i=1}^{M} \mathrm{tr}\left[\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes (\mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{J}_{\ell k} + \mathbf{J}_{k\ell}\mathbf{\Phi}_2\mathbf{\Lambda}_2')\right)\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes (\mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{J}_{sr} + \mathbf{J}_{rs}\mathbf{\Phi}_2\mathbf{\Lambda}_2')\right)\right].
\end{aligned}$$

From Proposition 5.11 it follows that the above equation may be simplified further, so that

$$\begin{aligned}
\mathbf{H}\left(\left[\gamma^{\mathbf{\Lambda}_2}\right]_u, \left[\gamma^{\mathbf{\Lambda}_2}\right]_v\right) &= \frac{1}{2}\sum_{i=1}^{M} n_i^2 \mathrm{tr}\left[\mathbf{S}_{n_i}(\mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{J}_{\ell k} + \mathbf{J}_{k\ell}\mathbf{\Phi}_2\mathbf{\Lambda}_2')\mathbf{S}_{n_i}(\mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{J}_{sr} + \mathbf{J}_{rs}\mathbf{\Phi}_2\mathbf{\Lambda}_2')\right] \\
&= \frac{1}{2}\sum_{i=1}^{M} n_i^2 \mathrm{tr}\left[\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{J}_{\ell k}\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{J}_{sr} + \mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{J}_{\ell k}\mathbf{S}_{n_i}\mathbf{J}_{rs}\mathbf{\Phi}_2\mathbf{\Lambda}_2'\right. \\
&\quad \left. + \mathbf{S}_{n_i}\mathbf{J}_{k\ell}\mathbf{\Phi}_2\mathbf{\Lambda}_2'\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{J}_{sr} + \mathbf{S}_{n_i}\mathbf{J}_{k\ell}\mathbf{\Phi}_2\mathbf{\Lambda}_2'\mathbf{S}_{n_i}\mathbf{J}_{rs}\mathbf{\Phi}_2\mathbf{\Lambda}_2'\right] \\
&= \sum_{i=1}^{M} n_i^2 \mathrm{tr}\left[\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{J}_{\ell k}\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{J}_{sr} + \mathbf{S}_{n_i}\mathbf{J}_{k\ell}\mathbf{\Phi}_2\mathbf{\Lambda}_2'\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{J}_{sr}\right] \\
&= \sum_{i=1}^{M} n_i^2 \left\{[\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2]_{r\ell}[\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2]_{ks} + [\mathbf{S}_{n_i}]_{rk}[\mathbf{\Phi}_2\mathbf{\Lambda}_2'\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2]_{\ell s}\right\} \\
&= \sum_{i=1}^{M} n_i^2 \left\{[(\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2) \otimes (\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2)]_{kr,s\ell} + [\mathbf{S}_{n_i} \otimes (\mathbf{\Phi}_2\mathbf{\Lambda}_2'\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2)]_{\ell r,sk}\right\}
\end{aligned}$$

which proves the proposition. □

Proposition 5.24

The (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{U}_2},(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1})')$ is given by

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{U}_2}\right]_u,\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}\right]_v\right)=\sum_{i=1}^{M}n_i\left\{[(\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2)\otimes(\mathbf{U}_2'\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)]_{\ell r,sk}+[(\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\mathbf{U}_2)\otimes(\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)]_{kr,s\ell}\right. \tag{5.34}$$

Proof

Suppose the (k,ℓ)-th element of $\mathbf{U}_2$ corresponds to the u-th element of vecs$(\mathbf{U}_2)$ and the (r,s)-th element of $\boldsymbol{\Lambda}_1$ corresponds to the v-th element of vec$(\boldsymbol{\Lambda}_1)$. Then, using (5.19), the (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{U}_2},(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1})')$ can be written as

$$\begin{aligned}
&\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{U}_2}\right]_u,\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}\right]_v\right)\\
=\;&\frac{1}{2}\sum_{i=1}^{M}\mathrm{tr}\left[\mathbf{W}_i^{-1}\frac{\partial\mathbf{W}_i}{\partial[\mathbf{U}_2]_{k\ell}}\mathbf{W}_i^{-1}\frac{\partial\mathbf{W}_i}{\partial[\boldsymbol{\Lambda}_1]_{rs}}\right]\\
=\;&\frac{1}{2}\sum_{i=1}^{M}\mathrm{tr}\left[\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}'\otimes(\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_2'+\boldsymbol{\Lambda}_2\mathbf{J}_{k\ell}\mathbf{U}_2'\boldsymbol{\Lambda}_2')\right)\mathbf{W}_i^{-1}\left(\mathbf{I}_{n_i}\otimes(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\mathbf{J}_{sr}+\mathbf{J}_{rs}\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1')\right)\right].
\end{aligned}$$

Using Proposition 5.12, it follows that

$$\begin{aligned}
&\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{U}_2}\right]_u,\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}\right]_v\right)\\
=\;&\frac{1}{2}\sum_{i=1}^{M}n_i\mathrm{tr}\left[\mathbf{S}_{n_i}(\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_2'+\boldsymbol{\Lambda}_2\mathbf{J}_{k\ell}\mathbf{U}_2'\boldsymbol{\Lambda}_2')\mathbf{S}_{n_i}(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\mathbf{J}_{sr}+\mathbf{J}_{rs}\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1')\right]\\
=\;&\sum_{i=1}^{M}n_i\mathrm{tr}\left[\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\mathbf{J}_{sr}+\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\mathbf{J}_{k\ell}\mathbf{U}_2'\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\mathbf{J}_{sr}\right]
\end{aligned}$$

$$= \sum_{i=1}^{M} n_i \{[\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\mathbf{U}_2]_{r\ell}[\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1]_{ks} + [\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2]_{rk}[\mathbf{U}_2'\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1]_{\ell s}\}$$

$$= \sum_{i=1}^{M} n_i \{[(\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\mathbf{U}_2) \otimes (\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)]_{kr,s\ell} + [(\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2) \otimes (\mathbf{U}_2'\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)]_{\ell r,sk}\}$$

which proves the proposition. □

Proposition 5.25

The (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{U}_2}, (\boldsymbol{\gamma}^{\mathbf{U}_1})')$ is given by

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{U}_2}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{U}_1}\right]_v\right) = \sum_{i=1}^{M} n_i \left\{[(\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2) \otimes (\mathbf{U}_2'\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1)]_{\ell s,rk} + [(\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\mathbf{U}_2) \otimes (\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1)]_{ks,r\ell}\right\}. \quad (5.35)$$

Proof

Suppose the (k,ℓ)-th element of $\mathbf{U}_2$ corresponds to the u-th element of $\text{vecs}(\mathbf{U}_2)$ and the (r,s)-th element of $\mathbf{U}_1$ corresponds to the v-th element of $\text{vecs}(\mathbf{U}_1)$. Then, using (5.19), the (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{U}_2}, (\boldsymbol{\gamma}^{\mathbf{U}_1})')$ can be written as

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{U}_2}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{U}_1}\right]_v\right)$$

$$= \frac{1}{2}\sum_{i=1}^{M} \text{tr}\left[\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{U}_2]_{k\ell}}\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{U}_1]_{rs}}\right]$$

$$= \frac{1}{2}\sum_{i=1}^{M} \text{tr}\left[\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes (\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_2' + \boldsymbol{\Lambda}_2\mathbf{J}_{k\ell}\mathbf{U}_2'\boldsymbol{\Lambda}_2')\right)\mathbf{W}_i^{-1}\left(\mathbf{I}_{n_i} \otimes (\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{sr}\boldsymbol{\Lambda}_1' + \boldsymbol{\Lambda}_1\mathbf{J}_{rs}\mathbf{U}_1'\boldsymbol{\Lambda}_1')\right)\right].$$

Using Proposition 5.12, it follows that

$$
\begin{aligned}
& \mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{U}_2}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{U}_1}\right]_v\right) \\
= & \frac{1}{2}\sum_{i=1}^{M} n_i \mathrm{tr}\left[\mathbf{S}_{n_i}(\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_2' + \boldsymbol{\Lambda}_2\mathbf{J}_{k\ell}\mathbf{U}_2'\boldsymbol{\Lambda}_2')\mathbf{S}_{n_i}(\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{sr}\boldsymbol{\Lambda}_1' + \boldsymbol{\Lambda}_1\mathbf{J}_{rs}\mathbf{U}_1'\boldsymbol{\Lambda}_1')\right] \\
= & \sum_{i=1}^{M} n_i \mathrm{tr}\left[\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{sr}\boldsymbol{\Lambda}_1' + \mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\mathbf{J}_{k\ell}\mathbf{U}_2'\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{sr}\boldsymbol{\Lambda}_1'\right] \\
= & \sum_{i=1}^{M} n_i \left\{[\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\mathbf{U}_2]_{s\ell}[\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1]_{kr} + [\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2]_{sk}[\mathbf{U}_2'\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1]_{\ell r}\right\} \\
= & \sum_{i=1}^{M} n_i \left\{[(\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\mathbf{U}_2)\otimes(\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1)]_{ks,r\ell} + [(\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2)\otimes(\mathbf{U}_2'\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1)]_{\ell s,rk}\right\}
\end{aligned}
$$

which proves the proposition. □

Proposition 5.26

The (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{U}_2}, (\boldsymbol{\gamma}^{\mathbf{D}_1})')$ is given by

$$
\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{U}_2}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{D}_1}\right]_v\right) = \sum_{i=1}^{M} n_i \left[(\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\mathbf{U}_2)\otimes(\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i})\right]_{kv,v\ell}. \tag{5.36}
$$

Proof

Let us consider the submatrix $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{U}_2}, (\boldsymbol{\gamma}^{\mathbf{D}_1})')$ of $\mathbf{H}(\boldsymbol{\gamma})$. Suppose the (k,ℓ)-th element of $\mathbf{U}_2$ corresponds to the u-th element of $\mathrm{vecs}(\mathbf{U}_2)$. The (u,v)-th element of the submatrix can now be written as (see (5.19))

$$\begin{aligned}\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{U}_2}\right]_u,\left[\boldsymbol{\gamma}^{\mathbf{D}_1}\right]_v\right) &= \frac{1}{2}\sum_{i=1}^{M}\mathrm{tr}\left[\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{U}_2]_{k\ell}}\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{D}_1]_{vv}}\right] \\ &= \frac{1}{2}\sum_{i=1}^{M}\mathrm{tr}\left[\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}'\otimes(\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_2'+\boldsymbol{\Lambda}_2\mathbf{J}_{k\ell}\mathbf{U}_2'\boldsymbol{\Lambda}_2')\right)\mathbf{W}_i^{-1}(\mathbf{I}_{n_i}\otimes\mathbf{J}_{vv})\right]\end{aligned}$$

and using Proposition 5.12, this simplifies to

$$\begin{aligned}\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{U}_2}\right]_u,\left[\boldsymbol{\gamma}^{\mathbf{D}_1}\right]_v\right) &= \frac{1}{2}\sum_{i=1}^{M} n_i\mathrm{tr}\left[\mathbf{S}_{n_i}(\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_2'+\boldsymbol{\Lambda}_2\mathbf{J}_{k\ell}\mathbf{U}_2'\boldsymbol{\Lambda}_2')\mathbf{S}_{n_i}\mathbf{J}_{vv}\right] \\ &= \sum_{i=1}^{M} n_i\mathrm{tr}\left[\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\mathbf{J}_{vv}\right] \\ &= \sum_{i=1}^{M} n_i\left[\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\mathbf{U}_2\right]_{v\ell}\left[\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\right]_{kv} \\ &= \sum_{i=1}^{M} n_i\left[(\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\mathbf{U}_2)\otimes(\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i})\right]_{kv,v\ell}\end{aligned}$$

which proves the proposition. □

Proposition 5.27

The (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{U}_2},(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2})')$ is given by

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{U}_2}\right]_u,\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2}\right]_v\right)=\sum_{i=1}^{M} n_i^2\left\{\left[(\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2)\otimes(\mathbf{U}_2'\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2)\right]_{\ell r,sk}+\left[(\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\mathbf{U}_2)\otimes(\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2)\right]_{kr,s\ell}\right\} \tag{5.37}$$

Proof

Consider the (u,v)-th element of the submatrix $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{U}_2},(\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2})')$ where the u-th element of vecs$(\mathbf{U}_2)$ corresponds to the (k,ℓ)-th element of $\mathbf{U}_2$ and the v-th element of $\boldsymbol{\Lambda}_2$ corresponds to the (r,s)-th elements of $\boldsymbol{\Lambda}_2$. Consequently (5.19) can be used to write

$$\begin{aligned}
&\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{U}_2}\right]_u,\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2}\right]_v\right)\\
=\;&\frac{1}{2}\sum_{i=1}^{M}\mathrm{tr}\left[\mathbf{W}_i^{-1}\frac{\partial\mathbf{W}_i}{\partial[\mathbf{U}_2]_{k\ell}}\mathbf{W}_i^{-1}\frac{\partial\mathbf{W}_i}{\partial[\boldsymbol{\Lambda}_2]_{rs}}\right]\\
=\;&\frac{1}{2}\sum_{i=1}^{M}\mathrm{tr}\left[\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}'\otimes(\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_2'+\boldsymbol{\Lambda}_2\mathbf{J}_{k\ell}\mathbf{U}_2'\boldsymbol{\Lambda}_2')\right)\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}'\otimes(\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2\mathbf{J}_{sr}+\mathbf{J}_{rs}\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2')\right)\right]
\end{aligned}$$

From Proposition 5.11 it follows that the above equation may be simplified further, so that

$$\begin{aligned}
&\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{U}_2}\right]_u,\left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2}\right]_v\right)\\
=\;&\frac{1}{2}\sum_{i=1}^{M}n_i^2\mathrm{tr}\left[\mathbf{S}_{n_i}(\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_2'+\boldsymbol{\Lambda}_2\mathbf{J}_{k\ell}\mathbf{U}_2'\boldsymbol{\Lambda}_2')\mathbf{S}_{n_i}(\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2\mathbf{J}_{sr}+\mathbf{J}_{rs}\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2')\right]\\
=\;&\sum_{i=1}^{M}n_i^2\mathrm{tr}\left[\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2\mathbf{J}_{sr}+\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\mathbf{J}_{k\ell}\mathbf{U}_2'\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2\mathbf{J}_{sr}\right]\\
=\;&\sum_{i=1}^{M}n_i^2\left\{[\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2]_{rk}[\mathbf{U}_2'\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2]_{\ell s}+[\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\mathbf{U}_2]_{r\ell}[\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2]_{ks}\right\}\\
=\;&\sum_{i=1}^{M}n_i^2\left\{[(\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2)\otimes(\mathbf{U}_2'\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2)]_{\ell r,sk}+[(\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\mathbf{U}_2)\otimes(\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2)]_{kr,s\ell}\right\}
\end{aligned}$$

which proves the proposition. □

Proposition 5.28

The (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{U}_2},(\boldsymbol{\gamma}^{\mathbf{U}_2})')$ is given by

$$\begin{aligned}\mathbf{H}\left([\boldsymbol{\gamma}^{\mathbf{U}_2}]_u,[\boldsymbol{\gamma}^{\mathbf{U}_2}]_v\right) &= \sum_{i=1}^{M} n_i^2 \left\{[(\mathbf{U}_2'\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\mathbf{U}_2)\otimes(\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2)]_{ks,r\ell}\right.\\ &\quad \left.+[(\mathbf{U}_2'\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2)\otimes(\mathbf{U}_2'\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2)]_{\ell s,rk}\right\}. \qquad (5.38)\end{aligned}$$

Proof

Consider the (u,v)-th element of the submatrix $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{U}_2},(\boldsymbol{\gamma}^{\mathbf{U}_2})')$ where the u-th and v-th element of vecs$(\mathbf{U}_2)$ respectively correspond to the (k,ℓ)-th and (r,s)-th elements of $\mathbf{U}_2$. Consequently (5.19) can be used to write

$$\begin{aligned}&\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{U}_2}\right]_u,\left[\boldsymbol{\gamma}^{\mathbf{U}_2}\right]_v\right)\\ &= \frac{1}{2}\sum_{i=1}^{M}\mathrm{tr}\left[\mathbf{W}_i^{-1}\frac{\partial\mathbf{W}_i}{\partial[\mathbf{U}_2]_{k\ell}}\mathbf{W}_i^{-1}\frac{\partial\mathbf{W}_i}{\partial[\mathbf{U}_2]_{rs}}\right]\\ &= \frac{1}{2}\sum_{i=1}^{M}\mathrm{tr}\left[\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}'\otimes(\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_2'+\boldsymbol{\Lambda}_2\mathbf{J}_{k\ell}\mathbf{U}_2'\boldsymbol{\Lambda}_2')\right)\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}'\otimes(\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{sr}\boldsymbol{\Lambda}_2'+\boldsymbol{\Lambda}_2\mathbf{J}_{rs}\mathbf{U}_2'\boldsymbol{\Lambda}_2')\right)\right].\end{aligned}$$

From Proposition 5.11 it follows that the above equation may be simplified further, so that

$$\begin{aligned}&\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{U}_2}\right]_u,\left[\boldsymbol{\gamma}^{\mathbf{U}_2}\right]_v\right)\\ &= \frac{1}{2}\sum_{i=1}^{M}n_i^2\mathrm{tr}\left[\mathbf{S}_{n_i}(\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{\ell k}\boldsymbol{\Lambda}_2'+\boldsymbol{\Lambda}_2\mathbf{J}_{k\ell}\mathbf{U}_2'\boldsymbol{\Lambda}_2')\mathbf{S}_{n_i}(\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{sr}\boldsymbol{\Lambda}_2'+\boldsymbol{\Lambda}_2\mathbf{J}_{rs}\mathbf{U}_2'\boldsymbol{\Lambda}_2')\right]\end{aligned}$$

$$= \sum_{i=1}^{M} n_i^2 \mathrm{tr}\left[\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{U}_2\mathbf{J}_{\ell k}\mathbf{\Lambda}_2'\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{U}_2\mathbf{J}_{sr}\mathbf{\Lambda}_2' + \mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{J}_{k\ell}\mathbf{U}_2'\mathbf{\Lambda}_2'\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{U}_2\mathbf{J}_{sr}\mathbf{\Lambda}_2'\right]$$

$$= \sum_{i=1}^{M} n_i^2 \left\{[\mathbf{U}_2'\mathbf{\Lambda}_2'\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{U}_2]_{s\ell}[\mathbf{\Lambda}_2'\mathbf{S}_{n_i}\mathbf{\Lambda}_2]_{kr} + [\mathbf{U}_2'\mathbf{\Lambda}_2'\mathbf{S}_{n_i}\mathbf{\Lambda}_2]_{sk}[\mathbf{U}_2'\mathbf{\Lambda}_2'\mathbf{S}_{n_i}\mathbf{\Lambda}_2]_{\ell r}\right\}$$

$$= \sum_{i=1}^{M} n_i^2 \left\{[(\mathbf{U}_2'\mathbf{\Lambda}_2'\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{U}_2)\otimes(\mathbf{\Lambda}_2'\mathbf{S}_{n_i}\mathbf{\Lambda}_2)]_{ks,r\ell} + [(\mathbf{U}_2'\mathbf{\Lambda}_2'\mathbf{S}_{n_i}\mathbf{\Lambda}_2)\otimes(\mathbf{U}_2'\mathbf{\Lambda}_2'\mathbf{S}_{n_i}\mathbf{\Lambda}_2)]_{\ell s,rk}\right\}$$

which proves the proposition. □

Proposition 5.29

The (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{D}_2},(\boldsymbol{\gamma}^{\mathbf{\Lambda}_1})')$ is given by

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{D}_2}\right]_u,\left[\boldsymbol{\gamma}^{\mathbf{\Lambda}_1}\right]_v\right) = \sum_{i=1}^{M} n_i\left[\mathbf{S}_{n_i}\otimes(\mathbf{S}_{n_i}\mathbf{\Lambda}_1\mathbf{\Phi}_1)\right]_{ur,su}. \tag{5.39}$$

Proof

Assume that the (r,s)-th element of $\mathbf{\Lambda}_1$ corresponds to the v-th element of $\mathrm{vec}(\mathbf{\Lambda}_1)$. Then, according to (5.19), we have the (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{D}_2},(\boldsymbol{\gamma}^{\mathbf{\Lambda}_1})')$ given by

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{D}_2}\right]_u,\left[\boldsymbol{\gamma}^{\mathbf{\Lambda}_1}\right]_v\right) = \frac{1}{2}\sum_{i=1}^{M}\mathrm{tr}\left[\mathbf{W}_i^{-1}\frac{\partial\mathbf{W}_i}{\partial[\mathbf{D}_2]_{uu}}\mathbf{W}_i^{-1}\frac{\partial\mathbf{W}_i}{\partial[\mathbf{\Lambda}_1]_{rs}}\right]$$

$$= \frac{1}{2}\sum_{i=1}^{M}\mathrm{tr}\left[\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}'\otimes\mathbf{J}_{uu}\right)\mathbf{W}_i^{-1}\left(\mathbf{I}_{n_i}\otimes(\mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{J}_{sr}+\mathbf{J}_{rs}\mathbf{\Phi}_1\mathbf{\Lambda}_1')\right)\right]$$

which simplifies further, using Proposition 5.12, to

$$\begin{aligned}
\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{D}_2}\right]_u, \left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_1}\right]_v\right) &= \frac{1}{2}\sum_{i=1}^{M} n_i \mathrm{tr}\left[\mathbf{S}_{n_i}\mathbf{J}_{uu}\mathbf{S}_{n_i}(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\mathbf{J}_{sr} + \mathbf{J}_{rs}\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1')\right] \\
&= \sum_{i=1}^{M} n_i \mathrm{tr}\left[\mathbf{S}_{n_i}\mathbf{J}_{uu}\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\mathbf{J}_{sr}\right] \\
&= \sum_{i=1}^{M} n_i [\mathbf{S}_{n_i}]_{ru}[\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1]_{us} \\
&= \sum_{i=1}^{M} n_i [\mathbf{S}_{n_i} \otimes (\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)]_{ur,su}
\end{aligned}$$

which proves the proposition. □

Proposition 5.30

The (u, v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{D}_2}, (\boldsymbol{\gamma}^{\mathbf{U}_1})')$ is given by

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{D}_2}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{U}_1}\right]_v\right) = \sum_{i=1}^{M} n_i \left[(\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i} \otimes (\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1)\right]_{us,ru}. \tag{5.40}$$

Proof

Assume that the (r, s)-th element of $\mathbf{U}_1$ corresponds to the v-th element of vecs$(\mathbf{U}_1)$. Then, according to (5.19), we have the (u, v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{D}_2}, (\boldsymbol{\gamma}^{\mathbf{U}_1})')$ given by

$$\begin{aligned}
\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{D}_2}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{U}_1}\right]_v\right) &= \frac{1}{2}\sum_{i=1}^{M} \mathrm{tr}\left[\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{D}_2]_{uu}}\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{U}_1]_{rs}}\right] \\
&= \frac{1}{2}\sum_{i=1}^{M} \mathrm{tr}\left[\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes \mathbf{J}_{uu}\right)\mathbf{W}_i^{-1}\left(\mathbf{I}_{n_i} \otimes (\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{sr}\boldsymbol{\Lambda}_1' + \boldsymbol{\Lambda}_1\mathbf{J}_{rs}\mathbf{U}_1'\boldsymbol{\Lambda}_1')\right)\right]
\end{aligned}$$

which simplifies further, using Proposition 5.12, to

$$\begin{aligned}
\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{D}_2}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{U}_1}\right]_v\right) &= \frac{1}{2}\sum_{i=1}^{M} n_i \mathrm{tr}\left[\mathbf{S}_{n_i}\mathbf{J}_{uu}\mathbf{S}_{n_i}(\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{sr}\boldsymbol{\Lambda}_1' + \boldsymbol{\Lambda}_1\mathbf{J}_{rs}\mathbf{U}_1'\boldsymbol{\Lambda}_1')\right] \\
&= \sum_{i=1}^{M} n_i \mathrm{tr}\left[\mathbf{S}_{n_i}\mathbf{J}_{uu}\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1\mathbf{U}_1\mathbf{J}_{sr}\boldsymbol{\Lambda}_1'\right] \\
&= \sum_{i=1}^{M} n_i [\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i}]_{su}[\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1]_{ur} \\
&= \sum_{i=1}^{M} n_i [(\mathbf{U}_1'\boldsymbol{\Lambda}_1'\mathbf{S}_{n_i}) \otimes (\mathbf{S}_{n_i}\boldsymbol{\Lambda}_1)]_{us,ru}
\end{aligned}$$

which proves the proposition. □

Proposition 5.31

The (u, v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{D}_2}, (\boldsymbol{\gamma}^{\mathbf{D}_1})')$ is given by

$$\mathbf{H}\left([\boldsymbol{\gamma}^{\mathbf{D}_2}]_u, [\boldsymbol{\gamma}^{\mathbf{D}_1}]_v\right) = \frac{1}{2}\sum_{i=1}^{M} n_i \left[\mathbf{S}_{n_i}\right]_{uv}^2 . \tag{5.41}$$

Proof

The (u, v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{D}_2}, (\boldsymbol{\gamma}^{\mathbf{D}_1})')$ follows from (5.19) as

$$\begin{aligned}
\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{D}_2}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{D}_1}\right]_v\right) &= \frac{1}{2}\sum_{i=1}^{M} \mathrm{tr}\left[\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{D}_2]_{uu}}\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{D}_1]_{vv}}\right] \\
&= \frac{1}{2}\sum_{i=1}^{M} \mathrm{tr}\left[\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes \mathbf{J}_{uu}\right)\mathbf{W}_i^{-1}\left(\mathbf{I}_{n_i} \otimes \mathbf{J}_{vv}\right)\right]
\end{aligned}$$

and the use of Proposition 5.12 leads to the further simplification

$$\begin{aligned}\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{D}_2}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{D}_1}\right]_v\right) &= \frac{1}{2}\sum_{i=1}^{M} n_i \mathrm{tr}\left[\mathbf{S}_{n_i}\mathbf{J}_{uu}\mathbf{S}_{n_i}\mathbf{J}_{vv}\right] \\ &= \frac{1}{2}\sum_{i=1}^{M} n_i \left[\mathbf{S}_{n_i}\right]_{uv}^2\end{aligned}$$

which proves the proposition. □

Proposition 5.32

The (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{D}_2}, (\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2})')$ is given by

$$\mathbf{H}([\boldsymbol{\gamma}^{\mathbf{D}_2}]_u, [\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2}]_v) = \sum_{i=1}^{M} n_i^2 \left[\mathbf{S}_{n_i} \otimes (\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2)\right]_{ur,su}. \tag{5.42}$$

Proof

Suppose that the (r,s)-th element of $\boldsymbol{\Lambda}_2$ corresponds to the v-th element of $\mathrm{vec}(\boldsymbol{\Lambda}_2)$. Then the (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{D}_2}, (\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2})')$ can, by using (5.19), be written as

$$\begin{aligned}\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{D}_2}\right]_u, \left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2}\right]_v\right) &= \frac{1}{2}\sum_{i=1}^{M} \mathrm{tr}\left[\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{D}_2]_{uu}}\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\boldsymbol{\Lambda}_2]_{rs}}\right] \\ &= \frac{1}{2}\sum_{i=1}^{M} \mathrm{tr}\left[\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes \mathbf{J}_{uu}\right)\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes (\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2\mathbf{J}_{sr} + \mathbf{J}_{rs}\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2')\right)\right]\end{aligned}$$

and, using Proposition 5.11, it now follows that

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{D}_2}\right]_u, \left[\boldsymbol{\gamma}^{\boldsymbol{\Lambda}_2}\right]_v\right) = \frac{1}{2}\sum_{i=1}^{M} n_i^2 \mathrm{tr}\left[\mathbf{S}_{n_i}\mathbf{J}_{uu}\mathbf{S}_{n_i}(\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2\mathbf{J}_{sr} + \mathbf{J}_{rs}\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2')\right]$$

$$= \sum_{i=1}^{M} n_i^2 \mathrm{tr}\left[\mathbf{S}_{n_i}\mathbf{J}_{uu}\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2\mathbf{J}_{sr}\right]$$

$$= \sum_{i=1}^{M} n_i^2 \left[\mathbf{S}_{n_i}\right]_{ru}\left[\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2\right]_{us}$$

$$= \sum_{i=1}^{M} n_i^2 \left[\mathbf{S}_{n_i} \otimes (\mathbf{S}_{n_i}\mathbf{\Lambda}_2\mathbf{\Phi}_2)\right]_{ur,su}$$

which proves the proposition. □

Proposition 5.33

The (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{D}_2}, (\boldsymbol{\gamma}^{\mathbf{U}_2})')$ is given by

$$\mathbf{H}([\boldsymbol{\gamma}^{\mathbf{D}_2}]_u, [\boldsymbol{\gamma}^{\mathbf{U}_2}]_v) = \sum_{i=1}^{M} n_i^2 \left[(\mathbf{U}_2'\mathbf{\Lambda}_2'\mathbf{S}_{n_i}) \otimes (\mathbf{S}_{n_i}\mathbf{\Lambda}_2)\right]_{us,ru}. \tag{5.43}$$

Proof

Suppose that the (r,s)-th element of $\mathbf{U}_2$ corresponds to the v-th element of $\mathrm{vecs}(\mathbf{U}_2)$. Then the (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{D}_2}, (\boldsymbol{\gamma}^{\mathbf{U}_2})')$ can, by using (5.19), be written as

$$\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{D}_2}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{U}_2}\right]_v\right)$$

$$= \frac{1}{2}\sum_{i=1}^{M} \mathrm{tr}\left[\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{D}_2]_{uu}}\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{U}_2]_{rs}}\right]$$

$$= \frac{1}{2}\sum_{i=1}^{M} \mathrm{tr}\left[\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes \mathbf{J}_{uu}\right)\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes (\mathbf{\Lambda}_2\mathbf{U}_2\mathbf{J}_{sr}\mathbf{\Lambda}_2' + \mathbf{\Lambda}_2\mathbf{J}_{rs}\mathbf{U}_2'\mathbf{\Lambda}_2')\right)\right]$$

and, using Proposition 5.11, it now follows that

$$\begin{aligned}
\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{D}_2}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{U}_2}\right]_v\right) &= \frac{1}{2}\sum_{i=1}^{M} n_i^2 \mathrm{tr}\left[\mathbf{S}_{n_i}\mathbf{J}_{uu}\mathbf{S}_{n_i}(\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{sr}\boldsymbol{\Lambda}_2' + \boldsymbol{\Lambda}_2\mathbf{J}_{rs}\mathbf{U}_2'\boldsymbol{\Lambda}_2')\right] \\
&= \sum_{i=1}^{M} n_i^2 \mathrm{tr}\left[\mathbf{S}_{n_i}\mathbf{J}_{uu}\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\mathbf{U}_2\mathbf{J}_{sr}\boldsymbol{\Lambda}_2'\right] \\
&= \sum_{i=1}^{M} n_i^2 \left[\mathbf{U}_2'\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}\right]_{su}\left[\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2\right]_{ur} \\
&= \sum_{i=1}^{M} n_i^2 \left[(\mathbf{U}_2'\boldsymbol{\Lambda}_2'\mathbf{S}_{n_i}) \otimes (\mathbf{S}_{n_i}\boldsymbol{\Lambda}_2)\right]_{us,ru}
\end{aligned}$$

which proves the proposition. □

Proposition 5.34

The (u,v)-th element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{D}_2}, (\boldsymbol{\gamma}^{\mathbf{D}_2})')$ is given by

$$\mathbf{H}([\boldsymbol{\gamma}^{\mathbf{D}_2}]_u, [\boldsymbol{\gamma}^{\mathbf{D}_2}]_v) = \frac{1}{2}\sum_{i=1}^{M} n_i^2 \left[\mathbf{S}_{n_i}\right]_{uv}^2 . \tag{5.44}$$

Proof

An expression for a typical element of $\mathbf{H}(\boldsymbol{\gamma}^{\mathbf{D}_2}, (\boldsymbol{\gamma}^{\mathbf{D}_2})')$, say the (u,v)-th element, follows from (5.19) as

$$\begin{aligned}
\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{D}_2}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{D}_2}\right]_v\right) &= \frac{1}{2}\sum_{i=1}^{M} \mathrm{tr}\left[\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{D}_2]_{uu}}\mathbf{W}_i^{-1}\frac{\partial \mathbf{W}_i}{\partial [\mathbf{D}_2]_{vv}}\right] \\
&= \frac{1}{2}\sum_{i=1}^{M} \mathrm{tr}\left[\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes \mathbf{J}_{uu}\right)\mathbf{W}_i^{-1}\left(\mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes \mathbf{J}_{vv}\right)\right]
\end{aligned}$$

and, using Proposition 5.11, this simplifies to

$$\begin{aligned}\mathbf{H}\left(\left[\boldsymbol{\gamma}^{\mathbf{D}_2}\right]_u, \left[\boldsymbol{\gamma}^{\mathbf{D}_2}\right]_v\right) &= \frac{1}{2}\sum_{i=1}^{M} n_i^2 \mathrm{tr}\left[\mathbf{S}_{n_i}\mathbf{J}_{uu}\mathbf{S}_{n_i}\mathbf{J}_{vv}\right] \\ &= \frac{1}{2}\sum_{i=1}^{M} n_i^2 \left[\mathbf{S}_{n_i}\right]_{uv}^2\end{aligned}$$

which proves the proposition. □

Expressions for calculating typical elements of all the submatrices of $\mathbf{H}(\boldsymbol{\gamma})$ have now been derived and are given by Propositions 5.13 to 5.34. These expressions can now be used to calculate $\mathbf{H}(\boldsymbol{\gamma})$ and $\mathbf{H}^{-1}(\boldsymbol{\gamma})$, where the latter is a necessary element in the Fisher scoring algorithm to obtain the maximum likelihood estimate $\hat{\boldsymbol{\gamma}}$ of the parameter vector $\boldsymbol{\gamma}$, and also, at convergence of the algorithm, provide an estimate of the covariance matrix of the parameter estimates.

5.5 Estimation of parameters and standard errors

Exploratory factor analysis

It has been pointed out in the previous chapter that, in an exploratory analysis, the indeterminacies in the model may be removed by having $\boldsymbol{\Phi}_1$ and $\boldsymbol{\Phi}_2$ as identity matrices and constraining $\boldsymbol{\Lambda}_1'\mathbf{D}_1^{-1}\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2'\mathbf{D}_2^{-1}\boldsymbol{\Lambda}_2$ to be diagonal matrices. This method has the disadvantages of non-linear constraints that have to be imposed, and uninterpretable solutions that will require some kind of rotation.

However, if one chooses to do exploratory factor analysis employing the above methodology, the Fisher scoring method may be used. This method reduces the discrepancy function by iteratively updating the vector of parameter estimates $\hat{\boldsymbol{\gamma}}$ using equation (3.29). At iteration $t+1$ the vector of estimates at the previous iteration, $\hat{\boldsymbol{\gamma}}_t$, is updated

by adding an increment vector $\boldsymbol{\delta}_t$ to it. The expression for calculating $\boldsymbol{\delta}_t$ is found in (3.28) when no constraints are imposed, i.e. only one factor at each of the two levels, and (3.54) when constraints are imposed, i.e. when there is more than one factor at any one of the levels.

In calculating $\boldsymbol{\delta}_t$, the gradient vector $\mathbf{g}(\hat{\gamma}_t)$ may be obtained by means of the expression provided in (5.18) while expressions for obtaining the elements of $\mathbf{H}(\hat{\gamma}_t)$ are provided by Propositions 5.13 to 5.34. In practical applications it is possible that $\mathbf{H}(\hat{\gamma}_t)$ becomes near singular in the iteration procedure. A solution to this problem is to replace $\mathbf{H}^{-1}(\hat{\gamma}_t)$ by a reflexive generalised inverse using a stepwise regression approach (see Browne and Du Toit, 1992) and, in unconstrained estimation, an estimate of the covariance matrix of the parameter estimators is given by $N^{-1}\mathbf{H}^{-1}(\hat{\gamma})$ where $\hat{\gamma}$ replaces $\hat{\gamma}_t$ at convergence.

In constrained estimation it does not matter if $\mathbf{H}_t$ $(= \mathbf{H}(\hat{\gamma}_t))$ becomes singular during the iteration procedure since $\mathbf{D}_t$ in (3.54) may be chosen such that the submatrix $\mathbf{H}_t + \mathbf{L}_t'\mathbf{D}_t\mathbf{L}_t$ will be positive definite, even if $\mathbf{H}_t$ is singular (Browne and Du Toit, 1992). Expressions for obtaining the approximate covariance matrix of the parameter estimators and of the vector of Lagrange multipliers are given by (3.55) and (3.56) respectively.

Confirmatory factor analysis

When one enters confirmatory factor analysis, one has to make sure that the model is specified in such a way that the free parameters are uniquely defined and that the scale of the factors are fixed before estimation of the model parameters can proceed. Once these issues have been satisfactorily solved, the Fisher scoring method may be employed in a similar way as described for exploratory factor analysis. The gradient vector is again calculated using (5.18), but with the difference that zeroes are entered in the positions that correspond to the fixed parameters. In the expected Hessian matrix, the corresponding rows and columns are replaced by zeroes.

At the point of convergence, the Fisher scoring method provides estimates of the standard errors of the parameter esimates. However, since the factor covariance matrices $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$ were factorized in terms of lower triangular matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ respectively, the estimated standard errors will be available for the elements of $\hat{\mathbf{U}}_1$ and $\hat{\mathbf{U}}_2$, and not for $\hat{\mathbf{\Phi}}_1$ and $\hat{\mathbf{\Phi}}_2$. Fortunately, the Delta method (see for example Bishop, Fienberg and Holland (1975)) can be used to obtain the required estimated standard errors. This is done as follows:

We have $\mathbf{\Phi}_1 = \mathbf{U}_1\mathbf{U}_1'$, and the Fisher scoring method will provide the estimates for the elements of $\text{Cov}(\text{Vecs}(\hat{\mathbf{U}}_1), \text{Vecs}'(\hat{\mathbf{U}}_1))$. In order to obtain the estimates for the elements of $\text{Cov}(\text{Vecs}(\hat{\mathbf{\Phi}}_1), \text{Vecs}'(\hat{\mathbf{\Phi}}_1))$, note that the elements of $\text{Vecs}(\mathbf{\Phi}_1)$ are functions of the elements of $\text{Vecs}(\mathbf{U}_1)$, and this relation can be written as

$$\text{Vecs}(\mathbf{\Phi}_1) = \text{f}(\text{Vecs}(\mathbf{U}_1)).$$

According to the Delta method, one may now obtain the covariance matrix of the elements of $\text{Vecs}(\mathbf{\Phi}_1)$ from the expression

$$\text{Cov}(\text{Vecs}(\mathbf{\Phi}_1), \text{Vecs}'(\mathbf{\Phi}_1)) = \mathbf{J}\text{Cov}(\text{Vecs}(\mathbf{U}_1)\text{Vecs}'(\mathbf{U}_1))\mathbf{J}', \tag{5.45}$$

where

$$\mathbf{J} = \frac{\partial \text{Vecs}(\mathbf{\Phi}_1)}{\partial \text{Vecs}'(\mathbf{U}_1)}$$

in which $\mathbf{U}_1$ is replaced by its maximum likelihood estimate.

In a practical application, the following procedure can therefore be followed to obtain the elements of $\hat{\mathbf{\Phi}}_1$ and their estimated standard errors. At the point where the Fisher scoring method has converged, $\hat{\mathbf{U}}_1$ and $\text{Cov}(\text{Vecs}(\hat{\mathbf{U}}_1), \text{Vecs}'(\hat{\mathbf{U}}_1))$ are known. It follows straightforward that $\hat{\mathbf{\Phi}}_1 = \hat{\mathbf{U}}_1\hat{\mathbf{U}}_1'$. The Jacobian matrix $\mathbf{J}$ can now be calculated since the relation between the elements of $\mathbf{\Phi}_1$ and $\mathbf{U}_1$ is known (see for example Graybill (1976)). Replacing $\mathbf{U}_1$ by $\hat{\mathbf{U}}_1$ in (5.45) will now yield the estimated covariance matrix of the elements of $\hat{\mathbf{\Phi}}_1$.

Exactly the same procedure as described above is appropriate for calculating $\hat{\mathbf{\Phi}}_2$ and

$\text{Cov}(\text{Vecs}(\hat{\boldsymbol{\Phi}}_2), \text{Vecs}'(\hat{\boldsymbol{\Phi}}_2))$ from $\hat{\mathbf{U}}_2$ and $\text{Cov}(\text{Vecs}(\hat{\mathbf{U}}_2), \text{Vecs}'(\hat{\mathbf{U}}_2))$.

5.6 Goodness of fit and hypothesis testing

In the traditional case, i.e. where a hierarchical structure is not present in the population under consideration, testing the goodness of fit of a model in which it is assumed that the $p \times p$ population covariance matrix has a certain structure, usually proceeds by testing the null-hypothesis

$$\text{H}_0 : \quad \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\gamma})$$

where $\boldsymbol{\gamma}$ is identified, against the general alternative

$$\text{H}_1 : \quad \boldsymbol{\Sigma} \text{ is any non} - \text{negative } p \times p \text{ matrix.}$$

If $\hat{\boldsymbol{\gamma}}$ is the maximum likelihood estimator of $\boldsymbol{\gamma}$ under H_0, then $\boldsymbol{\Sigma}(\hat{\boldsymbol{\gamma}})$ is the maximum likelihood estimator of $\boldsymbol{\Sigma}$ under H_0. Let $\mathbf{S}$ be the maximum likelihood estimator of $\boldsymbol{\Sigma}$ under H_1. The likelihood ratio test statistic for testing H_0 against H_1 is given by

$$\lambda = \frac{L(\boldsymbol{\Sigma}(\hat{\boldsymbol{\gamma}}))}{L(\mathbf{S})}$$

where $L(\boldsymbol{\Sigma})$ is the likelihood function for the $N \times p$ matrix variate $\mathbf{X}$ whose rows are independently and identically distributed according to the p-variate normal distribution with covariance matrix $\boldsymbol{\Sigma}$.

If H_0 is true, $-2\ell n\lambda$ is asymptotically distributed as a chi-squared variate with degrees of freedom equal to the difference in the number of parameters estimated under H_0 and H_1.

In order to use the likelihood ratio test described above for hypothesis testing, the models under H_0 and H_1 should be estimable and the parameter space for H_0 should be a subset of the full parameter space.

For the present model, i.e. the two-level factor analysis model, let

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_M \end{pmatrix}$$

be the $Np \times 1$ vector variate where each $\mathbf{y}_i$ is a $pn_i \times 1$ vector corresponding to the i-th level two unit. Following from the model (cf. (5.1)), the covariance matrix of $\mathbf{y}$ is given by the $Np \times Np$ matrix

$$\boldsymbol{\Sigma} = \text{Cov}(\mathbf{y}, \mathbf{y}') = \begin{pmatrix} \mathbf{W}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{W}_M \end{pmatrix} \tag{5.46}$$

where each $\mathbf{W}_i$ $(i = 1, 2, \ldots, M)$ is assumed to be structured according to (4.3). The null hypothesis for testing the goodness of fit of this covariance structure model, is

$$H_0 : \quad \boldsymbol{\Sigma} \text{ has the structure given in (5.46).}$$

This restrictive hypothesis may now be tested against an unrestrictive hypothesis along the lines of McDonald and Goldstein (1989). Their work in this regard will now briefly be discussed.

McDonald and Goldstein (1989) show that it is possible to express the log-likelihood,

in the balanced case, as a function of a set of sufficient statistics. They further show that the likelihood equations - i.e. the derivatives of the log-likelihood function with respect to the parameters - for unrestricted parameter matrices are satisfied by simple functions of these sufficient statistics, leading to closed form expressions for the maximum likelihood estimates of the parameters. In the unbalanced case, the log-likelihood is the sum of M terms, and the likelihood equations cannot be solved in closed form.

A discrepancy function that has a minimum of zero, where the minimum is attained if and only if the model fits perfectly, is proposed as a suitable testing procedure. This discrepancy function is based on the ratio of likelihoods and also yields an asymptotic chi-square test for identified models, and is given by

$$U = l_\omega - l_\Omega$$

where l_ω is the likelihood for the restricted parameter matrices and l_Ω is the likelihood for the unrestricted matrices, in both balanced and unbalanced cases.

In the present work, it is also possible to perform hypothesis testing on the number of factors required in the model to adequately fit the data. Suppose one wishes to test the hypothesis that r_1 and r_2 factors at levels one and two respectively, are sufficient. The null hypothesis in this case would be

H_0 : $\boldsymbol{\Sigma}$ has the structure given in (5.46) while the matrices on the diagonal have the structures $\mathbf{W}_i = \mathbf{I}_{n_i} \otimes \mathbf{V}_1 + \mathbf{j}_{n_i}\mathbf{j}'_{n_i} \otimes \mathbf{V}_2$ with $\boldsymbol{\Lambda}_1 : p \times r_1$ and $\boldsymbol{\Lambda}_2 : p \times r_2$.

This null hypothesis may now be tested against the alternative hypothesis that k_1 (≥ 0) additional factors are required at level one and k_2 (≥ 0) additional factors are required at level two. The integers k_1 and k_2 cannot both be zero since H_0 and H_1 will then be identical. This alternative may now be stated as

H_1 : $\boldsymbol{\Sigma}$ has the structure given in (5.46) while the matrices on the diagonal have the structures $\mathbf{W}_i = \mathbf{I}_{n_i} \otimes \mathbf{V}_1 + \mathbf{j}_{n_i}\mathbf{j}'_{n_i} \otimes \mathbf{V}_2$ with $\boldsymbol{\Lambda}_1 : p \times (r_1 + k_1)$ and $\boldsymbol{\Lambda}_2 : p \times (r_2 + k_2)$.

The likelihood ratio test statistic for testing H_0 against H_1 is obtained as the ratio

$$\lambda = \frac{L(\hat{\boldsymbol{\Sigma}}_0)}{L(\hat{\boldsymbol{\Sigma}}_1)}$$

where L is the likelihood function for $\mathbf{y}$ and $\hat{\boldsymbol{\Sigma}}_0$ and $\hat{\boldsymbol{\Sigma}}_1$ are the maximum likelihood estimators of $\boldsymbol{\Sigma}$ under H_0 and H_1 respectively. The test statistic λ may be obtained, if we use (5.2), as

$$\lambda = \frac{\prod_{i=1}^{M} |\mathbf{W}_{i0}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\mathrm{tr}\left[\mathbf{W}_{i0}^{-1}\mathbf{G}_i\right]\}}{\prod_{i=1}^{M} |\mathbf{W}_{i1}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\mathrm{tr}\left[\mathbf{W}_{i1}^{-1}\mathbf{G}_i\right]\}}$$

where $\mathbf{W}_{i0}$ and $\mathbf{W}_{i1}$ are the estimators of the block diagonal matrices respectively in $\hat{\boldsymbol{\Sigma}}_0$ and $\hat{\boldsymbol{\Sigma}}_1$.

If H_0 is true, the limiting distribution of

$$\begin{aligned} -2\ell n\lambda &= -2(\ell n L(\hat{\boldsymbol{\Sigma}}_0) - \ell n L(\hat{\boldsymbol{\Sigma}}_1)) \\ &= \sum_{i=1}^{M} \{\ell n|\mathbf{W}_{i0}| - \ell n|\mathbf{W}_{i1}| + \mathrm{tr}\left[(\mathbf{W}_{i0}^{-1} - \mathbf{W}_{i1}^{-1})\mathbf{G}_i\right]\} \end{aligned} \tag{5.47}$$

is the chi-square distribution with ν degrees of freedom where

$$\begin{aligned}\nu &= (pr_1^\star + pr_2^\star - r_1^\star(r_1^\star - 1)/2 - r_2^\star(r_2^\star - 1)/2) \\ &\quad -(pr_1 + pr_2 - r_1(r_1 - 1)/2 - r_2(r_2 - 1)/2)\end{aligned}$$

with $r_1^\star = r_1 + k_1$ and $r_2^\star = r_2 + k_2$.

5.7 Practical applications

The practical applications given in this section are continuations of those presented in Chapter 4. The parameter estimates obtained in the previous chapter are regarded only as rough estimates. These parameter estimates are subsequently used as starting values in the Fisher scoring method, using the computer program FSBIFAC - written in FORTRAN - to apply the theory derived in this chapter to real life data.

Example 5.7.1: One factor at each level (Exploratory analysis)

For this example, the same model is assumed as in the one-factor application in Chapter 4. The parameter estimates obtained there - provided in Table 4.2 - are used as starting values in FSBIFAC. The Fisher scoring iterations are terminated when the norm of the gradient vector becomes sufficiently small. The norm is calculated as

$$||\mathbf{g}(\boldsymbol{\gamma})|| = (\gamma_1^2 + \gamma_2^2 + \ldots + \gamma_q^2)^{\frac{1}{2}}$$

where q=30 in this case. It was decided that when the norm becomes less than 0,2 the parameter estimates have converged satisfactorily. Applying this criterion, the program FSBIFAC carried out seven iterations before convergence occurred. The parameter estimates at the point of convergence are given in Table 5.1 below. Their estimated standard errors are also calculated by the program and are given in brackets in the table. These estimates are obtained from the inverse of the approximate Hessian matrix

for the parameter estimates (cf. Section 5.5). This matrix is calculated at each iteration and, at the point of convergence, it provides the estimated standard errors, which are given in Table 5.1.

TABLE 5.1

Parameter estimates and standard errors

$\hat{\mu}$	$\hat{\Lambda}_1$	$\hat{D}_1$	$\hat{\Lambda}_2$	$\hat{D}_2$
2,744 (,037)	0,896 (,012)	0,392 (,009)	0,365 (,031)	0,023 (,005)
2,258 (,033)	0,776 (,012)	0,400 (,009)	0,312 (,028)	0,025 (,005)
2,554 (,034)	0,885 (,011)	0,273 (,007)	0,351 (,028)	0,011 (,003)
2,595 (,036)	0,814 (,012)	0,427 (,009)	0,355 (,030)	0,022 (,005)
2,637 (,030)	0,681 (,010)	0,258 (,006)	0,265 (,026)	0,032 (,005)
2,608 (,026)	0,719 (,011)	0,315 (,007)	0,246 (,023)	0,014 (,003)

It is evident that the convergence criterion used here worked well, since from iteration six to seven, no changes occurred in the parameter estimates in the first four decimal places. The next table gives, for each of the seven iterations, the norm of the gradient vector in order to show the speed of convergence.

TABLE 5.2

Norm of gradient vector

Iteration	Norm
1	782,745
2	34,347
3	8,521
4	2,266
5	0,876
6	0,296
7	0,121

It is evident from Table 5.1 that all the factor loading parameter estimates are significantly different from zero, indicating that all six observed variables are significantly related to the factor - at both levels.

Example 5.7.2: One factor at each level (Confirmatory analysis)

In this example, the scale of the factor is determined by fixing the first parameter in $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ - they are both 6×1 matrices - to unity. The parameter representing the factor variance is left free for estimation.

It was decided to terminate the iteration procedure when the norm of the gradient becomes smaller than 0,2. The EM solution (Table 4.4) is used as starting point. Running this application, the program FSBIFAC carried out eight iterations before convergence occurred. The parameter estimates at the point of convergence are given in Table 5.3 below. Their estimated standard errors are also calculated by the program and are given in brackets in the table.

TABLE 5.3

Parameter estimates and standard errors

$\hat{\mu}$	$\hat{\Lambda}_1$	$\hat{D}_1$	$\hat{\Lambda}_2$	$\hat{D}_2$	$\hat{\Phi}_1$	$\hat{\Phi}_2$
2,744 (,037)	1,000	0,392 (,009)	1,000	0,023 (,005)	0,803 (,022)	0,133 (,022)
2,258 (,033)	0,865 (,013)	0,400 (,009)	0,855 (,063)	0,025 (,005)		
2,554 (,034)	0,987 (,013)	0,273 (,007)	0,962 (,057)	0,011 (,003)		
2,595 (,036)	0,908 (,014)	0,427 (,009)	0,972 (,065)	0,022 (,005)		
2,637 (,030)	0,760 (,011)	0,258 (,006)	0,726 (,061)	0,032 (,005)		
2,608 (,026)	0,802 (,012)	0,315 (,007)	0,674 (,050)	0,014 (,003)		

The next table gives the norm of the gradient vector at each iteration in order to show the speed of convergence.

TABLE 5.4

Norm of gradient vector

Iteration	Norm
1	1334,912
2	506,466
3	119,221
4	32,778
5	7,695
6	2,271
7	0,488
8	0,175

Example 5.7.3: Two factors at each level (Exploratory analysis)

In this application the parameter estimates obtained by the EM algorithm, and set out in Table 4.6, are used as initial values in the computer program FSBIFAC. As in the previous example, the iterative procedure was terminated when the norm of the gradient vector became smaller than 0,2. This criterion was reached after 11 iterations, and the parameter estimates after convergence are given in Table 5.5 below. The estimated standard errors - obtained from the inverse of the approximate Hessian matrix after convergence - are provided in brackets.

TABLE 5.5

Parameter estimates and standard errors

$\hat{\mu}$	$\hat{\Lambda}_1$		$\hat{D}_1$
2,745 (,037)	0,176 (,011)	0,879 (,013)	0,391 (,009)
2,258 (,033)	0,167 (,011)	0,758 (,012)	0,400 (,009)
2,554 (,034)	0,159 (,010)	0,870 (,012)	0,273 (,007)
2,596 (,035)	0,118 (,011)	0,807 (,012)	0,425 (,009)
2,638 (,029)	0,125 (,009)	0,671 (,010)	0,258 (,006)
2,609 (,026)	0,142 (,010)	0,706 (,011)	0,315 (,007)
19,609 (,113)	1,304 (,039)	−0,911 (,044)	5,609 (,119)
13,125 (,152)	2,152 (,058)	−1,344 (,068)	11,993 (,261)
24,346 (,286)	2,301 (,050)	−1,752 (,065)	7,588 (,190)
10,738 (,224)	2,354 (,050)	−1,559 (,065)	7,573 (,190)
12,658 (,222)	2,474 (,051)	−1,437 (,067)	7,745 (,198)
20,087 (,199)	1,603 (,056)	−1,197 (,060)	12,096 (,248)

TABLE 5.5 (continued)
Parameter estimates and standard errors

$\hat{\Lambda}_2$		$\hat{D}_2$
0,145 (,052)	0,335 (,041)	0,022 (,005)
0,160 (,044)	0,273 (,041)	0,024 (,005)
0,188 (,044)	0,307 (,046)	0,006 (,003)
0,146 (,049)	0,314 (,040)	0,024 (,005)
0,014 (,043)	0,280 (,025)	0,021 (,004)
0,089 (,035)	0,221 (,028)	0,015 (,003)
0,514 (,155)	−0,914 (,136)	0,459 (,078)
0,633 (,200)	−1,109 (,180)	1,085 (,178)
1,702 (,406)	−2,534 (,387)	1,603 (,268)
1,495 (,307)	−1,985 (,327)	0,417 (,122)
1,452 (,307)	−1,939 (,320)	0,554 (,131)
0,531 (,298)	−1,828 (,195)	1,430 (,229)

It appears that convergence has also been obtained to a satisfactory degree in this application, since no changes occurred in the parameter estimates from the 10th to the 11th iterations in the first four decimal places.

The off-diagonal element of $\hat{\Lambda}_1'\hat{D}_1^{-1}\hat{\Lambda}_1$ at the point of convergence is 0,36E-14 with Lagrange multiplier equal to 0,30E-14, while the off-diagonal element of $\hat{\Lambda}_2'\hat{D}_2^{-1}\hat{\Lambda}_2$ is 0,73E-9 with Lagrange multiplier 0,43E-16. This indicates that the constraints have also converged at this point.

From the results presented in Table 5.5, it is apparent that all factor loading parameter estimates at level one are significantly different from zero. Two clear factors can be detected here - the one being identified by the first six variables, and the second one by the last six variables. This follows from the fact that all loadings are positive on the first factor and that the second factor is bipolar (the first six variables being positive and

the last six being negative). A similar interpretation follows for the second level results - only at this level the loadings on factor one for variables 5 and 12 are not significant.

For each of the 11 iterations, Table 5.6 below gives the norm of the gradient vector in order to show the speed of convergence.

TABLE 5.6

Norm of gradient vector

Iteration	Norm
1	352,795
2	61,638
3	18,045
4	15,293
5	7,352
6	4,194
7	2,129
8	1,075
9	0,529
10	0,259
11	0,126

Example 5.7.4: Two factors at each level (Confirmatory analysis)

In this application the parameter estimates obtained in the previous chapter by the EM algorithm (Table 4.9) are used as initial values in the computer program FSBIFAC. The criterion that the norm of the gradient vector should be smaller than 0,2 was reached after 12 iterations, and the parameter estimates after convergence are given in Table 5.7 below. The estimated standard errors are provided in brackets.

TABLE 5.7

Parameter estimates and standard errors

$\hat{\mu}$	$\hat{\Lambda}_1$		$\hat{D}_1$	$\hat{\Phi}_1$	
2,745 (,037)	1,000	0,000	0,391 (,009)	0,805 (,036)	
2,258 (,033)	0,873 (,015)	0,010 (,009)	0,400 (,009)	−0,571 (,027)	2,529 (,019)
2,554 (,034)	0,979 (,015)	−0,010 (,008)	0,273 (,007)		
2,596 (,035)	0,888 (,015)	−0,029 (,009)	0,425 (,009)		
2,638 (,029)	0,756 (,012)	−0,007 (,007)	0,258 (,006)		
2,609 (,026)	0,803 (,014)	−0,000 (,008)	0,315 (,007)		
19,609 (,113)	0,000	1,000	5,609 (,119)		
13,125 (,152)	0,160 (,092)	1,629 (,056)	11,993 (,261)		
24,346 (,286)	−0,143 (,089)	1,784 (,055)	7,588 (,190)		
10,738 (,224)	0,086 (,090)	1,794 (,055)	7,573 (,190)		
12,658 (,222)	0,291 (,092)	1,859 (,057)	7,745 (,198)		
20,087 (,199)	−0,077 (,080)	1,240 (,049)	12,096 (,248)		

TABLE 5.7 (continued)

Parameter estimates and standard errors

$\hat{\Lambda}_2$		$\hat{D}_2$	$\hat{\Phi}_2$	
1,000	0,000	0,022 (,005)	0,134 (,022)	
0,937 (,082)	0,046 (,027)	0,024 (,005)	−0,232 (,019)	1,099 (,011)
1,079 (,076)	0,060 (,025)	0,006 (,003)		
0,967 (,083)	0,012 (,028)	0,024 (,005)		
0,513 (,069)	−0,118 (,023)	0,021 (,004)		
0,640 (,062)	−0,007 (,021)	0,015 (,003)		
0,000	1,000	0,459 (,078)		
0,024 (,491)	1,223 (,163)	1,085 (,178)		
0,824 (,893)	3,075 (,280)	1,603 (,268)		
1,131 (,705)	2,587 (,236)	0,417 (,122)		
1,080 (,698)	2,518 (,233)	0,554 (,131)		
−1,490 (,575)	1,454 (,191)	1,430 (,229)		

It appears that convergence has also been obtained to a satisfactory degree in this application, since no changes occurred in the parameter estimates from the 11th to the 12th iterations in the first four decimal places.

For each iteration, Table 5.8 below gives the norm of the gradient vector in order to show the speed of convergence.

TABLE 5.8
Norm of gradient vector

Iteration	Norm
1	3611,098
2	1918,345
3	995,426
4	508,482
5	244,188
6	132,777
7	61,575
8	15,814
9	4,974
10	0,748
11	0,505
12	0,124

Example 5.7.5: Testing ordinary factor models against two-level models

In this example we formally test the null-hypothesis that ordinary factor analysis models adequately describe the data that were used in the previous examples, against the

alternative that two-level models are necessary. Two tests are performed: fitting a one-factor model to describe the covariances among six variables, and fitting a two-factor model to describe the covariances among twelve variables. The null- and alternative hypotheses can, for both examples, be written as

$$\mathrm{H}_0 \ : \ \mathbf{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\Lambda}_1 \mathbf{d}_{1,ij} + \mathbf{e}_{1,ij}$$

and

$$\mathrm{H}_1 \ : \ \mathbf{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\Lambda}_1 \mathbf{d}_{1,ij} + \mathbf{e}_{1,ij} + \boldsymbol{\Lambda}_2 \mathbf{d}_{2,i} + \mathbf{e}_{2,i}$$

The only difference between the two hypotheses is the additional parameters which have to be estimated under H_1.

As before, let the $Np \times 1$ vector of observations be represented in terms of the M clusters, by

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_M \end{pmatrix}$$

and let $\mathbf{V}_1$ and $\mathbf{V}_2$ be defined as in Example 4.6.1. Then H_0 and H_1 can equivalently be written as

$$\mathrm{H}_0 \ : \ \mathrm{Cov}(\mathbf{y}, \mathbf{y}') = \mathrm{Diag}(\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_M) \text{ where } \mathbf{W}_i = \mathbf{I}_{n_i} \otimes \mathbf{V}_1$$

and

$$\mathrm{H}_1 \ : \ \mathrm{Cov}(\mathbf{y},\mathbf{y}') = \mathrm{Diag}(\mathbf{W}_1,\mathbf{W}_2,\ldots,\mathbf{W}_M) \text{ where } \mathbf{W}_i = \mathbf{I}_{n_i} \otimes \mathbf{V}_1 + \mathbf{j}_{n_i}\mathbf{j}_{n_i}' \otimes \mathbf{V}_2$$

Under H_0, the log-likelihood is

$$\ell n\ L_0 = -\frac{1}{2}\sum_{i=1}^{M}\left(pn_i\ell n(2\pi) + \ell n|\mathbf{I}_{n_i} \otimes \mathbf{V}_1| + \mathrm{tr}\left[(\mathbf{I}_{n_i} \otimes \mathbf{V}_1)^{-1}\mathbf{G}_i\right]\right). \tag{5.48}$$

However, since $\mathbf{G}_i = (\mathbf{y}_i - \mathbf{j}_{n_i} \otimes \boldsymbol{\mu})(\mathbf{y}_i - \mathbf{j}_{n_i} \otimes \boldsymbol{\mu})'$, it follows that

$$\mathrm{tr}\left[(\mathbf{I}_{n_i} \otimes \mathbf{V}_1)^{-1}\mathbf{G}_i\right] = \mathrm{tr}\left[\mathbf{V}_1^{-1}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \boldsymbol{\mu})(\mathbf{y}_{ij} - \boldsymbol{\mu})'\right]$$

and, since $\ell n|\mathbf{I}_{n_i} \otimes \mathbf{V}_1| = n_i\ell n|\mathbf{V}_1|$, it follows that (5.33) can equivalently be written as

$$\ell n\ L_0 = -\frac{1}{2}N\left(p\ell n(2\pi) + \ell n|\mathbf{V}_1| + \mathrm{tr}[\mathbf{V}_1^{-1}\mathbf{A}]\right) \tag{5.49}$$

where

$$\mathbf{A} = \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \boldsymbol{\mu})(\mathbf{y}_{ij} - \boldsymbol{\mu})'.$$

Under H_1, the log-likelihood is

$$\ell n\ L_1 = -\frac{1}{2}\sum_{i=1}^{M}\left(pn_i\ell n(2\pi) + \ell n|\mathbf{W}_i| + \mathrm{tr}\left[\mathbf{W}_i^{-1}\mathbf{G}_i\right]\right).$$

Inspection of (5.34) shows that the estimator of $\mathbf{V}_1$ under H_0 is the ordinary maximum likelihood estimator obtained by maximising the likelihood function for a sample of N

observations. The estimators of $\mathbf{V}_1$ and $\mathbf{V}_2$ which are used to build $\hat{\mathbf{W}}_1, \hat{\mathbf{W}}_2, \ldots, \hat{\mathbf{W}}_M$ under H_1, have been determined in Chapters 4 and 5, using the hierarchical structure in the population.

The likelihood ratio test statistic, λ, for testing H_0 against H_1 is given by (5.32) and, if H_0 is true, -2 $\ell n\ \lambda$ is asymptotically distributed as a chi-squared variate with degrees of freedom equal to

$$\begin{aligned}\nu &= (3+r_1+r_2)p - r_1(r_1-1)/2 - r_2(r_2-1)/2 - [(2+r_1)p - r_1(r_1-1)/2] \\ &= (r_2+1)p - r_2(r_2-1)/2.\end{aligned}$$

It follows that

$$\begin{aligned}-2\ell n\ \lambda &= -2(\ell n\hat{L}_0 - \ell n\hat{L}_1) \\ &= -2\left(-\frac{1}{2}N\ell n|\hat{\mathbf{V}}_1| - \frac{1}{2}N\mathrm{tr}[\hat{\mathbf{V}}_1^{-1}\hat{\mathbf{A}}] + \frac{1}{2}\sum_{i=1}^{M}\ell n|\hat{\mathbf{W}}_i| + \frac{1}{2}\sum_{i=1}^{M}\mathrm{tr}[\hat{\mathbf{W}}_i^{-1}\hat{\mathbf{G}}_i]\right) \\ &= N\left(\ell n|\hat{\mathbf{V}}_1| + \mathrm{tr}[\hat{\mathbf{V}}_1^{-1}\hat{\mathbf{A}}]\right) - \sum_{i=1}^{M}\left(\ell n|\hat{\mathbf{W}}_i| + \mathrm{tr}[\hat{\mathbf{W}}_i^{-1}\hat{\mathbf{G}}_i]\right)\end{aligned}$$

where $\hat{\mathbf{V}}_1$ and $\hat{\mathbf{A}}$ are the maximum likelihood estimators of $\mathbf{V}_1$ and $\mathbf{A}$ under H_0, and $\hat{\mathbf{W}}_i$ and $\hat{\mathbf{G}}_i$ are the maximum likelihood estimators of $\mathbf{W}_i$ and $\mathbf{G}_i$ under H_1.

The test procedure described above was used to test if the ordinary factor analysis models fitted in Examples 2.6.1 and 2.6.2 are adequate to describe the data, or if the use of additional information - in this case the fact that the data were collected at two levels - used in Examples 5.7.1 and 5.7.2 produced significantly better fits.

In the case of six variables and one factor, the value of -2 $\ell n\ \hat{L}_0$ is 75 918,357 and the value of -2 $\ell n\ \hat{L}_1$ is 74 418,445. The difference between these two values results in a value of 1 499,912 for -2 $\ell n\ \lambda$, and with 12 degrees of freedom, this suggests that the evidence in favour of H_1 is highly significant. With twelve variables and two factors, the values of -2 $\ell n\ \hat{L}_0$ and -2 $\ell n\ \hat{L}_1$ are respectively 259 930,778 and 252 420,562 which results in the value of 7 510,216 for -2 $\ell n\ \lambda$, and with 35 degrees of freedom, this also

suggests that the two-level model provides a more adequate fit.

5.8 Summary

This chapter presents a discussion of how to obtain the maximum likelihood estimators of the unknown parameters in a two-level factor analysis model by means of the Fisher scoring method.

The likelihood function is given, and subsequently the gradient vector and expected Hessian matrix of this function with respect to the parameters are derived. It is then shown how this information is used in the iterative Fisher scoring method to obtain the maximum likelihood parameter estimators in unconstrained and constrained estimation.

The goodness of fit and hypothesis testing of the model are discussed next, pointing out some difficulties in the unbalanced case - i.e. when the n_i's are unequal. In the practical application section, the same data as in the examples in Chapters 2 and 4 are used to fit the two-level factor analysis model by means of the Fisher scoring method. The computer program FSBIFAC - written in FORTRAN - was applied to do the analyses. The same models are fitted as in Chapter 4, using the rough estimates obtained there as starting values for the program FSBIFAC. As a final application, formal hypothesis tests were performed to test whether two-level models provide better descriptions of the data than ordinary factor analysis models.

CHAPTER 6

SUGGESTIONS FOR FURTHER RESEARCH

This final chapter will be devoted to presenting a few topics or ideas that may lead to interesting further research in the field of multilevel factor analysis and its practical application.

One of the first things that may cross the mind of someone who reads this thesis is the issue of standardised versus unstandardised observed variables. It is common practice in ordinary factor analysis to perform the analysis on the correlation matrix of the observed variables, which is equivalent to analysing standardised variables. This practice simplifies the interpretation of results. In the case of multilevel latent variable models, however, their introduction and application have mostly been in terms of the unstandardised form of the observed variables, and the analyses are performed on the covariance matrices (see e.g. Goldstein and McDonald (1988) and Longford and Muthén (1992)). One exception is the work of McDonald (1994). Consequently it seems that further research is required in the area of standardisation of the observed variables in multilevel latent variable models.

When it comes to parameter estimators, it is well known that in ordinary factor analysis there are several methods that may be chosen from to obtain them. Some do not require extensive calculations and may be considered suitable for obtaining quick estimates to be used for exploration only. Here, the principal factor method is an example. Other methods, of which that of maximum likelihood is one, require many more calculations, but the estimators then obtained have desirable asymptotic distributional properties, and testing the goodness of fit of the model is also possible. Since only the maximum likelihood method was considered as a method of obtaining the parameters in a multilevel factor analysis model, there is much room for research on using different methods of parameter estimation in multilevel factor analysis, especially methods that can be used to obtain quick approximations of the parameter estimates.

In Section 2.4 a test is provided for testing statistically whether a specified number of factors in ordinary factor analysis is sufficient for the model to hold. Throughout the thesis, however, it is assumed that the number of factors on both levels is fixed and known. The problem when this number on either or both the levels is unknown, and has to be estimated from the data, has received no attention. Possibly, similar methods for deciding on the number of factors which are used in ordinary factor analysis, such as the scree test and Kaiser's rule of number of eigenvalues greater than one, may be used in multilevel factor analysis. There is, however, room for more research in this area.

The fitting of different two-level factor analysis models, using the estimation procedures provided in this thesis, may be worth considering. This area has been left untouched in the sense that estimation and testing procedures are provided in a very general framework and, subsequently, only one specific two-level factor analysis model is subjected to this theory. It would be a good idea to investigate the possibility of fitting different models into the framework. One such model is that for varying factor means (Muthén, 1994) which states that

$$\mathbf{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\lambda} d_{ij} + \mathbf{e}_{ij}$$

and that only the parameters for factor means vary across the groups. The means are now specified in terms of random effects and are written as

$$d_{ij} = \nu + \eta_{Bi} + \eta_{Wij}$$

where ν is the overall mean for d_{ij}, and η_{Bi} and η_{Wij} are the random between-group and within-group effects respectively, each having zero mean. In addition to considering this model, further research could also be aimed at extending it to include more than one factor.

The distributional assumptions that are made, where the data and the random pa-

rameters are concerned, do not deviate from standard procedures in that the normal distribution is used in all cases. Consequently there is obvious space for further research in the estimation and testing procedures provided here when non-standard distributional assumptions are involved.

There are three issues regarding the implementation of constraints in the two-level factor analysis model (or other multilevel latent variable models) that deserve mention as possible further research areas. The first issue - namely, when constraints are considered which simultaneously involve parameters on both levels (in the case of a two-level model) - is mentioned in Chapter 3 and is not considered in this thesis because such constraints are not necessary in the specific model being analysed here. Constraints of that kind may, however, be of importance in other multilevel latent variable models, which indicates a direction for further research.

The second issue is concerned with imposing constraints different from those considered in this thesis. For the two-level model, where the within-group and the between-group variation is assumed to follow factor structures, constraints of the kind provided by Browne and Du Toit (1992) may be extremely useful. The parameter estimates that satisfy these constraints are simultaneously those that maximise the varimax criterion. If these constraints are imposed separately on the parameters at both levels, the estimated factor matrices obtained will then have a simple structure and need not be rotated. Constraints are also provided by Browne and Du Toit (1992) that lead to the parameter estimates minimising the direct quartimin criterion. Further investigation in this area may yield useful results in the analysis of multilevel factor analysis models.

Thirdly, the method used here to obtain constrained parameter esimates in both the EM algorithm and Fisher scoring method has been to approximate the non-linear constraints by linear constraints and then to apply the method of Lagrange multipliers to estimate the constrained parameters. It may be worth looking into the possibility of using different methods to impose constraints on the parameters in this kind of model.

Applying the EM algorithm in practice requires initial estimates of the parameters to get the iteration procedure started. In the thesis these initial values are merely provided, and nothing is said about where they come from. No analytic method was used to obtain them - they are simply arbitrarily chosen. The first choice, however, was not always successful. From this experience gained in applying the EM algorithm in preparing the practical examples, it seems that the iteration procedure in this specific case is fairly sensitive to the initial values being used. As a consequence, difficulties are frequently encountered during the first few iterations - namely, negative variance estimates are obtained that lead to the inversion of negative definite matrices, and this causes all kinds of computational problems. The problems may be overcome by changing the initial values, but that involves an extremely tedious procedure just to get the iteration process going. A simple procedure was used in the applications to reduce this - namely, whenever a variance parameter became negative, it was replaced by its value in the previous iteration, or sometimes by a small positive value (e.g. 0,005). This is very much a practical difficulty, and further research to obtain a feasible solution will be useful.

References

ANDERSON, T.W. (1984). *An introduction to multivariate statistical analysis*. 2nd Edition. John Wiley & Sons, New York.

BENTLER, P.M. (1986). Structural modeling and Psychometrika: An historical perspective on growth and achievements. *Psychometrika* 51, 35-51.

BENTLER, P.M. and BONETT, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin* 88, 588-606.

BISHOP, Y.M.M., FIENBERG, S.E. and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis : Theory and Practice*. The MIT Press, Cambridge.

BOCK, R.D. (1990). Marginal maximum likelihood estimation for the two-stage random regressions model with second-stage covariates. *Technical report*, Department of Psychology, University of Chicago, Chicago.

BOCK, R.D. and AITKIN, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* 46, 443-445.

BOCK, R.D. and BARGMANN, R.E. (1966). Analysis of covariance structures. *Psychometrika* 31, 507-534.

BROWNE, M.W. (1968). A comparison of factor analytic techniques. *Psychometrika* 33, 267-334.

BROWNE, M.W. (1969). *Factor analysis models and their application to prediction problems*. Unpublished Ph.D Thesis, University of South Africa.

BROWNE, M.W. (1991). *STA* 401 – *Matrix methods in statistics*. Supplementary notes. University of South Africa, Pretoria.

BROWNE, M.W. and DU TOIT, S.H.C. (1992). Automated fitting of nonstandard models. *MultivariateBehavioralResearch* 27, 269-300.

BROWNE, M.W. and MELS, G. (1990). *RAMONA PC User's Guide.* University of South Africa, Department of Statistics.

BRYK, A.S. and RAUDENBUSH, S.W. (1992). *Hierarchical Linear Models : Applications and Data Analysis Methods.* Sage Publications, Inc., Newbury Park, California.

CARROLL, J.B. (1953). An analytical solution for approximating simple structure in factor analysis. *Psychometrika* 18, 23-38.

DEMPSTER, A.P., RUBIN, D.B. and TSUTAKAWA, R.K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association* 76, 341-353.

DUNN, J.E. (1973). A note on a sufficiency condition for uniqueness of a restricted factor matrix. *Psychometrika* 38, 141-143.

DU TOIT, S.H.C. (1993). *Analysis of multilevel models, Part 1 - Theoretical aspects.* Human Sciences Research Council, Pretoria, South Africa.

EVERITT, B.S. (1984). *An introduction to latent variable models.* Chapman and Hall, London.

GOLDSTEIN, H. and McDONALD, R.P. (1988). A general model for the analysis of multilevel data. *Psychometrika* 53, 455-467.

GRAYBILL, F.A. (1976). *TheoryandApplicationoftheLinearModel.* Wadsworth Publishing Company, Inc., Belmont, California.

HARMAN, H.H. (1976). *Modern factor analysis. 3rd Edition Revised.* University of Chicago Press, Chicago.

HOWE, W.G. (1955). *Some contributions to factor analysis.* Oak Ridge National Laboratory Report ORNL-1919, Oak Ridge, Tenn.

JACKSON, J.E. (1991). *A user's guide to principal components.* John Wiley & Sons, New York.

JENNRICH, R.I. (1978). Rotational equivalence of factor loading matrices with specified values. *Psychometrika* 31, 313-323.

JENNRICH, R.I. and SAMPSON, P.F. (1966). Rotation for simple loadings. *Psychometrika* 31, 313-323.

JöRESKOG, K.G. (1966). Testing a simple stucture hypothesis in factor analysis. *Psychometrika* 31, 165-178.

JöRESKOG, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* 34, 183-202.

JöRESKOG, K.G. (1979). A general approach to confirmatory maximum likelihood factor analysis - with Addendum. In *Advances in factor analysis and structural equation models.* (Ed J. Magidson) Abt Associates Inc., Cambridge, Massachusetts.

JöRESKOG, K.G. and SöRBOM, D. (1981). *Analysis of linear structural relationships by maximum likelihood and least squares methods.* Research Report 81-8, University of Uppsala, Sweden.

LAWLEY, D.N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proc. R. Soc. Edinb. A* 60, 64-82.

LAWLEY, D.N. and MAXWELL, A.E. (1971). *Factor analysis as a statistical method.* Butterworth & Co., London.

LEE, S.Y. and JENNRICH, R.I. (1979). A study of algorithms for covariance structure analysis with specific comparisons using factor analysis. *Psychometrika* 44, 99-113.

LINDLEY, D.V. and SMITH, A.F.M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, 34, 1-41.

LONG, J.S. (1983). *Confirmatory factor analysis.* Series: Quantitative applications in the social sciences. SAGE Publications Inc., Beverly Hills, California.

LONGFORD, N.T. and MUTHÉN, B.O. (1992). Factor analysis for clustered populations. *Psychometrika* 57, 581-597.

MAGNUS, J.R. and NEUDECKER, H. (1988). *Matrix differential calculus with applications in statistics and econometrics.* New York: Wiley.

MARTIN, J.K. and McDONALD, R.P. (1975). Bayesian estimation in unrestricted factor analysis: A treatment for Heywood cases. *Psychometrika* 40, 505-517.

McARDLE, J.J. and McDONALD, R.P. (1984). Some algebraic properties of the Reticular Action Model for moment structures. *Br. J. Math. and Stat. Psych.* 37, 234-251.

McDONALD, R.P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *Br. J. Math. and Stat. Psych.* 23, *Part* 1, 1-21.

McDONALD, R.P. (1985). *Factor analysis and related methods.* Lawrence Erlbaum Associates, Hillsdale, New Jersey.

McDONALD, R.P. (1993). A general model for two-level data with responses missing at random. *Psychometrika* 58, 575-585.

McDONALD, R.P. (1994). The bilevel reticular action model for path analysis with latent variables. *Sociological Methods & Research, Vol.* 22, *No.* 3, 399-413.

McDONALD, R.P. and GOLDSTEIN, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *Br. J. Math. and Stat. Psych.* 42, 215-232.

MORRISON, D.F. (1990). *Multivariate statistical methods.* 3rd Edition. McGraw-Hill, Inc., New York.

MULAIK, S.A. (1972). *The foundations of factor analysis.* McGraw-Hill, Inc., New York.

MUTHÉN, B.O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika* 54, 557-585.

MUTHÉN, B.O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, Vol.* 22, *No.* 3, 376-398.

RAO, C.R. (1955). Estimation and tests of significance in factor analysis. *Psychometrika* 20, 93-111.

STEIGER, J.H. and LIND, J.C. (1980). *Statistically based tests for the number of common factors.* Article presented at the annual meeting of the Psychometric Society, Iowa City, IA.