ABSTRACT

Title of Dissertation: Factor Analysis of Cross-Classified Data

Hsiao-Hui S. Tsou, Doctor of Philosophy, 2005

Dissertation directed by: Professor Eric V. Slud
Department of Mathematics
STAT Faculty

This thesis introduces a model hierarchy related to Principal Component Analysis and Factor Analysis, in which vector measurements are linearly decomposed into a relatively small set of hypothetical principal directions, for purposes of dimension reduction. The mathematical specification of unknown parameters in the models is unified. Identifiability of the suitably defined models is proved. The EM algorithm and the Newton-Raphson algorithm based on likelihoods and profile likelihoods are implemented to get computationally effective (maximum likelihood) estimators for the unknown parameters. A restricted model (with some error variances 0) and a sufficient condition for a local maximum likelihood estimate are established. Score tests are constructed to check whether error variances are 0, which is shown to be associated with non-identifiability of models. Statistical tests of goodness of fit of the models to data are established in a likelihood ratio testing framework, so that the most parsimoniously parameterized model consistent with the data can be chosen for purposes

of description and classification of the experimental settings. The results are applied on a real data set involving coronal cross-sectional ultrasound pictures of the human tongue surface during speech. The likelihood ratio test is used to test the fit of the PARAFAC model on the real coronal tongue data, leading to a finding of inadequacy of the PARAFAC model.

Factor Analysis of Cross-Classified Data

by

Hsiao-Hui S. Tsou

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2005

Advisory Committee:

Professor Eric V. Slud, Chairman/Advisor
Professor Paul J. Smith
Professor Benjamin Kedem
Professor Michael M. Boyle
Professor William Levine

DEDICATION

To My Parents and My Husband, Yuh-Lin Lou

ACKNOWLEDGEMENTS

This study would never have been accomplished without the contribution of the following people. First of all, I am deeply grateful to my thesis advisor, Dr. Eric Slud whose enlightening guidance, enormous support and infinite patience made this possible. I also appreciate the timely advice provide to me by Dr. Paul Smith, Dr. Benjamin Kedem, Dr. Michael Boyle, and Dr. William Levine. I also thank Dr. Maureen Stone of Dental School in University of Maryland at Baltimore who provided me with the tongue data and additional insight necessary to complete this research. I thank also Dr. Robert Jennrich, and Dr. Mortaza Jamshidian for useful correspondence about the computation in factor analysis models. I wish to thank Dr. Lawrence Washington for help in linear algebra and generous donation of his time. My research on the tongue data was partially supported by NIH Grant R01 DC 01758 with Dr. Maureen Stone as Principal Investigator. I appreciate the patience of my loving husband, Yuh-lin, who gave me the time I needed. The gift of unbounded love and support has no

equal. Finally, I would like to extend my sincere thanks to my mom and my mother in law for their endless support and unfailing tolerance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

In statistical practice, for investigations involving a large number of observed variables, it is often useful to simplify the analysis by considering a small number of linear combinations of the original variables. For example, scholastic achievement tests usually consist of a number of examinations in different subject areas. In attempting to rate students applying for admission, college administrators frequently attempt to reduce the scores from all subject areas to a single, overall score. If the reduction can be done with minimal information loss, it is better. Principal Component Analysis (PCA) is a method for data reduction. It is used to find linear combinations of the original variables which account for most of the variance in the original sample [2].

In many scientific fields, notably psychology and other social sciences, we are often interested in quantities, such as intelligence or social status, that are not directly measurable. However, it is often possible to measure other quantities which reflect the underlying variable of interest. Factor analysis is an attempt to explain the correlations between observable variables in terms of underlying factors, which are themselves not directly observable. For example, measurable quantities such as performance on a series of tests can be explained in terms of an underlying factor such as intelligence.

At first glimpse, factor analysis closely resembles principal components analy-

sis. Both use linear combinations of variables to explain sets of observations of many variables. In principal component analysis, the observed variables are themselves the quantities of interest. The combination of these variables in the principal components is primarily a tool for simplifying the interpretation of the observed variables. Principal components analysis is merely a transformation of the data. No assumptions are made about the form of the covariance matrix of the data. On the other hand, factor analysis assumes that the data comes from a statistical model which can be expressed in terms of a few underlying, but unobservable, random quantities called *factors* and some additional sources of variation called *error*. Factor analysis can be considered as an extension of principal components analysis. Both can be viewed as attempts to approximate the covariance matrix. Applications of PCA and factor analysis have become very popular in many fields such as psychology, economics, sociology, meteorology, medicine, political science, taxonomy and archaeology. Both of them have been successfully used in acoustic and phonetic research on tongue position by Harshman et al. (1977) , Jackson (1988), Nix et al. (1996), and Stone et al. (1997).

The PARAFAC model was pioneered by Harshman et al. (1977). It is a technique for extracting "articulatory prime" shapes from data allowing non-orthogonal components to scale differently for different speakers. The main concern underlying the PARAFAC model is how to modify the small set of prime shapes with large variance of sound production for different speakers, without requiring large numbers of parameters for all speaker and sound combinations. PCA might do well in reducing the dimension without extracting the behaviors for individual speaker differences. On the other hand, the PARAFAC model succeeds in decomposing tongue shape data into tongue shape factors. In my thesis, PCA, Factor Analysis and the PARAFAC model are introduced. A model hierarchy is defined, and then is applied to coronal tongue

cross-section ultrasound data of multiple subjects collected in the laboratory of Dr. M. Stone [22]. We also discuss the interpretation for the tongue data of the assumptions defining the models presented. Then we present data analytic results to distinguish which model is adequate.

1.1 Principal Component Analysis

PCA is concerned with explaining the variance-covariance structure through a few linear combinations of the original variables. The definition of Principal Components in the population is as follows.

Suppose the random vector

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix}$$

has the covariance matrix Σ. Since we will be interested only in the variance-covariance structure, we assume that the mean vector is $\mathbf{0}$. Let l be a p-component column vector such that $l^t l = 1$. The variance of $l^t Y$ is

$$E(l^t Y Y^t l) = l^t \Sigma l. \tag{1.1}$$

The i'th Principal Component, usually denoted by PC_i, can be defined inductively. The first principal component PC_1 is the linear combination $l_1^t Y$ where l_1 is the vector which maximizes $Var(l_1^t Y)$ subject to $l_1^t l_1 = 1$. The second principal component PC_2 is the linear combination $l_2^t Y$ where l_2 maximizes $Var(l_2^t Y)$ subject to $l_2^t l_2 = 1$ and $Cov(l_1^t Y, l_2^t Y) = 0$. Similarly, the i'th principal component PC_i= $l_i^t Y$ where l_i maximizes $Var(l_i^t Y)$ subject to $l_i^t l_i = 1$ and $Cov(l_k^t Y, l_i^t Y) = 0$ for $k < i$. Thus, the first principal component has the largest variance among all standardized linear

combinations of Y. Similarly, the second principal component has the largest variance among all standardized linear combinations of Y uncorrelated with the first principal component, and so on.

By the method of Lagrange multipliers, we can obtain that $PC_i = v_i^t Y$, where (λ_1, v_1), (λ_2, v_2), ..., (λ_p, v_p) are the eigenvalue-eigenvector pairs of Σ with

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$$

and

$$Cov(PC_i, PC_j) = \lambda_i \delta_{ij} \tag{1.2}$$

$$\sum_{i=1}^{p} Var(Y_i) = \lambda_1 + \ldots + \lambda_p. \tag{1.3}$$

Equation (1.3) is true when all the eigenvectors are distinct. It can be arranged to be true by the following two lemmas if some eigenvalues are the same [2].

Lemma 1. *Suppose $\lambda_{r+1} = \lambda_{r+2} = \ldots = \lambda_{r+m} = t$; then $(\Sigma - tI)$ is of rank $p - m$. Furthermore, the $p \times m$ matrix whose columns consist of an m-tuple of orthonormal eigenvectors v^*=$(v_{r+1} \ldots v_{r+m})$ of $(\Sigma - tI)$ is uniquely determined up to multiplication on the right by an orthogonal matrix.*

Lemma 2. *An orthogonal transformation $V = CY$ of a random vector Y leaves invariant the generalized variance and the sum of the variances of the components. The generalized variance of Y is defined as the determinant of EYY^t if $EY = 0$.*

The proof can be found in (Anderson 1984). The proportion of total variance due to the k'th principal component is

$$\frac{\lambda_k}{\lambda_1 + \ldots + \lambda_p}, \quad k = 1, \ldots, p$$

The vectors v_i used in defining the i'th principal component of the original variables are called Principal Directions. In general, there are as many principal components as variables. However, because of the way they are calculated, it is usually desirable to consider only a few of the principal components, which together explain most of the original variation. The most popular criterion to determine the number q of principal components to retain in describing data is

$$\frac{\sum_{k=1}^{q} \lambda_k}{\sum_{j=1}^{p} \lambda_j} \geq 1 - \alpha \tag{1.4}$$

for suitably defined constant α, usually, $.05$ or $.10$.

1.2 Factor Analysis

Factor analysis is a branch of statistical science. The origin of factor analysis is ascribed to Charles Spearman (1904). He was called the father of factor analysis because of his remarkable work in developing psychological theories involving factor analysis (Harman 1976). The further development of psychological theories and mathematical foundations of factor analysis was continued by Cyril Burt, Karl Pearson, Godfrey Thomson, etc. Applications of factor analysis in fields other than psychology have become very popular since 1950, along with the development of fast computers. The main applications of factor analytic techniques are to reduce the number of variables and to detect structure in the relationships between variables, that is, to classify variables. Therefore, factor analysis is applied as a data reduction or structure detection method. In order to analyze observed data, one approach is to provide a statistical model, to explain the underlying behavior of the data.

The general factor analysis model is defined as follows: let the observable vector

Y be written as

$$Y = \underline{\mu} + \Lambda \mathbf{f} + U \tag{1.5}$$

where Y, $\underline{\mu}$, and U are column vectors of p components, Λ is a $p \times q$ matrix of constants with q fixed and less than p, and $\mathbf{f}$ is a $q \times 1$ random vector. The elements of Λ are called $factor\ loadings$ and the matrix Λ is called the $loading\ matrix$. The elements of $\mathbf{f}$ are called $common$ factors and the elements of U are called $unique$ factors. We assume that $\mathbf{f} \sim N(0, I_q)$, $U \sim N(0, \Psi)$, $\mathbf{f}$ and U are independent, and Ψ is a $p \times p$ diagonal matrix. Therefore, the general random-effect factor model can be expressed as

$$Y \sim N(\underline{\mu},\ \Lambda\, \Lambda^t +\ \Psi).$$

We will present the parameter space of this model (M0) in the next chapter.

1.3 Overview of the Thesis

In Chapter 2, we introduce the general factor models and construct a model hierarchy (Figure 1.1 and Figure 1.2) for the application to tongue image data. For each model, we introduce the model assumptions and the parameter spaces, and then give a proof of identifiability of the model from data. The general sufficient condition for identifiability in the general random effect factor model (M0) has not been accomplished yet, but we find some new results related to the non-identifiable models and the parameters in the boundary of the parameter space.

In Chapter 3, we find the maximum likelihood estimators for the parameters (Λ, Ψ) in the factor models with error-matrix Ψ proportional to I_p (model M1) or to $diag(\underline{e})$ for a vector $\underline{e}$ with entries 0 or 1 (model M1R). In Section 3.2, we introduce the idea

of profile likelihood and use it to find the maximum likelihood estimators for the parameters under (M1R). In Section 3.3, we discuss profile likelihood optimization in (M0). In Section 3.4, we find a necessary condition to check the local maximum likelihood estimate. In Section 3.5, we consider the score test within (M1) for the problem $H_0 : \psi_{jj} = 0$ vs $H_A : \psi_{jj} > 0$. In Section 3.6, we discuss the likelihood ratio test for testing fit of the PARAFAC (M4a) against the fixed-effect factor model (M3). The PARAFAC model is a restricted model of (M3) in which each component of the fixed-effect factor is decomposed as a product of two terms. Details are in Chapter 2.

In Chapter 4, we introduce the EM algorithm and Newton-Raphson optimization method and develop an EM algorithm to compute the maximum likelihood estimator (MLE) for (M0). The performance of the algorithm on simulated data is described, particularly in relation to approximate non-identifiability. The Newton-Raphson method is also used to calculate the MLE of the profile likelihood function $l_p(\Psi)$ and is shown to give results for random effect factor models (M0) that agree with the EM algorithm. We find a new result that an MLE can be found on the boundary of the parameter space when the model is non-identifiable. Details of computations in MATLAB for (M4a) and Splus for (M3), are also given in this chapter.

In Chapter 5, we introduce a real data set of ultrasound cross-sectional images of the human tongue during speech. The PARAFAC (M4a) model had been successfully used in some tongue image data. However, Slud et al. [22] actually found (M4) which is similar to PARAFAC but with orthogonal loading matrix is inadequate to represent the data. Therefore, a more general model such as PARAFAC model (M4a) or fully general fixed effect factor model (M3) is needed for representing cross-classified data. Thus, the well-defined model hierarchy we constructed may help to rationalize the choice of models. In this chapter, the Likelihood Ratio Test (LRT) is used to test

whether the more general models (M3) or (M4a) represent the coronal tongue data better. We construct an algorithm and use a MATLAB toolbox to get the MLE for (M4a). We find that the more general model (M3) fits the coronal tongue data better than the PARAFAC (M4a) model.

In Chapter 6, we summarize the results from this research, and discuss future work.

1.4 Some Definitions and Notations

In this section, we define some notations that will be used in this thesis.

Notation 1.1. *Let $\mathcal{M}_{ab}$ denote the space of real $a \times b$ matrices and let $\mathcal{M}_{ab}^{+}$ denote the subset of matrices in $\mathcal{M}_{ab}$ satisfying the additional constraint that the first nonzero element in each column is positive.*

Notation 1.2. *The notation M^t for a matrix $M \in \mathcal{M}_{ab}$ denotes the transpose of M.*

Notation 1.3. *Let $\mathbf{R}^p$ denote the Euclidean p-dimensional space consists of all ordered p-tuples of real numbers. Symbolically,*

$$\mathbf{R}^p = \{(v_1, \ldots, v_p) : v_1, \ldots, v_p \in \mathbf{R}\}$$

We denote $\mathbf{R}_+^p$ as a subspace of $\mathbf{R}^p$ which consists of all ordered p-tuples of positive real numbers.

Notation 1.4. *The notation $diag(\mathbf{v})$ for a vector $\mathbf{v} \equiv (v_1, \ldots, v_p) \in \mathbf{R}^p$ denotes the square diagonal $p \times p$ matrix with $(v_1, \ldots, v_p)$ on the diagonal.*

Notation 1.5. *Let $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_q$ be q vectors in $\mathbf{R}^p$. The matrix W consisting of the q vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_q$ as its column vectors is denoted by $W = (\mathbf{w}_1|\mathbf{w}_2|\ldots|\mathbf{w}_q)$. Thus, $W \in \mathcal{M}_{pq}$.*

Notation 1.6. *Let $W \equiv (\mathbf{w}_1|\mathbf{w}_2|\ldots|\mathbf{w}_q)$. The notation $col(W)$ for the matrix W denotes the column space W. Then $col(W) = span\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_q\}$, which is the space spanned by the column vectors of W.*

Notation 1.7. *Let O_{pq} denote the space of $p \times q$ matrices M with real components and orthogonal columns, ordered by decreasing norm, (i.e., matrices satisfying M^tM = $diag(v_1, v_2, \ldots, v_q)$ with $v_1 > v_2 > \ldots > v_q > 0$), and let O^+_{pq} denote the subset of matrices in O_{pq} satisfying the additional constraint that the first nonzero element in each column is positive.*

Notation 1.8. *Let $M \in \mathcal{M}_{pq}$. The notation $range(M) \equiv \{M\mathbf{v} : \mathbf{v} \in \mathbf{R}^q\}$ is the range of the matrix M.*

Definition 1.9. *Let $M \in \mathcal{M}_{pp}$. If $\mathbf{v}^tM\mathbf{v} > 0$ for all non-zero vectors $\mathbf{v} \in \mathbf{R}^p$, then M is said to be positive definite on $\mathbf{R}^p$. If $\mathbf{v}^tM\mathbf{v} \geq 0$ for all $\mathbf{v} \in \mathbf{R}^p$, then M is said to be positive semidefinite (or non-negative definite). Positive definiteness (semidefiniteness) of a symmetric matrix is denoted by $M \succ 0$ ($M \succeq 0$).*

Definition 1.10. *Let $A = (a_{ij}) \in \mathcal{M}_{pp}$. The trace of A, denoted by $tr(A)$, is defined as $tr(A) = \sum_{j=1}^{p} a_{jj}$.*

Model Hierarchy (1)

General *Random Effect* Factor Model(M0)

$Y = \underline{\mu} + \Lambda f + U$

$F \sim N(0, I)$,

$U \sim N(0, \text{diag}(\varphi))$, $\varphi = (\varphi_1, \ldots, \varphi_p)$

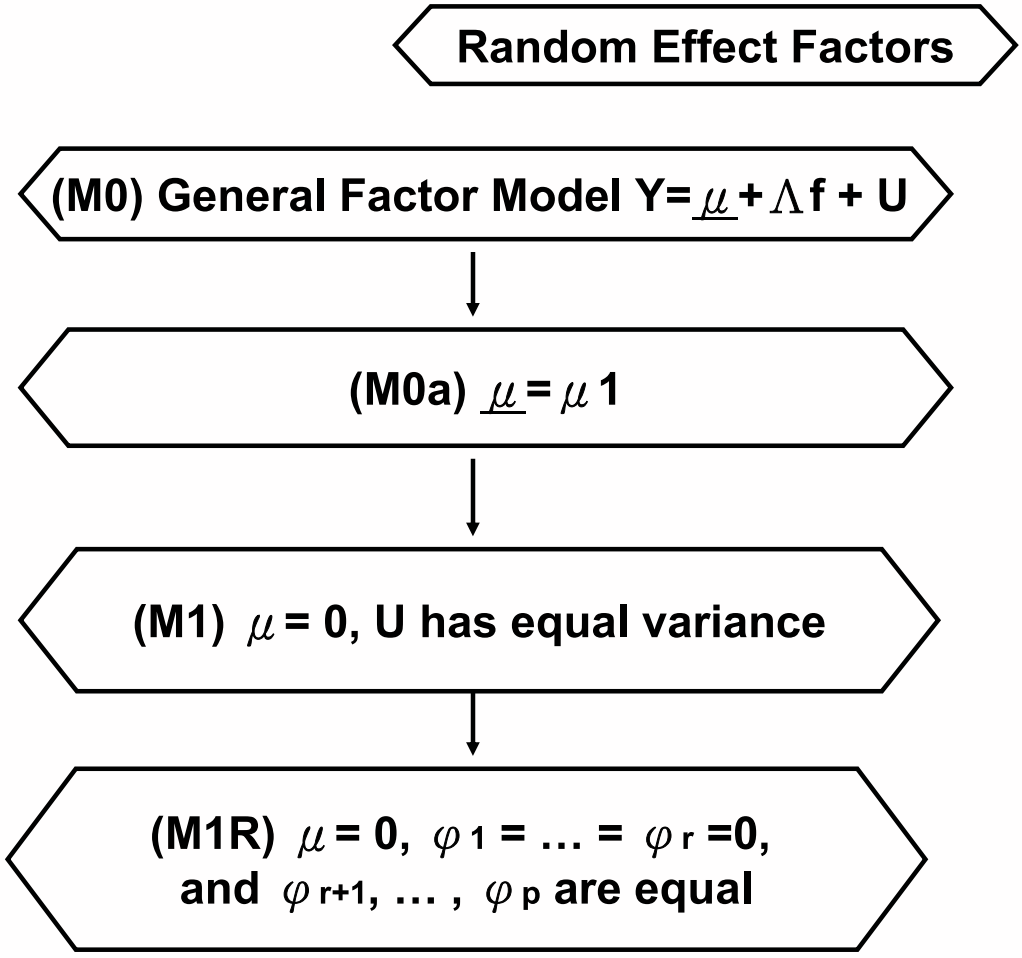

Figure 1.1: Model Hierarchy.

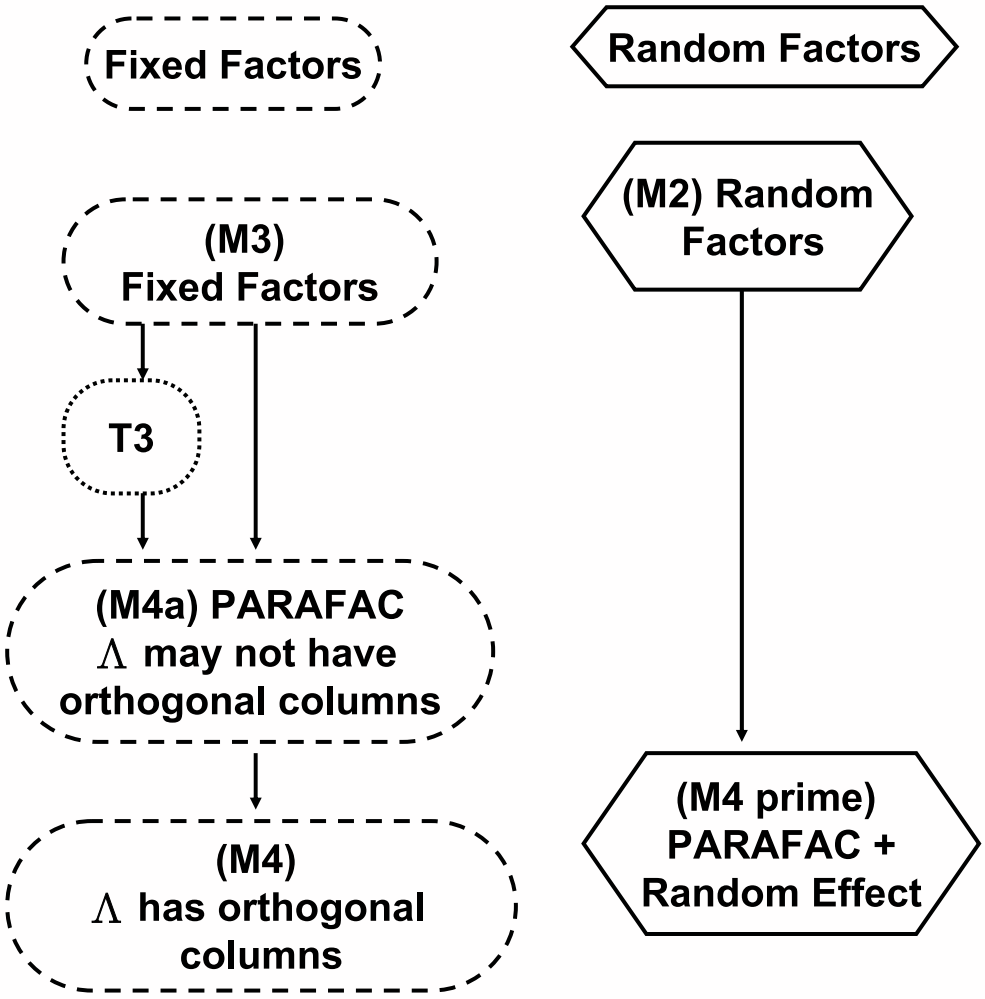

Figure 1.2: Cross-Classified Model Hierarchy.

Chapter 2

Factor Analysis Models and Model Hierarchy

In this chapter, we introduce the general factor models and construct a model hierarchy for application to tongue image data. In each model, we present the model assumptions and the parameter spaces, and then give the proof of identifiability.

2.1 General Factor Analysis Models

Let $\{Y^{(r)}; r = 1, \ldots, R\}$ be an independent sequence of random column vectors of p components with mean μ and covariance matrix Σ_y. Then we say that the q-factor model [15] holds for $Y^{(r)}$ if $Y^{(r)}$ can be written in the form

$$Y^{(r)} = \underline{\mu} + \Lambda \mathbf{f}^{(r)} + U^{(r)} \qquad (M0)$$

where $\underline{\mu}$ is a column vector of p components; Λ is a $p \times q$ matrix of constants with q fixed and less than p; $\mathbf{f}^{(r)}$ is a $q \times 1$ random vector; and $U^{(r)}$ is a $p \times 1$ random vector for $r = 1, \ldots, R$. The elements of Λ are called $factor\ loadings$, the elements of $\mathbf{f}^{(r)}$ are called $common$ factors and the elements of $U^{(r)}$ are called $unique$ factors.

Assume (as in Mardia Kent Bibby [15]) that $\mathbf{f}^{(r)} \sim N(0, I_q)$, $U^{(r)} \sim N(0, \Psi)$, and $\mathbf{f}^{(r)}$ and $U^{(r)}$ are independent where $\Psi = diag(\psi)$ is a $p \times p$ diagonal matrix with the vector $\psi \equiv (\psi_1, \ldots, \psi_p) \in \mathbf{R}^p$ on the diagonal. Therefore, the General Factor Model

can be expressed

$$Y^{(r)} \sim N(\underline{\mu},\ \Lambda\,\Lambda^t +\ \Psi)$$

The observed data always consist of $\{Y^{(r)}; r\ =\ 1, \ldots, R\}$. The parameter $\theta\ =\ (\underline{\mu},\ \Lambda,\ \psi)$ is assumed to belong to the space

$$\Theta_{M0} \equiv \mathbf{R}^p \times O^+_{pq} \times \mathbf{R}^p_+ \tag{2.1}$$

where O^+_{pq} is defined in Notation 1.7 and $\mathbf{R}^p_+$ is defined in Notation 1.3. This model is called model (M0).

2.1.1 Generality of Σ_Y produced by the model

In model (M0), $Y^{(r)}$ is normally distributed with mean μ and covariance

$$\Sigma_Y = \Lambda\Lambda^t + diag(\psi) \tag{2.2}$$

where $\Psi \equiv diag(\psi)$. In this case, there is a problem of existence of the model: for a normal population with mean μ^* and covariance matrix Σ^*, is there a factor model (M0) that can generate this population? The essential question is whether the equation $\Sigma^* = \Lambda\Lambda^t + diag(\psi)$ can be solved, or what condition is needed to solve the equation.

It is of interest to compare the number of parameters in Σ_Y with the number of free parameters in the factor model. There are p elements of ψ and pq elements of Λ. However, in any solution Λ can be replaced by ΛT, where T is any $q \times q$ orthogonal matrix and T has $q(q-1)/2$ independent elements. Thus, a solution $\Lambda \in O^+_{pq}$ must satisfy $q(q-1)/2$ additional column orthogonality constraints. Since the number of distinct elements of Σ_Y is $p(p+1)/2$, we see that the number of covariance parameters minus the number of additional independent constraints is

$$\begin{aligned} C(p,q) &= \frac{1}{2}p(p+1) - [pq + p - \frac{1}{2}q(q-1)] \\ &= \frac{1}{2}[(p-q)^2 - (p+q)] \end{aligned} \tag{2.3}$$

Usually, $C(p,q) > 0$, since p is much larger than q. In general, a solution (Λ, ψ) under the additional constraints can be unique only if $C(p,q) \leq 0$. Setting the quadratic $C(p,q)$ equal to zero and solving for q, the two roots are given by

$$q = \frac{1}{2}[(2p+1) \pm \sqrt{8p+1}] \tag{2.4}$$

For any fixed value of p, the plot of the quadratic function $C(p,q)$ is a parabola which opens up vertically. Hence the values of q such that $C(p,q) \leq 0$ are given by

$$\frac{1}{2}[(2p+1) + \sqrt{8p+1}] \geq q \geq \max(\frac{1}{2}[(2p+1) - \sqrt{8p+1}], 0) \tag{2.5}$$

2.1.2 Identifiability for (M0) model

A parameter θ for a family of probability density functions $\mathcal{P}_\theta \equiv \{p_\theta : \theta \in \Theta\}$ is said to be identifiable if the distinct values of θ correspond to distinct probability densities. That is, θ is identifiable if $\theta \neq \theta'$ implies $p_\theta \neq p_{\theta'}$. The existence of a consistent estimator of a parameter θ (in independent identically distributed samples from $\mathcal{P}_\theta$) implies identifiability of θ.

In the general factor analysis model (M0), $Y^{(r)}$ is assumed to be multivariate normally distributed with mean $\underline{\mu}$ and covariance matrix $\Sigma_Y \equiv \Lambda\Lambda^t + diag(\psi)$. Thus, identifiability of the model requires precisely that the mapping

$$(\underline{\mu}, \Lambda, \psi) \longmapsto (\underline{\mu}, \Lambda\Lambda^t + diag(\psi))$$

be one-to-one. Therefore, given covariance matrix Σ and a number q of factors, we ask whether there exists unique (Λ, ψ) to satisfy (2.2). It is clear that if (Λ,ψ) is a solution of (2.2), then (ΛT,ψ) is also a solution of (2.2), for any $q \times q$ orthogonal matrix T. So the problem is whether we can find constraints such that there is a unique solution under the constraints within O^+_{pq}.

As we count the number of equations and number of free parameters in the previous section, identifiability corresponds roughly to a solution set of dimension 0. However, the counting of equations does not really give enough information for a sufficient condition. We should investigate the problem more fully. Let us first start from observing some examples of non-identifiable models.

Example 2.1. *Let $\{e_1, e_2, \ldots, e_p\}$ denote the canonical basis of $\mathbf{R}^p$, e_j the vector with i-th component δ_{ij}, and let $(\widetilde{\Lambda}, \widetilde{\psi})$ be a solution of equation (2.2) satisfying the conditions that the columns of $\widetilde{\Lambda}$ are orthogonal and the first column of $\widetilde{\Lambda}$ is $a \cdot e_1$ for some scalar a. Let $\lambda^{(j)}$ denote the j-th column of $\widetilde{\Lambda}$ and write $\widetilde{\Lambda} = (a \cdot e_1 | \lambda^{(2)} | \ldots | \lambda^{(q)})$, $\widetilde{\psi}$=$(\psi_1, \ldots, \psi_p)$, and $\Sigma_Y = (\sigma_{ij})$. Substitute them in (2.2); then the (1,1) component of Σ_Y satisfies the equation*

$$\sigma_{11} = a^2 + \psi_1.$$

We can decompose σ_{11} as

$$\sigma_{11} = (a^2 - \epsilon) + (\epsilon + \psi_1) \text{ for any } \epsilon \in (0, a^2).$$

For any $\epsilon \in [-\psi_1, a^2)$, let $\Lambda_\epsilon \equiv (\sqrt{a^2 - \epsilon} \cdot e_1 | \lambda^{(2)} | \ldots | \lambda^{(q)})$ and $\psi_\epsilon \equiv (\epsilon + \psi_1, \psi_2, \ldots, \psi_p)$. Then $(\Lambda_\epsilon, \psi_\epsilon)$ is also a solution of equation (2.2) and is in a neighborhood of $(\widetilde{\Lambda}, \widetilde{\psi})$. Hence, there exist infinitely many solutions of (2.2) in a neighborhood of $(\widetilde{\Lambda}, \widetilde{\psi})$. Thus, the model is non-identifiable.

Example 2.2. *Consider the dimension $p = 2$ and $q = 1$, so that $p < 2q + 1$. Let $\Lambda_0 = (1, 1)^t$, $\Lambda_1 = (\sqrt{1.1}, \sqrt{\frac{1}{1.1}})^t$, $\psi_0 = (1, 1)$, and $\psi_1 = (0.9, \frac{1.2}{1.1})$. Then*

$$\Lambda_0 \Lambda_0^t + diag(\psi_0) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = \Lambda_1 \Lambda_1^t + diag(\psi_1) \tag{2.6}$$

Hence, there exist two solutions (Λ_0, ψ_0) and (Λ_1, ψ_1) of (2.2). Thus, the model is non-identifiable.

Example 2.3. *Consider the dimension $p = 3$ and $q = 2$ again satisfying $p < 2q + 1$. Let*

$$\Lambda_0 = \begin{pmatrix} 1 & 2 \\ 1 & -1 \\ 1 & 1 \end{pmatrix}, \quad \Lambda_1 = \begin{pmatrix} 1.1 & \sqrt{3.99} \\ 1 & \frac{-2.1}{\sqrt{3.99}} \\ 1 & \frac{1.9}{\sqrt{3.99}} \end{pmatrix} \tag{2.7}$$

$\psi_0 = (1, 1, 1)$, and $\psi_1 = (0.8, \frac{3.57}{3.99}, \frac{4.37}{3.99})$. Then

$$\Lambda_0\Lambda_0^t + diag(\psi_0) = \begin{pmatrix} 6 & -1 & 3 \\ -1 & 3 & 0 \\ 3 & 0 & 3 \end{pmatrix} = \Lambda_1\Lambda_1^t + diag(\psi_1) \tag{2.8}$$

Hence, there exist two solutions (Λ_0, ψ_0) and (Λ_1, ψ_1) of (2.2). Thus, the model is non-identifiable.

In Example 2.2, $p = 2$ and $q = 1$. There were 4 parameters to solve for but only 3 equations, which is why there was more than one solution of (2.2). In Example 2.3, $p = 3$ and $q = 2$. There were 8 parameters but only 6 equations. In general, the equation-counting result in (2.3) suggests that there will be identifiability only if $C(p, q) \geq 0$, or $(p - q)^2 \geq p + q$, and $p \geq 2q + 1$ is sufficient for this.

In 1956, Anderson and Rubin [1] gave a sufficient condition for identification of the general factor analysis model as follows:

Theorem 2.4. *A sufficient condition for identification of ψ and Λ up to multiplication on the right by an orthogonal matrix is that if any row of Λ is deleted there remain two disjoint sub-matrices of full rank.*

They also found a sufficient condition for local identification, which we now define.

Definition 2.5. *(Λ, ψ) is said to be locally identifiable within a subset $\mathcal{U}$ of $O_{pq}^+ \times \mathbf{R}_+^p$ if there exists a neighborhood $\mathcal{N}$ of (Λ, ψ) within $O_{pq}^+ \times \mathbf{R}_+^p$ such that $\Lambda\Lambda^t + diag(\psi)$ = $\Lambda_1\Lambda_1^t + diag(\psi_1)$ has the unique solution $(\Lambda_1, \psi_1) = (\Lambda, \psi)$ within $\mathcal{U} \bigcap \mathcal{N}$.*

Now the sufficient condition for local identification proposed by Anderson and Rubin [1] is as follows:

Theorem 2.6. *Let* $\Psi \equiv diag(\psi)$ *and* $\Phi \equiv \Psi - \Lambda(\Lambda^t\ \Psi^{-1}\Lambda)^{-1}\Lambda^t$. *If* $|\phi_{ij}^2| \neq 0$, *that is, the matrix* Ξ *with elements* $\xi_{ij} = \phi_{ij}^2$ *is nonsingular, then* Λ *and* ψ *are locally identified under the restriction that* $\Lambda^t\ \Psi^{-1}\Lambda$ *is diagonal and the non-diagonal elements are different and arranged in descending order of size.*

However, the condition for local identification in the previous theorem is hard to check. We should find other more practical conditions on the parameter space such that the parameter in this parameter space is identifiable under (M0).

We start with a special case, denoted (M0a), of the (M0) model when $\underline{\mu} = \mu\mathbf{1}$:

$$Y = \mu\mathbf{1} + \Lambda\mathbf{f} + U \qquad (M0a)$$

Let $p \geq 2q + 1$ and let the parameter $\theta \equiv (\mu,\ \Lambda,\ \psi)$. We define the parameter space of (M0a) as

$$\Theta_{M0a} \equiv \mathbf{R} \times O_{pq}^{+} \times \mathbf{R}_{+}^{p} \qquad (2.9)$$

where O_{pq}^{+} is defined in Notation 1.7. However, the parameter in the parameter space Θ_{M0a} is not identifiable under (M0a). Thus, we need additional constraints on Λ such that the model (M0a) is identifiable. The constraint could be one of the following two cases:

$$(i) \quad \Lambda \text{ contains a column proportional to } \mathbf{1} \qquad (2.10)$$

$$(ii) \quad \Lambda^t\mathbf{1} = \mathbf{0} \qquad (2.11)$$

Therefore, we can redefine the parameter space as either of the two spaces:

$$\Theta_{M0a1} = \Theta_{M0a} \cap \{\Lambda \text{ contains a column proportional to } \mathbf{1}\} \qquad (2.12)$$

$$\Theta_{M0a2} = \Theta_{M0a} \cap \{\Lambda^t\mathbf{1} = \mathbf{0}\}. \qquad (2.13)$$

We will show identifiability under (M0a).

Lemma 2.7. *The model (M0a) is identifiable in either* Θ_{M0a1} *or* Θ_{M0a2}.

Proof. Suppose there exist two pairs (Λ, ψ), (Λ^*, ψ^*) both in either Θ_{M0a1} or Θ_{M0a2} and satisfying (2.2), and let $\Psi \equiv diag(\psi)$, $\Psi^* \equiv diag(\psi^*)$. Then

$$\Sigma_Y = \Lambda\Lambda^t + \Psi = \Lambda^*\Lambda^{*t} + \Psi^* \tag{2.14}$$

and

$$(\Lambda\Lambda^t - \Lambda^*\Lambda^{*t})\mathbf{1} = (\Psi^* - \Psi)\mathbf{1}. \tag{2.15}$$

Now the left hand side of this equation is either $\mathbf{0}$, if both (Λ, ψ), $(\Lambda^*, \psi^*) \in \Theta_{M0a2}$, or is a constant times $\mathbf{1}$, if both (Λ, ψ), $(\Lambda^*, \psi^*) \in \Theta_{M0a1}$. In the first case, we conclude that the diagonal matrix $\Psi^* - \Psi$ is the zero matrix, and then from (2.14) it follows also that $\Lambda\Lambda^t = \Lambda^*\Lambda^{*t}$. Since both Λ and Λ^* belong to the space O^+_{pq}, it follows that $\Lambda = \Lambda^*$.

In the second case (if (2.10) holds for both Λ, Λ^*), we have $\mathbf{1}$ as an eigenvector of $\Lambda\Lambda^t - \Lambda^*\Lambda^{*t}$ with possibly nonzero eigenvalue, which is a contradiction unless $\Psi^* - \Psi = cI_p$ for some possibly nonzero constant c. Since $p > 2q$, this is possible only if $c = 0$, $\Psi^* = \Psi$, $\Lambda\Lambda^t = \Lambda^*\Lambda^{*t}$ because

$$\text{rank}(\Lambda\Lambda^t),\ \text{rank}(\Lambda^*\Lambda^{*t}) < p/2$$

implies

$$\text{rank}(\Lambda\Lambda^t - \Lambda^*\Lambda^{*t}) < p = \text{rank}(I_p).$$

Thus, we conclude in either case that $\Psi = \Psi^*,\ \Lambda\Lambda^t = \Lambda^*\Lambda^{*t}$. Since both Λ and Λ^* belong to O^+_{pq}, it follows that $\Lambda = \Lambda^*$. $\square$

Remark 2.8. *The vector* $\mathbf{1}$ *could be replaced by any other vector* v_0 *with all non-zero entries which is known in the sense that it is written into the parameter space into conditions like (2.10) or (2.11), playing the same role for both* Λ *and any potential* Λ^* *in (2.14).*

Remark 2.9. *The assumption in (2.13) automatically implies that the canonical basis vectors cannot lie in the column space of* Λ. *In the case (2.12), the restriction to* Λ *matrices containing a column proportional to* $\mathbf{1}$ *means that we have identifiability despite allowing possibly that a canonical basis vector might lie in the column space of* Λ.

2.1.3 Identifiability for (M0) (continued)

Now, let us go back to model (M0) and give some conditions such that the parameter is identifiable under (M0). In Example 2.1, we explained that if Λ contains a column proportional to any element of the canonical basis, then the parameter is not identifiable under (M0). Thus, in order to make the model identifiable, the canonical basis must be excluded from the column space of Λ. Therefore, we have the the following Lemma.

Lemma 2.10. *If* $p \geq 2q+1$ *and if, for some* $\underline{\mu} \in \mathbf{R}^p$, *there exist* $(\underline{\mu}, \Lambda_0, \psi_0)$, $(\underline{\mu}, \Lambda_1, \psi_1)$ $\in \Theta_{M0}$ *defined in (2.1) such that that* $\{e_1, e_2, \ldots, e_p\} \cap col(\Lambda_j) = \emptyset$ *for* $j = 0, 1$, $col(\Lambda_0) \subseteq col(\Lambda_1)$, *and satisfying the condition (2.2), then* (Λ_0, ψ_0)=(Λ_1, ψ_1).

Proof. Suppose there exist $(\underline{\mu}, \Lambda_0, \psi_0)$, $(\underline{\mu}, \Lambda_1, \psi_1)$ for some $\underline{\mu} \in \mathbf{R}^p$ such that $\{e_1, e_2, \ldots, e_p\} \cap col(\Lambda_j)$=$\emptyset$ for $j = 0, 1$, $col(\Lambda_0) \subseteq col(\Lambda_1)$, and satisfying the condition

$$\Sigma_Y = \Lambda_j \Lambda_j^t + diag(\psi_j) \tag{2.16}$$

Let $\Psi_0 \equiv diag(\psi_0)$ and $\Psi_1 \equiv diag(\psi_1)$. Then

$$\Lambda_0\Lambda_0^t - \Lambda_1\Lambda_1^t = \Psi_1 - \Psi_0 \tag{2.17}$$

which implies that the range of $(\Psi_1 - \Psi_0)$ must be contained in the space spanned by the columns of Λ_0 and Λ_1. That is,

$$range(\Psi_1 - \Psi_0) \subseteq span\{col(\Lambda_0), col(\Lambda_1)\}. \tag{2.18}$$

Here $(\Psi_1 - \Psi_0)$ is diagonal since both Ψ_0 and Ψ_1 are diagonal. Thus,

$$range(\Psi_1 - \Psi_0) = span\{e_j : (\Psi_1 - \Psi_0)_{jj} \neq 0\}. \tag{2.19}$$

Through equation (2.18) and (2.19), we have

$$span\{e_j : (\Psi_1 - \Psi_0)_{jj} \neq 0\} \subseteq span\{col(\Lambda_0), col(\Lambda_1)\}. \tag{2.20}$$

Under the restriction $col(\Lambda_0) \subseteq col(\Lambda_1)$, the above equation becomes

$$span\{e_j : (\Psi_1 - \Psi_0)_{jj} \neq 0\} \subseteq col(\Lambda_1). \tag{2.21}$$

This contradicts the assumption that $\{e_1, e_2, \ldots, e_p\} \cap col(\Lambda_1) = \emptyset$. $\square$

The spaces O_{pq} and O_{pq}^+ have been defined in Notation 1.7. We now define more general spaces O_{pq}^* and O_{pq}^{*+}.

Notation 2.11. *Let O_{pq}^* denote the space of $p \times q$ matrices M with real components and orthogonal columns, ordered by non-increasing norm, (i.e., satisfying M^tM = $diag(v_1, v_2, \ldots, v_q)$ with $v_1 \geq v_2 \geq \ldots \geq v_q \geq 0$), and let O_{pq}^{*+} denote the subset of matrices in O_{pq}^* satisfying the additional constraint that the first nonzero element in each column is positive.*

The space O_{pq} is a subspace of O^*_{pq}, and O^+_{pq} is a subspace of ${O^*_{pq}}^+$. Now, we define a more general parameter space. Let $\theta \equiv (\underline{\mu},\ \Lambda,\ \psi)$. The parameter space Θ^*_{M0}, which contains Θ_{M0}, is defined as

$$\Theta^*_{M0} \equiv \mathbf{R}^p \times {O^*_{pq}}^+ \times \mathbf{R}^p_+. \tag{2.22}$$

The parameter $\theta \in \Theta_{M0}$ in (2.1) was shown non-identifiable under (M0) if $e_j \in col(\Lambda)$ for any $j = 1, 2, \ldots, q$. We now have the following Lemma connecting the non-identifiability to a solution of (2.2) in the boundary of the parameter space Θ^*_{M0}.

Lemma 2.12. *If $e_j \in col(\Lambda)$ and if (Λ, ψ) satisfies the condition (2.2), then there exists (Λ^*, ψ^*) in the boundary of the parameter space Θ^*_{M0}, possibly with larger q, and also satisfying the condition (2.2).*

Proof. Since (Λ, ψ) satisfies the condition (2.2), we can decompose Σ_Y as

$$\begin{aligned}
\Sigma_Y &= \Lambda\Lambda^t + diag(\psi) \\
&= (\Lambda\Lambda^t + \psi_j e_j e_j^t) + (-\psi_j e_j e_j^t + diag(\psi)) \\
&= (\Lambda\Lambda^t + \psi_j e_j e_j^t) \\
&\quad + diag(\psi - \psi_{j-1},\ 0,\ \psi_{j+1}, \ldots, \psi_p)
\end{aligned} \tag{2.23}$$

The first term in (2.23) is positive definite and symmetric. By the spectral decomposition theorem, it can be written as

$$\Lambda\Lambda^t + \psi_j e_j e_j^t = \Lambda_1 \Lambda_1^t \tag{2.24}$$

where Λ_1 has orthogonal columns and positive norms of columns, but the norms of columns may not be all distinct and ordered. The norms can be made ordered non-increasing if we multiply Λ_1 by a permutation matrix R from the right. That is,

$$\Lambda^* {\Lambda^*}^t = (\Lambda_1 R)(\Lambda_1 R)^t = \Lambda_1 \Lambda_1^t = \Lambda\Lambda^t + \psi_j e_j e_j^t \tag{2.25}$$

where $\Lambda^* \equiv \Lambda_1 R \in {O^*_{pq}}^+$.

Denote $\Psi \equiv diag(\psi)$ and denote the second term in (2.23) as

$$\Psi^* \equiv diag(\psi^*) \tag{2.26}$$

where $\psi^* \equiv (\psi_1, \ldots, \psi_{j-1}, 0, \psi_{j+1}, \ldots, \psi_p)$. The $p \times p$ diagonal matrix Ψ^* is just like Ψ but the j-th diagonal element is 0. This reduces the number of parameters in Ψ from p to $(p-1)$. Therefore, if $(\Lambda, \psi) \in \Theta_{M0}$ is a solution of $\Sigma_Y = \Lambda\Lambda^t + \Psi$ with $\Psi = diag(\psi)$, then there exists (Λ^*, ψ^*) in the boundary of Θ^*_{M0} such that, with $\Psi^* = diag(\psi^*)$,

$$\Sigma_Y = \Lambda\Lambda^t + \Psi = \Lambda^*{\Lambda^*}^t + \Psi^*. \tag{2.27}$$

That is, if $e_1 \in col(\Lambda)$ and $(\Lambda, \psi) \in \Theta_{M0}$ is a solution of $\Sigma_Y = \Lambda\Lambda^t + \Psi$, then there exists another solution (Λ^*, ψ^*) in the boundary of Θ^*_{M0}. $\square$

Now we will explore a relationship between a non-identifiable model and the parameterization in which not the dimension of Ψ but the column space of Λ is reduced. We need the following Lemma for this purpose.

Lemma 2.13. *Let $A \succeq 0$ be a $p \times p$ symmetric, positive semidefinite matrix (cf Definition 1.9) and let $v \in \mathbf{R}^p$ be a vector in the range of A, $v \neq 0$. Then*

$$\sup\{\alpha \in \mathbf{R} : A - \alpha v v^t \succeq 0\} > 0$$

Proof. Since A is nonnegative definite and symmetric, using the singular value decomposition, A can be decomposed as

$$A = WDW^t \tag{2.28}$$

where $D = diag(d_1, \ldots, d_s)$ and $d_1, \ldots, d_s$ are the non-zero eigenvalues of A with the corresponding unit eigenvectors $w_1, w_2, \ldots, w_s$, and $W = (w_1|w_2|\ldots|w_s)$. Note that $range(A) = span\{w_1, \ldots, w_s\} = col(W)$.

Note that

$$\inf\{x^t Ax : x \in range(A), \|x\| = 1\} = \min_{1 \le k \le s} d_k \tag{2.29}$$

Also, $A : range(A) \to range(A)$ is linear, symmetric, invertible and positive definite. Then any α with $0 < \alpha < \min\{d_j : 1 \le j \le s\}$ results in $A - \alpha I_p : range(A) \to range(A)$ which is invertible and positive definite by (2.28).

Given any vector v in the range of A, we can construct an orthonormal basis $\{v, v_2, v_3, \ldots, v_s\}$ of $range(A)$ such that $I_p = vv^t + \sum_{j=2}^{s} v_j v_j^t$ as an operator on $range(A)$. Therefore, $A - \alpha vv^t$ can be decomposed as

$$A - \alpha vv^t = (A - \alpha I_p) + \alpha \sum_{j=2}^{s} v_j v_j^t \tag{2.30}$$

which is positive definite on $range(A)$ since $A - \alpha I_p$ is positive definite on $range(A)$ and $\alpha \sum_{j=2}^{s} v_j v_j^t$ is nonnegative definite. Let $(range(A))^{\perp}$ denote the orthogonal complement of $range(A)$. Since $A - \alpha vv^t$ maps $(range(A))^{\perp}$ to 0 and $A - \alpha vv^t \succeq 0$ on $range(A)$, we have $A - \alpha vv^t \succeq 0$. $\square$

We can now make a statement on the relationship between non-identifiable models (M0) involving parameters with reduced column space for Λ. We have the following lemma.

Lemma 2.14. *If $(\Lambda, \psi) \in \Theta_{M0}$, and satisfies (2.2), and if $e_j \in col(\Lambda)$, then there exists another solution $(\Lambda^*, \psi^*) \in \Theta^*_{M0}$, which is defined in (2.37), such that e_j does not belong to $col(\Lambda^*)$.*

Proof. If $e_j \in col(\Lambda)$, then we have $e_j \in Range(\Lambda\Lambda^t)$. By Lemma 2.13, there exists a number $\hat{\alpha} \equiv \sup\{\alpha : \Lambda\Lambda^t - \alpha e_j e_j^t \succeq 0\}$ which is positive. Let $Q \equiv \Lambda\Lambda^t - \hat{\alpha} e_j e_j^t$. Then Q is non-negative definite and $e_j \notin col(Q)$ as we shall prove below. The covariance matrix $\Sigma_y = \Lambda\Lambda^t + \Psi$ can be decomposed as

$$\begin{aligned} \Sigma &= \Lambda\Lambda^t + \Psi \\ &= Q + (\hat{\alpha} e_j e_j^t + \Psi) \end{aligned} \tag{2.31}$$

The matrix Q is symmetric and non-negative definite. By the spectral decomposition theorem, and using the same idea as in the proof of Lemma 2.12, it can be written as

$$Q = \Lambda^* \Lambda^{*\,t} \tag{2.32}$$

where $\Lambda^* \in O_{pq}^{*\,+}$.

Next, we show that e_j does not belong to $range(Q)$. If $e_j \in range(Q)$, then by Lemma 2.13, there exists $\alpha > 0$ such that $(Q - \alpha e_j e_j^t) \succeq 0$ which contradicts the definition of Q. Therefore, e_j does not belong to $range(Q) = col(\Lambda^*)$. Hence, Λ^* does not contain e_j in its column space. $\square$

We defined the parameter spaces Θ_{M0} in (2.1) and Θ_{M0}^* in (2.37). To prevent confusion in the dimension of $col(\Lambda)$, we redefine the notations

$$\Theta_{M0}^q \equiv \Theta_{M0} \text{ and } \Theta_{M0}^{*\,q} \equiv \Theta_{M0}^* \tag{2.33}$$

to specify $dim(col(\Lambda)) = q$ in Θ_{M0} and Θ_{M0}^*, respectively.

Based on Lemma 2.12 and Lemma 2.14, we conclude that, if $e_j \in col(\Lambda)$ and (Λ, Ψ) is a solution of $\Sigma = \Lambda\Lambda^t + \Psi$, then there exist two other solutions (Λ^*, Ψ^*). One has $\Psi^* = diag(\psi_1, \ldots, \psi_{j-1},\ 0\ , \psi_{j+1}, \ldots, \psi_p)$ in the boundary, and for the other, Λ^* does not contain e_j in its column space. Therefore, we have the following lemma.

Lemma 2.15. *If $p > 2q$ and $\Sigma_Y = \Lambda\Lambda^t + diag(\psi)$ for model (M0) parameters which are non-identifiable, then there exists (Λ^*, ψ^*) in the boundary of the parameter space $\Theta_{M0}^{*\,\widetilde{q}}$, for some $\widetilde{q} \geq q$, satisfying the condition (2.2).*

Proof. If the model is non-identifiable, then there exist two distinct pairs (Λ_0, ψ_0), (Λ_1, ψ_1), with $\Psi_0 \equiv diag(\psi_0)$ and $\Psi_1 \equiv diag(\psi_1)$, satisfying the condition (2.2). Then we have

$$\Sigma_Y = \Lambda_0\Lambda_0^t + \Psi_0 = \Lambda_1\Lambda_1^t + \Psi_1. \tag{2.34}$$

Given any $s \in (0, 1)$, the convex mixture $(1 - s)(\Lambda_0\Lambda_0^t) + s(\Lambda_1\Lambda_1^t)$ is non-negative definite. Therefore, using the Singular Value Decomposition theorem, we can define $\Lambda_s \in O_{pq}^{*\,+}$ such that

$$\Lambda_s\Lambda_s^t \equiv (1 - s) \cdot (\Lambda_0\Lambda_0^t) + s \cdot (\Lambda_1\Lambda_1^t). \tag{2.35}$$

Also, define

$$\psi_s \equiv (1 - s) \cdot \psi_0 + s \cdot \psi_1 \text{ and } \Psi_s \equiv diag(\psi_s) \tag{2.36}$$

Then (Λ_s, ψ_s) is also a solution of (2.2). Let $q_s \equiv dim(col(\Lambda_s))$. Note that $q_0 = q$. Then $col(\Lambda_0) \subseteq col(\Lambda_s)$ and $q \leq q_s$ by (2.35). Applying Lemma 2.10, there must exist $e_j \in col(\Lambda_s)$ for some j. Then, by Lemma 2.12, there exists (Λ^*, ψ^*) in the boundary of the parameter space $\Theta_{M0}^{*\,q_s}$ also satisfying the condition (2.2), where

$$\Theta_{M0}^{*\,q_s} \equiv \mathbf{R}^p \times O_{p,q_s}^{*\,+} \times \mathbf{R}_+^p. \tag{2.37}$$

which is the same as Θ_{M0}^*, but with possibly different dimension of $col(\Lambda)$. $\square$

2.2 Factor Model (M1)

Consider the special case of (M0) when $\Psi = \sigma^2 I_p$ and $\mu = 0$:

$$Y = \Lambda \mathbf{f} + U \tag{$M1$}$$

where $\mathbf{f} \sim N(0, I_q)$, $U \sim N(0, \sigma^2 I_p)$, and where Λ is a $p \times q$ matrix such that $\Lambda^t \Lambda$ is diagonal with distinct ordered-decreasing elements.

Under (M1), the covariance Σ can be expressed in terms of Λ and σ^2 through the equation

$$\Sigma = \Lambda \Lambda^t + \sigma^2 I_p \tag{2.38}$$

Now, let us define the parameter space for (M1).

Let $\theta = (\Lambda, \sigma^2)$. We define the parameter space as

$$\Theta_{M1} = O^+_{pq} \times \mathbf{R}_+ \tag{2.39}$$

where $\mathbf{R}_+$ denotes the set of all positive real numbers. We first show that our parameter θ is identifiable from the observed data in (M1).

Lemma 2.16. *Model (M1) is identifiable if the parameter θ is assumed to belong to Θ_{M1}.*

Proof. In model (M1), the covariance matrix of $Y^{(r)}$ is given by (2.38). Here σ^2 can be identified by the minimum eigenvalue of Σ_y since $q < p$. Therefore, $\Lambda\Lambda^t$ is identifiable. By the uniqueness of the Singular Value Decomposition, Λ is identified in O^+_{pq}. Therefore, model (M1) is identifiable. $\square$

2.3 Factor Model (M1R)

Consider the reduced form of the (M1) model:

$$Y = \Lambda \mathbf{f} + U \qquad (M1R)$$

where $\mathbf{f} \sim N(0, I_q)$, $U \sim N(0, \Psi)$, Λ is a p by q matrix and

$$\Psi = \begin{pmatrix} 0_r & O \\ O^t & \sigma^2 I_{p-r} \end{pmatrix} \qquad (2.40)$$

where $r < q < p$, 0_r is the $r \times r$ zero matrix, O is a $r \times (p-r)$ zero matrix and O^t denotes the transpose of O. Under (M1R), the covariance parameter Σ_Y can be expressed in terms of Λ and Ψ through (2.2). For simplicity of notation, partition

$$\Sigma_Y = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \text{ and } \Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} \qquad (2.41)$$

where Σ_{11}, Σ_{12} and Σ_{22} are $r \times r$, $r \times (p-r)$ and $(p-r) \times (p-r)$ sub-matrices of Σ_Y, respectively, and Λ_{11}, Λ_{12}, Λ_{21} and Λ_{22} are $r \times r$, $r \times (q-r)$, $(p-r) \times r$ and $(p-r) \times (q-r)$ sub-matrices of Λ. Now, let us define the parameter space for (M1R).

Let $\theta = (\Lambda, \sigma^2)$ where Λ is partitioned as in (2.41). We define the parameter space as

$$\begin{aligned} \Theta_{M1R} = \{ \quad \theta = (\Lambda, \sigma^2) : \quad & \Lambda_{11} \in O^+_{rr},\ \Lambda_{12} = 0,\ \Lambda_{21} \in \mathcal{M}_{p-r,r}, \\ & \Lambda_{22} \in O^+_{p-r,\, q-r},\ 0 < \sigma^2 < \infty \}. \end{aligned} \qquad (2.42)$$

where $\mathcal{M}_{ab}$ is defined in Notation 1.1. Thus, Θ_{M1R} is a subset of $\mathbf{R}^{pq} \times \mathbf{R}_+$. We next show that our parameter θ is identifiable from the observed data in (M1R).

Theorem 2.17. *The parameter $\theta \equiv (\Lambda, \sigma^2) \in \Theta_{M1R}$ is identifiable under model (M1R).*

Proof. Write $\Sigma_Y = \Lambda\Lambda^t + \Psi$, and partition Σ_Y and Λ as in (2.41), obtaining

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} \Lambda_{11}\Lambda_{11}^t + \Lambda_{12}\Lambda_{12}^t & \Lambda_{11}\Lambda_{21}^t + \Lambda_{12}\Lambda_{22}^t \\ \Lambda_{21}\Lambda_{11}^t + \Lambda_{22}\Lambda_{12}^t & \Lambda_{21}\Lambda_{21}^t + \Lambda_{22}\Lambda_{22}^t + \sigma^2 I_{p-r} \end{pmatrix}$$

Thus, $\Sigma_{11} = \Lambda_{11}\Lambda_{11}^t + \Lambda_{12}\Lambda_{12}^t$. Since $\Lambda_{12} = 0$, we have $\Sigma_{11} = \Lambda_{11}\Lambda_{11}^t$. By the uniqueness of singular value decomposition of $\Lambda_{11} \in O_{rr}^+$, Λ_{11} is uniquely determined such that $\Lambda_{11}^t\Lambda_{11} = B$ where $B \equiv diag(b_1, \ldots, b_r)$ with $b_1 > b_2 > \ldots > b_r > 0$. Moreover, since $\Lambda_{12} = 0$, also $\Sigma_{21} = \Lambda_{21}\Lambda_{11}^t + \Lambda_{22}\Lambda_{12}^t = \Lambda_{21}\ \Lambda_{11}^t$. Multiplying the last equation by Λ_{11} from the right, we have

$$\Sigma_{21}\Lambda_{11} = \Lambda_{21}\Lambda_{11}^t\Lambda_{11} = \Lambda_{21}B. \tag{2.43}$$

Therefore,

$$\Lambda_{21} = \Sigma_{21}\Lambda_{11}(\Lambda_{11}^t\Lambda_{11})^{-1} = \Sigma_{21}\Lambda_{11}B^{-1} \tag{2.44}$$

is also uniquely determined. Since Λ_{11} has full rank, $(\Lambda_{11}^t\Lambda_{11})^{-1}$=$(\Lambda_{11})^{-1}\ (\Lambda_{11}^t)^{-1}$. Then (2.44) can be simplified as

$$\Lambda_{21} = \Sigma_{21}(\Lambda_{11}^t)^{-1}. \tag{2.45}$$

Finally,

$$\Sigma_{22} \;=\; \Lambda_{21}\Lambda_{21}^t + \Lambda_{22}\Lambda_{22}^t + \sigma^2 I_{p-r}. \tag{2.46}$$

Substituting $\Lambda_{21} = \Sigma_{21}(\Lambda_{11}^t)^{-1}$ in equation (2.46), we have

$$\begin{aligned} \Sigma_{22} &= \Sigma_{21}(\Lambda_{11}^t)^{-1}\Lambda_{11}^{-1}\Sigma_{21}^t + \Lambda_{22}\Lambda_{22}^t + \sigma^2 I_{p-r} \\ &= \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} + \Lambda_{22}\Lambda_{22}^t + \sigma^2 I_{p-r}. \end{aligned} \tag{2.47}$$

Subtract $\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ from (2.47) on both sides, leaving

$$\begin{aligned} \Lambda_{22}\Lambda_{22}^t + \sigma^2 I_{p-r} &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \\ &= \Sigma_{22.1}. \end{aligned} \tag{2.48}$$

Next we show that $\Lambda_{22}\Lambda_{22}^t$ and $\Lambda_{22}^t\Lambda_{22}$ have the same the nonzero eigenvalues, and the nonzero eigenvalues are distinct. Applying the singular value decomposition of Λ_{22}, we have $\Lambda_{22} = UDV^t$, where U ($(p-r)\times(q-r)$) and V ($(q-r)\times(q-r)$) are column orthonormal matrices ($U^tU = V^tV = I_{q-r}$) and D=$diag(d_1,\ldots,d_{q-r})$ with positive elements d_j for $1 \le j \le (q-r)$. Since $\Lambda_{22} \in O^+_{p-r,q-r}$, $d_1 \ge \ldots \ge d_{q-r}$. Then $\Lambda_{22}\Lambda_{22}^t = UD^2U^t$ and $\Lambda_{22}^t\Lambda_{22} = VD^2V^t$. Therefore, $\Lambda_{22}\Lambda_{22}^t$ and $\Lambda_{22}^t\Lambda_{22}$ have the same the nonzero eigenvalues d_1^2, ..., d_{q-r}^2, and

$$d_1^2 \ge \ldots \ge d_{q-r}^2. \tag{2.49}$$

Let $\Sigma_{22.1} = Q\Delta Q^t$ be the singular value decomposition of $\Sigma_{22.1}$, where $Q \in \mathcal{M}_{p-r,q-r}$, Δ=$diag(\delta_1,\delta_2,\ldots,\delta_{q-r})$. The values $\delta_1,\delta_2,\ldots,\delta_{q-r}$ are the eigenvalues of $\Sigma_{22.1}$ and the columns of Q are the corresponding standardized eigenvectors. By (2.48) and (2.49), we have $U = Q$ and the eigenvalues of $\Sigma_{22.1}$, $\{\delta_j = d_j^2+\sigma^2, 1 \le j \le q-r\}$, are all distinct. Then σ^2 can be identified by the minimum eigenvalue δ_{q-r}, and then

$$\begin{aligned}\Lambda_{22}\Lambda_{22}^t &= \Sigma_{22.1} - \sigma^2 I_{p-r} \\ &= Q\Delta Q^t - Q\sigma^2 I_{p-r} Q^t \\ &= Q(\Delta - \sigma^2 I_{p-r})Q^t\end{aligned}$$

is identifiable. By the uniqueness of singular-value-decomposition of $\Sigma_{22.1}$, Λ_{22} is identifiable in $O^+_{p-r,q-r}$. $\square$

Remark 2.18. *The matrix Λ in the parameter $\theta \in \Theta_{M1R}$ which is defined in (2.42) does not have orthogonal columns any more. We can always transform Λ to have orthogonal columns by applying the singular-value-decomposition to $\Lambda\Lambda^t$.*

2.4 Cross-Classified Factor Model (M2)

In this section, we start to formulate a cross-classified or multi-group model. Let $\{y^{(r,m)} : r = 1, \ldots, R; m = 1, \ldots, M\}$ be a set of vector observations, where the observations $y^{(r,m)}$ are vectors in $\mathbf{R}^p$ which represent vector measurements on an experimental system ; let $r = 1, \ldots, R$ index the identically distributed replications and $m = 1, 2, \ldots, M$ index the experimental settings. We are interested in the common situation where experimental settings are doubly indexed by (a, s), for example, to reflect cross-classification by *treatment* and *subject*. The models we consider all have the following structure:

$$Y^{(r,a,s)} = \begin{pmatrix} y_{1ras} \\ \vdots \\ y_{pras} \end{pmatrix} = \mu_{as} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \Lambda \mathbf{f}^{(r,a,s)} + U^{(r,a,s)} \tag{2.50}$$

That is,

$$y_{iras} = \mu_{as} + \sum_{k=1}^{q} \lambda_{ik} f_{kras} + u_{iras} \tag{2.51}$$

We assume that $\mathbf{f}^{(r,a,s)} \sim N(G^{as}, I_q)$, and $F^{(r)}_{q \times AS} = (\mathbf{f}^{(r,1,1)} | \mathbf{f}^{(r,1,2)} | \cdots | \mathbf{f}^{(r,A,S)})$ is a sequence of $q \times AS$ matrices where $r = 1, \ldots, R$. The elements of $U^{(r,a,s)}$, u_{iras} are independent $N(0, \sigma^2_{u,as})$. Now let us define the parameter space for (M2) by

$$\begin{aligned} \Theta_{M2} \;=\; & \{(\mu, \Lambda, G, \psi) : \Lambda \in O^+_{pq} \text{ and } \Lambda^t \mathbf{1} = 0, \\ & \mu = \{\mu_{as} : a = 1, \ldots, A, s = 1, \ldots, S\} \in \mathbf{R}^{AS}, \\ & \psi = (\sigma^2_{u,as}, (a, s) \in \{1, \ldots, A\} \times \{1, \ldots, S\}) \in \mathbf{R}^{AS}_+, \\ & G = (G^{as}) \in \mathcal{M}_{q,AS}\}. \end{aligned} \tag{2.52}$$

The model in (2.50) under this parameter space is called model (M2). We will show that the parameter θ is identifiable from the observed data in (M2).

Lemma 2.19. *Model (M2) is identifiable.*

Proof. In model (M2), the expectation of $Y^{(r,a,s)}$ is

$$E(Y^{(r,a,s)}) = \mu_{as}\mathbf{1} + \Lambda G^{(a,s)}, \tag{2.53}$$

and the covariance matrix of $Y^{(r,a,s)}$ is

$$\Sigma_Y^{(a,s)} = \Lambda\Lambda^t + \sigma_{u,as}^2 I. \tag{2.54}$$

Here μ_{as} can be estimated consistently by

$$\widetilde{\mu_{as}} = \frac{1}{pR}\sum_{r=1}^{R}\sum_{i=1}^{p} y_{iras}$$

when $R \to \infty$. In equation (2.54), $\sigma_{u,as}^2$ can be identified by the minimum eigenvalues of $\Sigma_Y^{(a,s)}$ since $q < p$. Therefore, $\Lambda\Lambda^t$ is also identifiable.

Next, we want to identify Λ. Using the uniqueness of the Singular Value Decomposition in Lemma A.3 for $\Lambda \in O_{pq}^+$, we can uniquely determine Λ.

Finally, we want to identify $G^{(a,s)}$. Multiplying (2.53) by Λ^t from the left on both sides and using the fact that the columns of Λ are orthogonal to the $\mathbf{1}$ vector, we have

$$E(\Lambda^t Y^{(r,a,s)}) = \Lambda^t\Lambda G^{(a,s)}. \tag{2.55}$$

Therefore, $G^{(a,s)} = E[(\Lambda^t\Lambda)^{-1}(\Lambda^t Y^{(r,a,s)})]$ is identifiable. □

2.5 Cross-Classified Factor Model (M3)

In (M2), we are interested in the common situation where experimental settings are doubly indexed by (a, s), for example, to reflect cross-classification by *treatment* and *subject*. The common factors $\mathbf{f}^{(r,a,s)}$ in (M0)-(M2) are considered as *random effects*,

that is, random variables. Then the specific $\mathbf{f}^{(r,a,s)}$ would not be of interest in themselves, because another set of batches in a subsequent experiment would provide different $\mathbf{f}^{(r,a,s)}$. What might be of interest is the size of the variation in the $\mathbf{f}^{(r,a,s)}$. In (M3), we will think of the factors $\mathbf{f}^{(r,a,s)}$ as being *fixed effects*, each associated specifically with one of the experimental settings. Then the specific $\mathbf{f}^{(r,a,s)}$ would be of interest.

In the case of fixed factors, we assume that $\mathbf{f}^{(r,a,s)} \equiv \mathbf{f}^{(a,s)}$, and

$$F = (\mathbf{f}^{(1,1)}|\mathbf{f}^{(1,2)}|\cdots|\mathbf{f}^{(A,S)})$$

defines a non-random $q \times AS$ matrix. Define the parameter space to be

$$\begin{aligned}
\Theta_{M3} \;=\; & \{(\mu, \Lambda, F, \psi) : \Lambda \in O^{+}_{pq} \text{ and } \Lambda^t \mathbf{1} = 0, \\
& \mu = \{\mu_{as} : a = 1, \ldots, A, s = 1, \ldots, S\} \in \mathbf{R}^{AS}, \\
& \psi = (\sigma^2_{u,as}, (a,s) \in \{1, \ldots, A\} \times \{1, \ldots, S\}) \in \mathbf{R}^{AS}_{+}, \\
& F = (\mathbf{f}^{(a,s)}) \in \mathcal{M}_{q,AS} \text{ with rows orthonormal}, \\
& rank(F) = rank(\Lambda) = rank(\Lambda F) = q\}. \qquad (2.56)
\end{aligned}$$

The model in (2.50) under this parameter space is called model (M3). We will show that our parameter $\theta \in \Theta_{M3}$ is identifiable from the observed data in (M3).

Lemma 2.20. *Model (M3) is identifiable.*

Proof. In model (M3), the expectation of $Y^{(r,a,s)}$ is

$$E(Y^{(r,a,s)}) = \mu_{as}\mathbf{1} + \Lambda \mathbf{f}^{(a,s)} \qquad (2.57)$$

and the covariance matrix of $Y^{(r,a,s)}$ is

$$\Sigma_Y^{(a,s)} = \sigma^2_{u,as} I. \qquad (2.58)$$

Here $\sigma^2_{u,as}$ can be identified by the minimum eigenvalues of $\Sigma_Y^{(a,s)}$ since $q < p$. In equation (2.57), the parameter μ_{as} can be estimated consistently by

$$\widetilde{\mu_{as}} = \frac{1}{pR} \sum_{r=1}^{R} \sum_{i=1}^{p} y_{iras}$$

when $R \to \infty$. Therefore, $\Lambda \mathbf{f}^{(a,s)}$ is also identifiable. That is, ΛF is identifiable.

Now we identify Λ and F. Note that Λ and F are $p \times q$ and $q \times AS$ matrices, respectively. Applying the singular value decomposition on ΛF, there exist unique matrices U, D and V (expect for possible changes of sign of the columns) such that

$$\Lambda F = U\, D\, V^t \tag{2.59}$$

where $D = diag(d_1, \ldots, d_q)$ such that $d_1^2, \ldots, d_q^2$ are eigenvalues of $\Lambda F (\Lambda F)^t$, U is a $p \times q$ matrix and the columns of U are standardized eigenvectors of $\Lambda F (\Lambda F)^t$, and V is a $AS \times q$ matrix with columns are standardized eigenvectors of $(\Lambda F)^t\ \Lambda F$. Since F has orthonormal rows in Θ_{M3}, by the uniqueness of singular value decomposition in (2.59), Λ and F are uniquely determined. Therefore, the model (M3) is identifiable. □

2.6 Factor Model (M4)

Assume that the fixed effects $\mathbf{f}^{(a,s)}$ now have the factorized form

$$\mathbf{f}^{(a,s)} = \begin{pmatrix} f_{1as} \\ \vdots \\ f_{qas} \end{pmatrix} \text{ and } f_{kas} = w_{ak} v_{sk} \text{ for } k = 1, \ldots, q \tag{2.60}$$

The parameter space is defined as

$$\begin{aligned}\Theta_{M4} \;=\; & \{(\mu,\Lambda,W,V,\psi) : \Lambda \in O^{+}_{pq} \text{ and } \Lambda^t\mathbf{1}=0,\\ & \mu=\{\mu_{as} : a=1,\ldots,A, s=1,\ldots,S\}\in \mathbf{R}^{AS},\\ & \psi=(\sigma^2_{u,as},(a,s)\in\{1,\ldots,A\}\times\{1,\ldots,S\})\in \mathbf{R}^{AS}_{+},\\ & W=(w_{ak})\in\mathcal{M}^{+}_{Aq} \text{ with } \sum_{a=1}^{A} w^2_{ak}=1,\\ & V=(v_{sk})\in\mathcal{M}_{Sq} \text{ with } \sum_{s=1}^{S} v^2_{sk}=1,\\ & rank(\Lambda)=rank(W)=rank(V)=q\le A,S\}. \end{aligned} \tag{2.61}$$

where $\mathcal{M}_{Aq}$ and $\mathcal{M}^{+}_{Aq}$ are defined in Notation 1.1. The model in (2.50) under this parameter space is called (M4) model, which can be written as

$$y_{iras}=\mu_{as}+\sum_{k=1}^{q}\lambda_{ik}w_{ak}v_{sk}+u_{iras}. \tag{$M4$}$$

We next show that our parameter θ is identifiable from the observed data Y in (M4).

Lemma 2.21. *Model (M4) is identifiable.*

Proof. In model (M4), the expectation of $Y^{(r,a,s)}$ is

$$E(Y^{(r,a,s)})=\mu_{as}\mathbf{1}+\Lambda\begin{pmatrix} w_{a1}v_{s1}\\ \vdots\\ w_{aq}v_{sq}\end{pmatrix} \tag{2.62}$$

and the covariance matrix of $Y^{(r,a,s)}$ is

$$\Sigma_Y^{(a,s)}=\sigma^2_{u,as}I. \tag{2.63}$$

Here $\sigma^2_{u,as}$ can be identified by the minimum eigenvalue of $\Sigma_Y^{(a,s)}$ for each (a,s) since $p>q$. Then μ_{as} can be estimated consistently by

$$\widetilde{\mu_{as}}=\frac{1}{pR}\sum_{r=1}^{R}\sum_{i=1}^{p}y_{iras}$$

when $R \to \infty$, and then $\Lambda \begin{pmatrix} w_{a1}v_{s1} \\ \vdots \\ w_{aq}v_{sq} \end{pmatrix}$ is identified through the equation (2.62). That is, $\sum_{k=1}^{q} \lambda_{ik} w_{ak} v_{sk}$ is identified for all $a = 1, \ldots, A, s = 1, \ldots, S, i = 1, \ldots, p$.

Next, we identify Λ, W and V using Jennrich's Basic Uniqueness Theorem stated in Lemma A.4. Suppose that there exist $\lambda_{ik}, w_{ak}, v_{sk}$ and $\lambda^*_{ik}, w^*_{ak}, v^*_{sk}$ such that

$$\sum_{k=1}^{q} \lambda_{ik} w_{ak} v_{sk} = \sum_{k=1}^{q} \lambda^*_{ik} w^*_{ak} v^*_{sk}$$

where $\Lambda^t \Lambda = \text{diag}(b_1, \ldots, b_q)$, $b_1 > b_2 > \ldots > b_q > 0$, ${\Lambda^*}^t \Lambda^* = \text{diag}(b^*_1, \ldots, b^*_q)$, $b^*_1 > b^*_2 > \ldots > b^*_q > 0$, $\sum_{a=1}^{A} w^2_{ak} = 1 = \sum_{a=1}^{A} (w^*_{ak})^2$, $\sum_{s=1}^{S} v^2_{sk} = 1 = \sum_{s=1}^{S} (v^*_{sk})^2$, and $rank(\Lambda) = rank(W) = rank(V) = rank(\Lambda^*) = rank(W^*) = rank(V^*) = q \leq A, S$. By Jennrich's Uniqueness Theorem, we have

$$\Lambda^* = \Lambda R D_1, W^* = W R D_2, V^* = V R D_3 \tag{2.64}$$

where R is a permutation matrix and D_1, D_2, D_3 are diagonal matrices with $D_1 D_2 D_3 = I$. Let $w^{(k)}$ be the k-th column vector of W, $w^{(k)*}$ be the k-th column vector of W^* and $R = (R_{jk})$. Since $W^* = WRD_2$, we have

$$(w^{(1)*} | w^{(2)*} | \cdots | w^{(q)*}) = (w^{(1)} | w^{(2)} | \cdots | w^{(q)}) R D_2$$

where RD_2 is a row permuted matrix of D_2. This implies $w^{(k)*} = (D_2)_{kk} w^{(j)}$ for some j which depends on the permutation matrix R such that $R_{kj} = 1$. Using the condition that $\sum_{a=1}^{A} w^2_{ak} = 1 = \sum_{a=1}^{A} (w^*_{ak})^2$, we have

$$1 = \sum_{a=1}^{A} (w^*_{ak})^2 = \|w^{(k)*}\|^2 = (D_2)^2_{kk} \|w^{(j)}\|^2 = (D_2)^2_{kk}. \tag{2.65}$$

Therefore, $D_2 = diag(d^{(2)}_1, d^{(2)}_2, \cdots, d^{(2)}_q)$ where $d^{(2)}_k = +1$ or -1. Similarly, we have $D_3 = diag(d^{(3)}_1, d^{(3)}_2, \cdots, d^{(3)}_q)$ where $d^{(3)}_k = +1$ or -1. Then $D_1 D_2 D_3 = I$

implies $1 = d_k^{(1)} d_k^{(2)} d_k^{(3)}$ where $D_1 = diag(d_1^{(1)}, d_2^{(1)}, \cdots, d_q^{(1)})$. Therefore, $d_k^{(1)} = +1$ or -1. By the conditions that $\Lambda^t\Lambda =$ diag($b_1, \ldots, b_q$), $b_1 > b_2 > \ldots > b_q > 0$, and $(\Lambda^*)^t(\Lambda^*) =$ diag($b_1^*, \ldots, b_q^*$), $b_1^* > b_2^* > \ldots > b_q^* > 0$, we have

$$\begin{aligned} diag(b_1^*, \ldots, b_q^*) &= (\Lambda^*)^t \Lambda^* \\ &= (\Lambda R D_1)^t (\Lambda R D_1) \\ &= D_1 R^t \, diag(b_1, \ldots, b_q) R D_1 \\ &= (D_1)^2 R^t \, diag(b_1, \ldots, b_q) R \\ &= R^t \, diag(b_1, \ldots, b_q) R. \end{aligned} \tag{2.66}$$

Applying the uniqueness of the Singular Value Decomposition Theorem to equation (2.66), we have $R = I$. Therefore, $\Lambda^* = \Lambda R D_1 = \Lambda D_1$ where D_1 is a diagonal matrix with $+1$ or -1 as the diagonal elements. Since we assume that the first nonzero element in each column of Λ and Λ^* is positive, then $D_1 = I$. Similarly, since $W^* = W D_2$ and the first nonzero element in each column of W and W^* is positive, $D_2 = I$ and $W = W^*$.

Finally, V is identified once $D_1 = D_2 = I$ and $D_1 D_2 D_3 = I$, so that $D_3 = I$ and $V^* = V D_3 = V_3$. $\square$

2.7 PARAFAC Model (M4a)

In this section, we consider a model which is similar to (M4) having fixed common factors $\mathbf{f}^{(a,s)}$ but without the orthogonality of columns of loading matrix Λ. Consider the model

$$Y^{(r,a,s)} = \mu_{as} \, \mathbf{1} + \Lambda_\star \, \mathbf{f}^{(a,s)} + U^{(r,a,s)} \tag{2.67}$$

where $\Lambda_\star$ is a $p \times q$ matrix with non-orthogonalized columns. Assume that the fixed factors $\mathbf{f}^{(a,s)}$ can be written as

$$\mathbf{f}^{(a,s)} = \begin{pmatrix} f_{1as} \\ \vdots \\ f_{qas} \end{pmatrix} \text{ and } f_{kas} = w_{ak} v_{sk} \text{ for } k = 1, \ldots, q. \tag{2.68}$$

The elements of $U^{(r,a,s)}$, u_{iras} are independent $N(0, \sigma^2_{u,as})$. We define the parameter space

$$\begin{aligned} \Theta_{M4a} \;=\; & \{(\mu, \Lambda_\star, W, V, \psi) : \mu = \{\mu_{as} : a = 1, \ldots, A, s = 1, \ldots, S\} \in \mathbf{R}^{AS}, \\ & \Lambda_\star \in \mathcal{M}^+_{pq},\ \Lambda^t_\star \mathbf{1} = 0, \text{ with column norms in decreasing order,} \\ & \psi = (\sigma^2_{u,as}, (a,s) \in \{1, \ldots, A\} \times \{1, \ldots, S\}) \in \mathbf{R}^{AS}_+, \\ & W = (w_{ak}) \in \mathcal{M}^+_{Aq} \text{ with } \sum_{a=1}^{A} w^2_{ak} = 1, \\ & V = (v_{sk}) \in \mathcal{M}_{Sq} \text{ with } \sum_{s=1}^{S} v^2_{ks} = 1, \\ & rank(\Lambda_\star) = rank(W) = rank(V) = q \leq A, S\}. \end{aligned} \tag{2.69}$$

Model in (5.2) under this parameter space Θ_{M4a} is called the PARAFAC model [**?**], denoted by (M4a), which can be written as

$$y_{iras} = \mu_{as} + \sum_{k=1}^{q} (\lambda_{ik\star}) w_{ak} v_{sk} + u_{iras}. \tag{$M4a$}$$

Lemma 2.22. *Model (M4a) is identifiable.*

Proof. Using the same proof as in Lemma 2.21, we can identify μ, $\Psi^{(a,s)}$, and then $\Lambda_\star \begin{pmatrix} w_{a1} v_{s1} \\ \vdots \\ w_{aq} v_{sq} \end{pmatrix}$ is identified. That is, $\sum_{k=1}^{q} \lambda_{ik\star} w_{ak} v_{sk}$ is identified for all $a = 1, \ldots, A, s = 1, \ldots, S, i = 1, \ldots, p$.

Suppose that there exist $\lambda_{ik\star}, w_{ak}, v_{sk}$ and $\bar{\lambda}_{ik\star}, \bar{w}_{ak}, \bar{v}_{sk}$ such that

$$\sum_{k=1}^{q} \lambda_{ik\star} w_{ak} v_{sk} = \sum_{k=1}^{q} \bar{\lambda}_{ik\star} \bar{w}_{ak} \bar{v}_{sk}$$

where both $\Lambda_\star$ and $\bar{\Lambda}_\star$ are in $\mathcal{M}_{pq}^+$ and have column norms ordered decreasing, $\sum_{a=1}^A w_{ak}^2 = 1 = \sum_{a=1}^A (\bar{w}_{ak})^2$, $\sum_{s=1}^S v_{sk}^2 = 1 = \sum_{s=1}^S (\bar{v}_{sk})^2$, $rank(\Lambda_\star) = rank(W) = rank(V) = rank(\bar{\Lambda}_\star) = rank(\bar{W}) = rank(\bar{V}) = q \leq A, S$. By Jennrich's Uniqueness Theorem (Lemma A.4), we have

$$\bar{\Lambda}_\star = \Lambda_\star R D_1,\ \bar{W} = W R D_2,\ \bar{V} = V R D_3 \tag{2.70}$$

where R is a permutation matrix and D_1, D_2, D_3 are diagonal matrices with $D_1 D_2 D_3 = I$. Following the same proof as in Lemma 2.21 and using (2.65), implies D_i is a diagonal matrix with $+1$ or -1 as the diagonal elements, for $i = 1, 2, 3$. Since $\bar{\Lambda}_\star = \Lambda_\star R D_1$ and we assume that, in Θ_{M4a}, the first nonzero element in each column of $\Lambda_\star$ and $\bar{\Lambda}_\star$ is positive, then $D_1 = I$. Also we assume that the column norms of both $\Lambda_\star$ and $\bar{\Lambda}_\star$ are ordered decreasing, this implies the permutation $R = I_q$. Therefore, $\bar{\Lambda}_\star = \Lambda_\star$. Similarly, we have $\bar{W} = W D_2$ and the first nonzero element in each column of W is positive, so that $D_2 = I$ and $\bar{W} = W$. Finally, V is identified once $D_1 = D_2 = I$ and $D_1 D_2 D_3 = I$, so that $D_3 = I$ and $\bar{V} = V$. $\square$

2.8 PARAFAC Random Model (M4′)

In this section, we are interested in the case that the common factors consist both of fixed and random effects. We assume that

$$\mathbf{f}^{(r,a,s)} = \begin{pmatrix} f_{1ras} \\ \vdots \\ f_{qras} \end{pmatrix} \text{ and } f_{kras} = w_{ak} v_{sk} + e_{kras} \text{ for } k = 1, \ldots, q \tag{2.71}$$

where the elements $\{e_{kras}, k = 1, \ldots, q\}$ are independent $N(0, \sigma_{e,kas}^2)$. That is,

$$\mathbf{f}^{(r,a,s)} \sim N(G^{as}, \Sigma_e^{(a,s)}) \tag{2.72}$$

where

$$G^{as} = \begin{pmatrix} g_{1as} \\ \vdots \\ g_{qas} \end{pmatrix} = \begin{pmatrix} w_{a1}v_{s1} \\ \vdots \\ w_{aq}v_{sq} \end{pmatrix} \text{ and } \Sigma_e^{(a,s)} = diag(\sigma^2_{e,1as}, \ldots, \sigma^2_{e,qas}). \quad (2.73)$$

Define the parameter space by

$$\begin{aligned} \Theta_{M4'} \;=\; & \{(\mu, \Lambda, W, V, \psi, \epsilon) : \Lambda \in O^+_{pq} \text{ and } \Lambda^t \mathbf{1} = 0, \\ & \mu = \{\mu_{as} : a = 1, \ldots, A, s = 1, \ldots, S\} \in \mathbf{R}^{AS}, \\ & \psi = (\sigma^2_{u,as}, (a,s) \in \{1, \ldots, A\} \times \{1, \ldots, S\}) \in \mathbf{R}^{AS}_+, \\ & \epsilon = (\sigma^2_{e,kas}, (k,a,s) \in \{1, \ldots, q\} \times \{1, \ldots, A\} \times \{1, \ldots, S\}) \in \mathbf{R}^{qAS}_+, \\ & W = (w_{ak}) \in \mathcal{M}^+_{Aq} \text{ with } \sum_{a=1}^{A} w^2_{ak} = 1, \\ & V = (v_{sk}) \in \mathcal{M}_{Sq} \text{ with } \sum_{s=1}^{S} v^2_{sk} = 1, \\ & rank(\Lambda) = rank(W) = rank(V) = q \leq A, S\}. \end{aligned} \quad (2.74)$$

The model in (2.50) under this parameter space is called PARAFAC random model, denoted by (M4′). We next show that our parameter θ is identifiable from the observed data in (M4′).

Lemma 2.23. *Model (M4′) is identifiable.*

Proof. In model (M4′), the expectation of $Y^{(r,a,s)}$ is

$$E(Y^{(r,a,s)}) = \mu_{as}\mathbf{1} + \Lambda \begin{pmatrix} w_{a1}v_{s1} \\ \vdots \\ w_{aq}v_{sq} \end{pmatrix} \quad (2.75)$$

and the covariance matrix of $Y^{(r,a,s)}$ is

$$\Sigma_Y^{(a,s)} = \Lambda \Sigma_e^{(a,s)} \Lambda^t + \sigma^2_{u,as} I \quad (2.76)$$

where $\Sigma_e^{(a,s)} = diag(\sigma^2_{e,1as}, \sigma^2_{e,2as}, \ldots, \sigma^2_{e,qas})$. Here $\sigma^2_{u,as}$ can be identified by the minimum eigenvalue of $\Sigma_Y^{(a,s)}$ for each a and s since $p > q$. Hence, $\Lambda\Sigma_e^{(a,s)}\Lambda^t$ is identified through equation (2.76). The parameter μ_{as} can be estimated consistently by

$$\widetilde{\mu}_{as} = \frac{1}{pR}\sum_{r=1}^{R}\sum_{i=1}^{p} y_{iras}$$

when $R \to \infty$. Then $\Lambda \begin{pmatrix} w_{a1}v_{s1} \\ \vdots \\ w_{aq}v_{sq} \end{pmatrix}$ is identified through the equation (2.75). That is, $\sum_{k=1}^{q} \lambda_{ik} w_{ak} v_{sk}$ is identified for all $a = 1, \ldots, A, s = 1, \ldots, S, i = 1, \ldots, p$.
Using Jennrich's Uniqueness Theorem (Theorem A.4), we can identify Λ, W and V as we did in Lemma 2.21.

Finally, we need to identify $\Sigma_e^{(a,s)}$. Since $\sigma^2_{u,as}$ is identified by the minimum eigenvalue of $\Sigma_Y^{(a,s)}$ for each a and s, the equation in (2.76) can be written as

$$\Sigma_Y^{(a,s)} - \sigma^2_{u,as} I = \Lambda\Sigma_e^{(a,s)}\Lambda^t. \tag{2.77}$$

Since Λ is identified and $\Lambda^t\Lambda = \text{diag}(b_1, \ldots, b_q)$, with $b_1 > b_2 > \ldots > b_q > 0$, the equation (2.77) can be written as

$$\begin{aligned} \Lambda^t(\Sigma_Y^{(a,s)} - \sigma^2_{u,as} I)\Lambda &= (\Lambda^t\Lambda)\Sigma_e^{(a,s)}(\Lambda^t\Lambda) \\ &= diag(b_1^2, \ldots, b_q^2)\ \Sigma_e^{(a,s)}. \end{aligned} \tag{2.78}$$

Since $\Lambda^t(\Sigma_Y^{(a,s)} - \sigma^2_{u,as} I)\Lambda$ and $diag(b_1^2, \ldots, b_q^2)$ are identified, so is $\Sigma_e^{(a,s)}$. □

2.9 Relationship among models in the model hierarchy

For application purposes, we have constructed a hierarchical family of factor models in this chapter. The model (M0) is the most general factor analysis model with the

form in (1.5). The model (M0a) is a special case of (M0) in which the mean level of the observations is proportional to the $\mathbf{1}$ vector. The model (M1) is a more restrictive model of (M0a) in which the mean level μ is assumed to be zero and Ψ is a scalar multiple of I_p. The model (M1R) is called the reduced model of (M1) which means that the covariance matrix of the error measurement under (M1R) has lower rank than the one under (M1). The hierarchy of models (M2), (M3), (M4), (M4a) and (M4′) are models for cross-classified data and can be applied to tongue image data. In model (M2), the common factors are considered as random effects. Model (M3) is similar to (M2) but the common factors are fixed for a specific experimental setting. The models (M4) and (M4a) have similar model assumptions. Both of them have fixed factors which can be decomposed in a specified form, so that both are nested in (M3). The difference between (M4) and (M4a) lies in the model assumption on the factor loading matrix. The loading matrix is assumed to have orthogonal columns in model (M4), but could have non-orthogonal columns in the PARAFAC model (M4a). Thus, (M4) is nested in (M4a). The model (M4′) differs from (M4) in having common factors which have both fixed and random effects.

Let "(Mb) $\subset$ (Ma)" denotes the model (Mb) is nested in (Ma). The relationships among the models are as follows: (M1R) $\subset$ (M1) $\subset$ (M0a) $\subset$ (M0). In cross-classified models, we have (M4) $\subset$ (M4a) $\subset$ (M3), and (M4′) $\subset$ (M2).

Chapter 3

ML Estimates for Factor Analysis Models

In this chapter, we find the maximum likelihood (ML) estimators for the parameters (Λ, ψ) under the models (M1) and (M1R) defined in Sections 2.2 and 2.3. In Section 3.2, we introduce the idea of profile likelihood and use it to find the maximum likelihood estimators for the parameters under (M1R). In Section 3.3, we discuss profile likelihood optimization in (M0). In Section 3.4, we find a necessary condition to check the local maximum likelihood estimate. In Section 3.5, we consider the score test for the problem $H_0 : \psi_{jj} = 0$ vs $H_A : \psi_{jj} > 0$. In Section 3.6, we discuss the likelihood ratio test for testing the adequacy of the PARAFAC model versus the general fixed effect factor model (M3).

3.1 Maximum likelihood estimate for (M1)

Consider the special case, denoted (M1), of the model (M0) when $\mu = 0$ and $\Psi = \sigma^2 I_p$:

$$Y = \Lambda f + U \qquad (M1)$$

where $f \sim N(0, I_q)$, $U \sim N(0, \sigma^2 I_p)$, and where $\Lambda \in O^+_{pq}$.

The parameter space Θ_{M1} for (M1) is defined in (2.39). The probability density

function of Y under (M1) is

$$f(y) = \frac{\exp\{-\frac{1}{2}y^t(\Lambda\Lambda^t + \sigma^2 I_p)^{-1}y\}}{(2\pi)^{p/2}|\Lambda\Lambda^t + \sigma^2 I_p|^{1/2}} \tag{3.1}$$

where $|A|$ means the determinant of the matrix A.

The maximum likelihood estimator for Λ and σ can be uniquely determined by the following lemma [24].

Lemma 3.1. *Consider the model*

$$Y = \mu + \Lambda f + U$$

where $f \sim N(0, I_q)$, $U \sim N(0, \sigma^2 I)$, and Λ has orthogonal columns. The maximum likelihood estimators for μ, Λ and σ^2 are given by

$$\widehat{\mu} = (1/n)\sum_{i=1}^{n} y_i \equiv \bar{y}, \tag{3.2}$$

$$\widehat{\Lambda} = Q_q(W_q - \sigma^2 I_q)^{\frac{1}{2}}R, \tag{3.3}$$

and

$$\widehat{\sigma}^2 = \frac{1}{p-q}\sum_{j=q+1}^{p} w_j. \tag{3.4}$$

where $Q_q \in \mathcal{M}_{pq}$ has the columns which are the principal eigenvectors of C_{yy}; $W_q = diag(w_1, \ldots, w_q)$ such that the entries w_j are the corresponding eigenvalues of C_{yy}; and R is an arbitrary $q \times q$ orthogonal matrix. Here C_{yy} is defined as

$$C_{yy} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{y})(Y_i - \bar{y})^t$$

Remark 3.2. *The equation in (3.4) has a clear interpretation as the variance "lost" in the projection, averaged over the lost dimensions.*

3.2 Maximum likelihood estimate for (M1R)

Consider the reduced form of the (M1) model:

$$Y = \Lambda f + U \qquad (M1R)$$

where $f \sim N(0, I_q)$, $U \sim N(0, \Psi)$, Λ is a p by q matrix and

$$\Psi = \begin{pmatrix} 0_r & O \\ O^t & \sigma^2 I_{p-r} \end{pmatrix} \tag{3.5}$$

where $r < q < p$, 0_r is a $r \times r$ matrix of zeroes, and O is a $r \times (p-r)$ matrix of zeroes. The parameter space Θ_{M1R} is defined in (2.42). To find the maximum likelihood estimators for the parameters Λ and σ^2 under the model (M1R), we start with the probability density function of Y.

3.2.1 Simplifying the probability density function for model (M1R)

The probability density function of Y under (M1R) is

$$f(y) = \frac{\exp\{-\frac{1}{2} y^t (\Lambda\Lambda^t + \sigma^2 I_p)^{-1} y\}}{(2\pi)^{p/2} |\Lambda\Lambda^t + \sigma^2 I_p|^{1/2}}. \tag{3.6}$$

For simplicity of notation, partition

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, \ \Lambda = \begin{pmatrix} \Lambda_1 \\ \Lambda_2 \end{pmatrix} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} \text{ and } U = \begin{pmatrix} 0_r \\ U_2 \end{pmatrix} \tag{3.7}$$

so that

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \Lambda_1 \\ \Lambda_2 \end{pmatrix} f + \begin{pmatrix} 0_r \\ U_2 \end{pmatrix} \tag{3.8}$$

Here $Y_1 \in \mathbf{R}^r, Y_2 \in \mathbf{R}^{p-r}$, Λ_1 is a $r \times q$ matrix and Λ_2 is a $(p-r) \times q$ matrix.

Projecting Λ_2 to the space generated by rows of Λ_1, we can write Λ_2 uniquely as

$$\Lambda_2 = B\Lambda_1 + \Lambda_2^* \tag{3.9}$$

where $B \in \mathcal{M}_{p-r,r}$ and the rows of Λ_1 are orthogonal to the rows of Λ_2^*, i.e., $\Lambda_2^*\Lambda_1^t = 0$. Then Y can be split into

$$Y_1 = \Lambda_1 f \tag{3.10}$$

$$Y_2 = \Lambda_2 f + U_2 = BY_1 + \Lambda_2^* f + U_2 \tag{3.11}$$

It follows that the conditional probability distribution of Y_2 given Y_1 is

$$Y_2|Y_1 \sim N_{p-r}(BY_1, \Lambda_2^*(\Lambda_2^*)^t + \sigma^2 I_{p-r}). \tag{3.12}$$

Therefore, the probability density function of Y under (M1R) model can be written as $f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1}(y_1) f_{Y_2|Y_1}(y_2|y_1)$ with

$$f_{Y_1}(y_1) = \frac{\exp\{-\frac{1}{2} y_1^t (\Lambda_1\Lambda_1^t)^{-1} y_1\}}{(2\pi)^{r/2}|\Lambda_1\Lambda_1^t|^{1/2}} \tag{3.13}$$

and

$$\begin{aligned} & f_{Y_2|Y_1}(y_2|y_1) \\ &= \frac{\exp\{-\frac{1}{2}(y_2 - By_1)^t(\Lambda_2^*(\Lambda_2^*)^t + \sigma^2 I_{p-r})^{-1}(y_2 - By_1)\}}{(2\pi)^{(p-r)/2}|\Lambda_2^*(\Lambda_2^*)^t + \sigma^2 I_{p-r}|^{1/2}}. \end{aligned} \tag{3.14}$$

So we have

$$\begin{aligned} & f_{Y_1,Y_2}(y_1, y_2) \\ &= \frac{\exp\{-\frac{1}{2} y_1^t (\Lambda_1\Lambda_1^t)^{-1} y_1\}}{(2\pi)^{r/2}|\Lambda_1\Lambda_1^t|^{1/2}} \\ &\times \frac{\exp\{-\frac{1}{2}(y_2 - By_1)^t(\Lambda_2^*(\Lambda_2^*)^t + \sigma^2 I_{p-r})^{-1}(y_2 - By_1)\}}{(2\pi)^{(p-r)/2}|\Lambda_2^*(\Lambda_2^*)^t + \sigma^2 I_{p-r}|^{1/2}} \end{aligned} \tag{3.15}$$

$$\begin{aligned} &= \frac{\exp\{-\frac{1}{2} y_1^t (\Lambda_1\Lambda_1^t)^{-1} y_1\}}{(2\pi)^{r/2}|\Lambda_1\Lambda_1^t|^{1/2}} \\ &\times \frac{\exp\{-\frac{1}{2}(y_2 - By_1)^t(\Lambda_{22}\Lambda_{22}^t + \sigma^2 I_{p-r})^{-1}(y_2 - By_1)\}}{(2\pi)^{(p-r)/2}|\Lambda_{22}\Lambda_{22}^t + \sigma^2 I_{p-r}|^{1/2}} \end{aligned} \tag{3.16}$$

To show that the equations (3.15) and (3.16) are equivalent, we need to show that $\Lambda_2^*(\Lambda_2^*)^t = \Lambda_{22}\Lambda_{22}^t$. First, we note that in the parameter space under (M1R), Λ_{11} is a

$r \times r$ matrix with rank r and Λ_{12} is a $r \times (q-r)$ zero matrix. Using $\Lambda_2^* \Lambda_1^t = 0$ and the equation (3.9), we have

$$\begin{aligned} 0 &= \Lambda_2^* \Lambda_1^t \\ &= (\Lambda_2 - B\Lambda_1)\Lambda_1^t \\ &= [(\Lambda_{21}|\Lambda_{22}) - B(\Lambda_{11}|0)](\Lambda_{11}|0)^t \\ &= \Lambda_{21}\Lambda_{11}^t - B(\Lambda_{11}\Lambda_{11}^t) \\ &= (\Lambda_{21} - B\Lambda_{11})\Lambda_{11}^t. \end{aligned} \tag{3.17}$$

Since Λ_{11} is a $r \times r$ matrix with full rank and therefore invertible, the last equation can be simplified as

$$\Lambda_{21} - B\Lambda_{11} = 0. \tag{3.18}$$

Now, let us show that $\Lambda_2^*(\Lambda_2^*)^t = \Lambda_{22}\Lambda_{22}^t$ using the equation (3.18):

$$\begin{aligned} \Lambda_2^*(\Lambda_2^*)^t &= [(\Lambda_{21}|\Lambda_{22}) - B(\Lambda_{11}|0)][(\Lambda_{21}|\Lambda_{22}) - B(\Lambda_{11}|0)]^t \\ &= [(\Lambda_{21} - B\Lambda_{11}|\Lambda_{22})][(\Lambda_{21} - B\Lambda_{11}|\Lambda_{22})]^t \\ &= [(0|\Lambda_{22})][(0|\Lambda_{22})]^t \\ &= \Lambda_{22}\Lambda_{22}^t \end{aligned} \tag{3.19}$$

Therefore, the equations (3.15) and (3.16) are equivalent.

3.2.2 Likelihood function and ML equation

Let $y_1, \ldots, y_n$ be a sample of n independent observations of Y. The joint probability density function $f(y_1, \theta) \cdots f(y_n, \theta)$, evaluated at $\mathbf{y} = (y_1, \ldots, y_n)$, can be considered as a function of θ, say $L(\theta)$. We call it the likelihood function. Let $\{y_i, i = 1, \ldots n\}$

be a sample. Partition

$$y_i = \begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} \tag{3.20}$$

where y_{1i} and y_{2i} are $q \times 1$ and $(p-q) \times 1$ column vectors, respectively. Under (M1R), the likelihood function for this sample is

$$L(\theta) = \prod_{i=1}^{n} f(y_i, \theta) = \prod_{i=1}^{n} f_{Y_{1i}Y_{2i}}(y_{1i}, y_{2i}; \theta) \tag{3.21}$$

The maximum likelihood estimates of θ are values $\hat{\theta}$ of θ which maximize the likelihood function $L(\theta)$, or equivalently, maximize the logarithm of the likelihood function, denoted by $l(\theta)$ with

$$\begin{aligned} l(\theta) &\equiv log(L(\theta)) = \sum_{i=1}^{n} log(f_{Y_{1i}Y_{2i}}(y_{1i}, y_{2i}; \theta)) \\ &= -\frac{np}{2} log(2\pi) - \frac{n}{2} log(|\Lambda_{11}\Lambda_{11}^t|) - \frac{1}{2}\sum_{i=1}^{n} y_{1i}^t (\Lambda_{11}\Lambda_{11}^t)^{-1} y_{1i} \\ &\quad -\frac{n}{2} log(|\Lambda_{22}\Lambda_{22}^t + \sigma^2 I_{p-r}|) \\ &\quad -\frac{1}{2}\sum_{i=1}^{n} (y_{2i} - By_{1i})^t (\Lambda_{22}\Lambda_{22}^t + \sigma^2 I_{p-r})^{-1} (y_{2i} - By_{1i}) \end{aligned} \tag{3.22}$$

To get maximum likelihood estimates for model (M1R), we first optimize the log-likelihood on B with the other parameters, Λ and σ^2 fixed. Once we get $\widehat{B} = \widehat{B}(\Lambda, \sigma^2)$, we plug it into the log-likelihood $l(\Lambda, \sigma^2, \widehat{B}(\Lambda, \sigma^2))$ and then optimize the likelihood on Λ, σ^2. This is the idea of the **profile log-likelihood**.

3.2.3 The profile log-likelihood

The idea of the profile log-likelihood is similar to the concentrated likelihood from Anderson (1984). The profile likelihood approach is as follows. Let Θ be the parameter space. We decompose the parameter space Θ into two subspaces Θ_1 and Θ_2

such that $\Theta = \Theta_1 \times \Theta_2$. Let $l(\theta)$ be the log-likelihood function on Θ. We optimize the log-likelihood on one subspace, say Θ_1, first with the other parameter component fixed. Let $\hat{\theta_1}(\theta_2)$ be the maximum likelihood estimate of θ_1 for fixed θ_2. **The profile log-likelihood** for θ_2 is defined as

$$l(\theta_2) = l(\hat{\theta_1}(\theta_2), \theta_2). \tag{3.23}$$

The maximum likelihood estimate $\hat{\theta_1}(\theta_2)$ is unique for many generalized linear models. Under certain conditions, the profile log-likelihood may be used just like any other log-likelihood as a function of the remaining parameter θ_2. Also, the maximized profile likelihood is equal to the overall maximized likelihood. That is,

$$\sup_{\theta \in \Theta} l(\theta) = \sup_{\theta_2 \in \Theta_2} \{ \max_{\theta_1 \in \Theta_1} l(\theta | \theta_2) \} = \sup_{\theta_2 \in \Theta_2} l(\hat{\theta_1}(\theta_2), \theta_2). \tag{3.24}$$

The following Lemma (Cheng [4]) shows that a sufficient condition for equation (3.24) to hold is that a unique maximum likelihood estimate $\hat{\theta_1}(\theta_2)$ exists when θ_2 is given.

Lemma 3.3. *Let $l(\theta)$ be a continuous log-likelihood function and $\theta = (\theta_1, \theta_2)$. If there exists a unique continuous function $\hat{\theta_1}(\theta_2)$ such that*

$$\max_{\theta_1 \in \Theta_1} l(\theta_1; \theta_2) = l(\hat{\theta_1}(\theta_2), \theta_2) \equiv l_p(\theta_2) \tag{3.25}$$

then we have

$$\sup_{\theta \in \Theta} l(\theta) = \sup_{\theta_2 \in \Theta_2} l_p(\theta_2). \tag{3.26}$$

Furthermore, if $l_p(\theta_2)$ is continuous, and Θ_2 is compact, then the right hand side of equation (3.26) is a maximum. That is,

$$\sup_{\theta_2 \in \Theta_2} l_p(\theta_2) = \max_{\theta_2 \in \Theta_2} l_p(\theta_2). \tag{3.27}$$

Now we will maximize the log-likelihood function in (3.22) for B with Λ and σ^2 fixed. This is equivalent to minimizing

$$\sum_{i=1}^{n}(y_{2i} - By_{1i})^t(\Lambda_{22}\Lambda_{22}^t + \sigma^2 I_{p-r})^{-1}(y_{2i} - By_{1i}). \tag{3.28}$$

As shown in Theorem 2.17, $\Lambda_{22}\Lambda_{22}^t + \sigma^2 I_{p-r} = \Sigma_{22.1}$ where $\Sigma_{22.1}$ is defined as $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. Expand the formula in (3.28) to get

$$\sum_{i=1}^{n}[y_{2i}^t(\Sigma_{22.1}^{-1})y_{2i} - 2y_{1i}^tB^t(\Sigma_{22.1}^{-1})y_{2i} + y_{1i}^tB^t(\Sigma_{22.1}^{-1})By_{1i}]. \tag{3.29}$$

The first summand is a constant independent of B, so minimizing (3.29) is equivalent to minimize the function $g(B)$ which is given by

$$g(B) = \sum_{i=1}^{n}[y_{1i}^tB^t(\Sigma_{22.1}^{-1})By_{1i} - 2y_{2i}^tB^t(\Sigma_{22.1}^{-1})y_{1i}]. \tag{3.30}$$

Next, find $\hat{B} = \arg\min_B g(B)$ by setting $\nabla_B g = 0$. Let B and K be two $(p-r)\times r$ matrices and δ be very small. Consider a small perturbation $B + \delta K$ of B. Then

$$< \nabla_B g, K >= \frac{d}{d\delta}g(B + \delta K)|_{\delta=0} \tag{3.31}$$

$$\begin{aligned} &= \textstyle\sum_{i=1}^{n}[\quad y_{1i}^tK^t(\Sigma_{22.1}^{-1})By_{1i} + y_{1i}^tB^t(\Sigma_{22.1}^{-1})Ky_{1i} \\ &\qquad -2y_{2i}^t(\Sigma_{22.1}^{-1})Ky_{1i}] + \delta\sum_{i=1}^{n}[y_{1i}^tK^t(\Sigma_{22.1}^{-1})Ky_{1i}]|_{\delta=0} \end{aligned} \tag{3.32}$$

yielding

$$\frac{1}{n}\sum_{i=1}^{n}y_{1i}^t[K^t(\Sigma_{22.1}^{-1})B + B^t(\Sigma_{22.1}^{-1})K]y_{1i} = \frac{2}{n}\sum_{i=1}^{n}y_{2i}^t(\Sigma_{22.1}^{-1})Ky_{1i}. \tag{3.33}$$

Let $C_{lm} = \frac{1}{n}\sum_{i=1}^{n}y_{li}y_{mi}^t$ for $l = 1, 2$ and $m = 1, 2$, where $\{y_i, i = 1, \ldots n\}$ is a sample and each $y_i^t = (y_{1i}, y_{2i})^t$ is partitioned as in (3.20) with y_{1i} and y_{2i} are $q\times 1$ and $(p-q)\times 1$ column vectors, respectively. Note that both C_{11} and $\Sigma_{22.1}^{-1}$ are symmetric.

By definition of trace in Definition 1.10, the left hand side of (3.33) can be written as

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} y_{1i}^t[K^t(\Sigma_{22.1}^{-1})B + B^t(\Sigma_{22.1}^{-1})K]y_{1i} \\
&= \frac{1}{n}\sum_{i=1}^{n} tr\{[K^t(\Sigma_{22.1}^{-1})B + B^t(\Sigma_{22.1}^{-1})K]y_{1i}y_{1i}^t\} \\
&= tr\{[K^t(\Sigma_{22.1}^{-1})B + B^t(\Sigma_{22.1}^{-1})K]C_{11}\} \\
&= 2\, tr\{[K^t(\Sigma_{22.1}^{-1})B]C_{11}\}. \qquad (3.34)
\end{aligned}$$

Similarly, the right hand side of (3.33) can be written as

$$\begin{aligned}
\frac{2}{n}\sum_{i=1}^{n} y_{2i}^t(\Sigma_{22.1}^{-1})Ky_{1i} &= \frac{2}{n}\sum_{i=1}^{n} tr\{(\Sigma_{22.1}^{-1})Ky_{1i}y_{2i}^t\} \\
&= 2\, tr\{(\Sigma_{22.1}^{-1})KC_{12}\} \qquad (3.35)
\end{aligned}$$

Since K is an arbitrary matrix, (3.34) and (3.35) gives us

$$(\Sigma_{22.1}^{-1})BC_{11} = (\Sigma_{22.1}^{-1})C_{12}^t \qquad (3.36)$$

Since both C_{11} and $\Sigma_{22.1}^{-1}$ are invertible, the solution of equation (3.36) is

$$\widehat{B} = C_{12}^t C_{11}^{-1} \qquad (3.37)$$

which is totally independent of Λ and Ψ.

Next, plugging $\hat{B}$ into the log-likelihood in (3.22), we can write $l_p(\theta_2) = l(\hat{\theta}_1(\theta_2), \theta_2) = constant + \log L_1 + \log L_2$ with

$$\log L_1 = -\frac{n}{2}\log(|\Lambda_{11}\Lambda_{11}^t|) - \frac{1}{2}\sum_{i=1}^{n} y_{1i}^t(\Lambda_{11}\Lambda_{11}^t)^{-1}y_{1i} \qquad (3.38)$$

and

$$\begin{aligned}
\log L_2 = &-\frac{n}{2}\log(|\Lambda_{22}\Lambda_{22}^t + \sigma^2 I_{p-r}|) \\
&-\frac{1}{2}\sum_{i=1}^{n}(y_{2i} - \hat{B}y_{1i})^t(\Lambda_{22}\Lambda_{22}^t + \sigma^2 I_{p-r})^{-1}(y_{2i} - \hat{B}y_{1i}). \qquad (3.39)
\end{aligned}$$

Now let us simplify $\log L_2$ after substitution of $\Sigma_{22.1} = \Lambda_{22}\Lambda_{22}^t + \sigma^2 I_{p-r}$ into expression (3.39) and using the definition of trace, we have

$$\begin{aligned}
\log L_2 &= -\frac{n}{2}\{\log(|\Sigma_{22.1}|) + \frac{1}{n}\sum_{i=1}^{n}(y_{2i} - \hat{B}y_{1i})^t(\Sigma_{22.1}^{-1})(y_{2i} - \hat{B}y_{1i})\} \\
&= -\frac{n}{2}\{\log(|\Sigma_{22.1}|) + tr[(\Sigma_{22.1}^{-1}) \cdot \frac{1}{n}\sum_{i=1}^{n}(y_{2i} - \hat{B}y_{1i})(y_{2i} - \hat{B}y_{1i})^t]\} \\
&= -\frac{n}{2}\{\log(|\Sigma_{22.1}|) + tr[(\Sigma_{22.1}^{-1})C_{xx}]\} \qquad (3.40)
\end{aligned}$$

where

$$C_{xx} = \frac{1}{n}\sum_{i=1}^{n}(y_{2i} - \hat{B}y_{1i})(y_{2i} - \hat{B}y_{1i})^t.$$

Note that $\log L_1$ is a function only depending on data and the parameters Λ_{11}, while $\log L_2$ is a function of Λ_{22} and σ^2 only. Thus, optimizing $l_p(\theta_2)$ is equivalent to optimizing $\log L_1$ for Λ_{11} and to optimizing $\log L_2$ for (Λ_{22}, σ^2).

In $\log L_1$, the log-likelihood is maximized, according to Lemma 3.1 [24], when

$$\widehat{\Lambda}_{11} = U_r D_r^{1/2}$$

where we decompose $\Lambda_{11}\Lambda_{11}^t = U_r D_r U_r^t$, D_r is a $r \times r$ diagonal matrix of eigenvalues $d_1, d_2, \ldots, d_r$ with $d_1 > d_2 > \ldots > d_r$, and the columns of U_r are standardized eigenvectors of $\Lambda_{11}\Lambda_{11}^t$ with the first nonzero element in each column being positive.

In $\log L_2$, the maximum-likelihood estimator for Λ_{22} and σ^2 can be uniquely determined by Lemma 3.1 [24] with

$$\widehat{\Lambda}_{22} = U_{q-r}(W_{q-r} - \sigma^2 I_{q-r})^{\frac{1}{2}} R$$

$$\widehat{\sigma}^2 = \frac{1}{p-q}\sum_{j=q+1-r}^{p-r} w_j$$

where $U_{q-r} \in O^{+}_{p-r,q-r}$, $W_{q-r} = diag(w_1, \ldots, w_{q-r})$ and R is an arbitrary $(q-r) \times (q-r)$ permutation matrix. The column vectors of U_{q-r} are the principal eigenvectors of C_{xx} corresponding to the eigenvalues $w_1, \ldots, w_{q-r}$.

3.3 Profile likelihood optimization in (M0)

Consider the (M0) model as defined in Chapter 2, Section 1:

$$Y^{(r)} = \underline{\mu} + \Lambda \mathbf{f}^{(r)} + U^{(r)} \qquad (M0)$$

where $\mathbf{f}^{(r)} \sim N(0, I_q)$, $U^{(r)} \sim N(0, \Psi)$, $\mathbf{f}^{(r)}$ and $U^{(r)}$ are independent. The matrix $\Psi = diag(\psi)$, is a $p \times p$ diagonal matrix with the vector $\psi \equiv (\psi_1, \ldots, \psi_p) \in \mathbf{R}^p$ on the diagonal. Therefore, the general factor model can be expressed

$$Y^{(r)} \sim N(\underline{\mu},\ \Lambda\, \Lambda^t +\ \Psi).$$

The observed data consists of $\{Y^{(r)}; r = 1, \ldots, R\}$. The log-likelihood is

$$\begin{aligned} l_R(\mu, \Lambda, \psi) &= -\frac{Rp}{2} \log 2\pi - \frac{R}{2} \log |\Sigma_y| \\ &- \frac{1}{2} \sum_{r=1}^{R} (y^{(r)} - \mu)^t\ \Sigma_y^{-1}\ (y^{(r)} - \mu). \end{aligned} \qquad (3.41)$$

Maximizing $l_R(\mu, \Lambda, \psi)$ with respect to μ yields

$$\hat{\mu} = (1/R) \sum_{r=1}^{R} y^{(r)} \equiv \bar{y}. \qquad (3.42)$$

Substituting $\hat{\mu}$ into (3.41) yields the profile likelihood

$$l_{R,prof}(\mu, \Lambda, \psi) = -\frac{Rp}{2} \log 2\pi - \frac{R}{2} (\log |\Sigma_y| + tr(\Sigma_y^{-1} C_{yy})) \qquad (3.43)$$

with

$$C_{yy} = (1/R) \sum_{r=1}^{R} (y^{(r)} - \bar{y})(y^{(r)} - \bar{y})^t. \qquad (3.44)$$

Clearly maximizing (3.43) is equivalent to minimizing $\log|\Sigma_y| + tr(\Sigma_y^{-1}C_{yy})$ with respect to Λ and ψ.

Now we consider $\Psi = \sigma^2\Gamma$ with Γ known and diagonal. Since Γ is known, we can multiply $Y^{(r)}$ by $\Gamma^{-1/2}$ from the left and transform the (M0) model to

$$Y^{(r)*} = \underline{\mu}^* + \Lambda^*\mathbf{f}^{(r)} + U^{(r)*} \tag{3.45}$$

where $Y^{(r)*} = \Gamma^{-1/2}\ Y^{(r)}$, $\underline{\mu}^* = \Gamma^{-1/2}\ \underline{\mu}$, $\Lambda^* = \Gamma^{-1/2}\ \Lambda$ and $U^{(r)*} = \Gamma^{-1/2}\ U^{(r)}$. Then $U^{(r)*} \sim N_p(0, \sigma^2 I_p)$ just as in (M1). The covariance matrix of $Y^{(r)*}$ is

$$\Sigma_Y^* = \Lambda^*\Lambda^{*\,t} + \sigma^2 I_p. \tag{3.46}$$

The log-likelihood is

$$\begin{aligned} l_R^*(\mu^*, \Lambda^*, \Psi^*) &= -\frac{Rp}{2}\log 2\pi - \frac{R}{2}\log|\Sigma_y^*| \\ &- \frac{1}{2}\sum_{r=1}^{R}(y^{(r)*} - \mu)^t\ (\Sigma_y^*)^{-1}\ (y^{(r)*} - \mu). \end{aligned} \tag{3.47}$$

By Lemma 3.1, we obtain, for the model (M0) on $\{Y^{(r)*}\}$ with Γ known,

$$\widehat{\mu}^* = (1/R)\sum_{r=1}^{R} y^{(r)*}, \tag{3.48}$$

$$\widehat{\Lambda}^* = Q_q(W_q - \sigma^2 I_q)^{\frac{1}{2}}T, \tag{3.49}$$

and

$$\widehat{\sigma}^2 = \frac{1}{p-q}\sum_{j=q+1}^{p} w_j \tag{3.50}$$

where $Q_q \in \mathcal{M}_{pq}$ has the columns which are the principal eigenvectors of C_{yy}^*; $W_q = diag(w_1, \ldots, w_q)$ such that the entries w_j are the corresponding eigenvalues of C_{yy}^*; and T is an arbitrary $q \times q$ orthogonal matrix. Here C_{yy}^* is a function of Γ defined as

$$C_{yy}^* \equiv C_{yy}^*(\Gamma) = \Gamma^{-1/2}\ C_{yy}\ \Gamma^{-1/2}$$

with

$$C_{yy} \equiv (1/R)\sum_{r=1}^{R}(y^{(r)} - \bar{y})(y^{(r)} - \bar{y})^t.$$

Therefore, the estimators of μ^* in (3.48), Λ^* in (3.49), and σ^2 in (3.50) are all functions of Γ.

Substituting $\hat{\mu}^*$ into (3.47) yields the profile likelihood

$$\begin{aligned} l^*_{prof}(\Gamma) &\equiv l_{R,prof}(\Lambda^*(\Gamma), \psi^*(\Gamma)) \\ &= -\frac{Rp}{2}\log 2\pi - \frac{R}{2}(\log|\Sigma^*_y| + tr((\Sigma^*_y)^{-1}C^*_{yy})). \end{aligned} \tag{3.51}$$

The idea of profile likelihood maximization in the general factor analysis model in terms of Γ for the vector ψ is discussed by Magnus and Neudecker [14].

Maximizing (3.51) is equivalent to minimizing

$$g_{prof}(\Lambda^*(\Gamma), \sigma^2(\Gamma)) \equiv (\log|\Sigma^*_y| + tr((\Sigma^*_y)^{-1}C^*_{yy}) \tag{3.52}$$

with respect to Γ. Substituting (3.49) and (3.50) in (3.52), and using the Newton-Raphson method to minimize $g_{prof}(\widehat{\Lambda}^*, \widehat{\sigma}^2)$ iteratively over Γ so that the current value of Γ is used as above, we can finally get $\widehat{\Gamma}$. The Newton-Raphson method will be introduced in the next chapter.

3.3.1 Why it is good to use the profile likelihood?

The profile likelihood method allowed us to reduce the parameter dimension by working on the two separate subspaces of parameters when we deal with high dimensional problems. For example, in (M0), we optimize log-likelihood $l(\theta)$ on Λ first with the other parameter component ψ fixed. Let $\hat{\Lambda}(\psi)$ be the maximum likelihood estimate of Λ for fixed ψ. Then the profile log-likelihood for ψ is defined as $l(\psi) = l(\hat{\Lambda}(\psi), \psi)$ which is only dependent on ψ. Since $\Lambda \in O^+_{pq}$ and ψ is a vector with p components,

the number of free parameters in Λ and in ψ are $p \cdot q - \frac{q(q-1)}{2}$ and p, respectively. Thus, the parameter dimension is reduced from $p \cdot q - \frac{q(q-1)}{2} + p$ to p if the profile likelihood method is used.

In model (M1R), we will be able to find the maximum likelihood estimator through the profile likelihood method. However, as in most multivariate analysis problems, the maximum of the profile log-likelihood does not have a closed-form analytic form for θ_2. That is, the profile log-likelihood equation can not be solved directly. We will use a numerical procedure to compute the maximum likelihood estimates iteratively. There are various iterative procedures such as the Newton-Raphson method, the EM (expectation-maximization) algorithm, and the steepest descent method. We will discuss these in the next chapter.

3.4 Condition to check the local maximum likelihood estimate

In Lemma 2.15, we show that if $p > 2q$ and the model is non-identifiable, then there exists (Λ^*, ψ^*) in the boundary of the parameter space satisfying the condition (2.2). We found the same situation in our simulated data, that is, the maximum likelihood estimator (Λ^*, ψ^*) may have $\psi_j^* = 0$ for some j, where $\psi^* = (\psi_1^*, \ldots, \psi_p^*)$. Therefore, we want to ask whether the estimator that we found in the boundary of the parameter space achieves the local maximum of the likelihood function. To verify this, we need the condition to check whether the log-likelihood is decreasing when we approach in a certain direction, for example, approach from the interior of the parameter space.

Suppose that $\theta^* \equiv (\Lambda^*, \psi^*)$ in the boundary of the parameter space Θ and $\psi^* = (\psi_1^*, \ldots, \psi_p^*)$ with $\psi_1^* = \ldots = \psi_r^* = 0$. Denote the log-likelihood function as $l(\theta)$ and

define

$$\nabla_{\psi^r} l(\widehat{\theta^r}) \equiv (\frac{\partial l(\theta)}{\partial \psi_1}, \ldots, \frac{\partial l(\theta)}{\partial \psi_r})^t \mid_{\theta=\widehat{\theta^r}} \tag{3.53}$$

where $\widehat{\theta^r}$ is the maximum likelihood estimator in the restricted model (M1R) defined in (2.42) with $\Psi \equiv diag(\psi)$ and $\psi = diag(0, \ldots, 0, \psi_{r+1}, \ldots, \psi_p)$. To verify whether the log-likelihood is decreasing when we approach in a certain direction is equivalent to checking the condition

$$\nabla_{\psi^r} l(\widehat{\theta^r}) \cdot e_j < 0 \text{ for } j = 1, \ldots, r \tag{3.54}$$

where $\{e_1, \ldots, e_p\}$ is the canonical basis of $\mathbf{R}^p$.

In (M0), the log-likelihood function is

$$l(\theta) = -\frac{pR}{2} \log 2\pi - \frac{R}{2} \log |\Sigma_y| - \frac{R}{2} tr(C_{yy}\ \Sigma_y^{-1}). \tag{3.55}$$

Here $C_{yy} \equiv \frac{1}{R} \sum_{r=1}^{R} (y^{(r)} - \bar{y})(y^{(r)} - \bar{y})^t$. Then the partial derivative of (3.55) with regard to ψ_j is (Anderson 1984)

$$\frac{\partial l(\theta)}{\partial \psi_j} = -\frac{R}{2} [\sigma^{jj} - \sum_{k=1}^{p} \sum_{m=1}^{p} c_{km} \sigma^{mj} \sigma^{jk}],\ j = 1, \ldots, r. \tag{3.56}$$

where $\Sigma_y^{-1} = (\sigma^{ij})$ and $C_{yy} = (c_{ij})$. Thus, the condition in (3.54) is equivalent to

$$(\widehat{\Sigma_y}^{-1} - \widehat{\Sigma_y}^{-1} C_{yy} \widehat{\Sigma_y}^{-1})_{jj} > 0 \ \text{ for } j = 1, 2, \ldots, r \tag{3.57}$$

where $\widehat{\Sigma_y} = \widehat{\Lambda}_r\ \widehat{\Lambda}_r^t + \widehat{\Psi}_r$ and $(\widehat{\Lambda}_r, \widehat{\psi}_r) = \widehat{\theta}_r$.

3.5 Score Test for $H_0 : \psi_j = 0$ versus $H_A : \psi_j > 0$

The efficient score test (Cox and Hinkley [5], p. 324), also called *Locally Most Powerful (LMP) test* (Lehmann [12]), is a widely applicable method of test construction that

provides a convenient alternative to the likelihood ratio test. Based on the likelihood, score tests are asymptotically equivalent to likelihood ratio tests but do not require calculation of maximum likelihood estimates from the full, unconstrained model. This property makes the score test an ideal alternative to the likelihood ratio tests when maximum likelihood estimates from the full model are difficult to obtain. Especially when parameter is not in the interior of the parameter but on the boundary, LRT is non-standard and has different distribution (Self and Liang [20]).

This section summarizes briefly the theory of likelihood score tests. Further background on score test can be found in Cox and Hinkley [5]. Let $l(y;\theta_1,\theta_2)$ be a log-likelihood function depending on a response vector y and parameter vectors θ_1 and θ_2. We wish to test the composite hypothesis $H_0 : \theta_1 = \theta_{10}$ against the general alternative $H_A : \theta_1$ is unrestricted. The components of θ_2 are so-called nuisance parameters because they are not of interest in the test but values must be estimated for them in order for a test statistic to be computed. The likelihood score vectors for θ_1 and θ_2 are the partial derivatives

$$S_1 = \frac{\partial l}{\partial \theta_1} \text{ and } S_2 = \frac{\partial l}{\partial \theta_2} \tag{3.58}$$

respectively. The observed information matrix for the parameters is $-H(\theta)$ with

$$H(\theta) = \frac{\partial^2 l}{\partial \theta \partial \theta^t} = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}. \tag{3.59}$$

The Fisher information matrix is $\mathcal{I} = E(-H)$, which is partitioned into the same blocks as H, yielding

$$\mathcal{I} = \begin{pmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{pmatrix}. \tag{3.60}$$

The score test statistic is based on the fact that the score vector $S = (S_1, S_2)$ has mean zero and covariance matrix $\mathcal{I}$. If the nuisance vector θ_2 is known, then the score

test statistic of H_0 is

$$Z = \mathcal{I}_{11}^{-1/2}\, S_1, \tag{3.61}$$

where $\mathcal{I}_{11}^{1/2}$ stands for any factor such that $\mathcal{I}_{11}^{1/2}\ (\mathcal{I}_{11}^{1/2})^t = \mathcal{I}_{11}$, or equivalently

$$T = Z^t\ Z = S_1^t\ \mathcal{I}_{11}^{-1}\ S_1 \tag{3.62}$$

with S_1 and $\mathcal{I}_{11}$ evaluated at $\theta_1 = \theta_{10}$. The score vector S is a sum of independent terms corresponding to individual observations and so is asymptotically normal under standard regularity conditions. It follows that Z is asymptotically a standard normal p_1-vector under the null hypothesis H_0 and that T is asymptotically chi-square distributed on p_1 degrees of freedom, where p_1 is the dimension of θ_1.

If the nuisance parameters are not known, then the score test substitutes for them their so-called 'restricted' maximum likelihood estimators $\hat{\theta}_2^{(r)}$ under the null hypothesis. Setting $\theta_2 = \hat{\theta}_2^{(r)}$ is equivalent to setting $S_2 = 0$, so we need the asymptotic distribution of S_1 conditional on $S_2 = 0$, which is normal with mean zero and covariance matrix

$$\mathcal{I}_{11.2} = \mathcal{I}_{11} - \mathcal{I}_{12}\, \mathcal{I}_{22}^{-1}\, \mathcal{I}_{21}. \tag{3.63}$$

The score test statistic becomes

$$T = S_1^t\, \mathcal{I}_{11.2}^{-1}\, S_1 \tag{3.64}$$

with S_1 and $\mathcal{I}_{11.2}$ evaluated at $\theta_2 = \hat{\theta}_2$ and $\theta_1 = \theta_{10}$. Under the null hypothesis H_0, T is asymptotically chi-square distributed on p_1 degrees of freedom, where p_1 is the dimension of θ_1.

Neyman [16] and Neyman and Scott [17] show that the asymptotic distribution and efficiency of the score statistic T is unchanged if an estimator other than the maximum

likelihood estimator is used for the nuisance parameters, provided that the estimator is consistent with convergence rate at least $O(n^{-1/3})$, where n is the number of observations.

If $\mathcal{I}_{21} = 0$, then θ_1 and θ_2 are said to be orthogonal. In this case, S_1 and S_2 are asymptotically independent and $\mathcal{I}_{11.2} = \mathcal{I}_{11}$, meaning that the information matrix $\mathcal{I}_{11}$ does not need to be adjusted for estimation of θ_2.

The p-value is defined as

$$Pr(T_n \geq t)|_{t=T_n^*}$$

where T_n is defined in (3.64) and $T_n^* = S_{1n}^t(0, \hat{\theta}_{20})[\mathcal{I}_{11.2}(0, \hat{\theta}_{20})]^{-1} S_1(0, \hat{\theta}_{20})$. Let α be the level of significance. If the p-value $\leq \alpha$, then we reject H_0.

3.5.1 Score Test for $H_0 : \psi_j = 0$ vs $H_A : \psi_j > 0$ under (M0) with $\mu = 0$

In this section, we will find the score test statistic and Fisher information under (M0) with $\mu = 0$. To test the composite hypothesis $H_0 : \psi_j = 0$ against $H_A : \psi_j > 0$, we start by calculating the score statistic. The parameter is $\theta \equiv (\Lambda, \psi)$, where $\Lambda = (\lambda_{ij})$ is the loading matrix and $\psi = (\psi_1, \ldots, \psi_p)$ is a vector such that the covariance of the error is $\Psi \equiv diag(\psi)$. Let $\theta_1 = \psi_j$ and θ_2 be the vector with components $\{\lambda_{ij}, \psi_k : i = 1, \ldots, p\,; j = 1, \ldots q\,; k = 1, \ldots, j-1, j+1, \ldots p\}$. The log-likelihood function of θ is

$$l(\theta) = -\frac{Rp}{2}\log(2\pi) - \frac{R}{2}\log|\Sigma| - \frac{R}{2}tr[C_{yy}(\Sigma)^{-1}]$$

where $\Sigma = \Lambda\Lambda^t + \Psi$. The derivative of Σ with respect to λ_{lm} is

$$\frac{\partial \sigma_{ij}}{\partial \lambda_{lm}} = \begin{cases} 2\lambda_{lm} & \text{if } i = j = l \\ \lambda_{jm} & \text{if } i = l,\ i \neq j \\ \lambda_{im} & \text{if } j = l,\ i \neq j \\ 0 & \text{otherwise.} \end{cases} \tag{3.65}$$

The derivative of Σ with respect to ψ_i is

$$\frac{\partial \sigma_{ij}}{\partial \psi_i} = diag(e_i). \tag{3.66}$$

From $\Sigma\Sigma^{-1} = I$, we obtain for any parameter θ

$$\frac{\partial \Sigma^{-1}}{\partial \theta} = -\Sigma^{-1}\ \frac{\partial \Sigma}{\partial \theta}\ \Sigma^{-1}. \tag{3.67}$$

Let $\Sigma^{-1} = (\sigma^{ij})$ and use (3.67) to get the partial derivative of Σ^{-1} with regard to ψ_j

$$\frac{\partial \sigma^{km}}{\partial \psi_j} = -\sigma^{kj}\ \sigma^{jm}. \tag{3.68}$$

Similarly, the partial derivative of Σ^{-1} with respect to λ_{mk} is

$$\frac{\partial \sigma^{km}}{\partial \lambda_{ij}} = -[(\Sigma^{-1})_{im}\ (\Sigma^{-1}\Lambda)_{kj}\ + (\Sigma^{-1})_{ik}\ (\Sigma^{-1}\Lambda)_{mj}]. \tag{3.69}$$

By A.6, we have

$$\frac{\partial \log|\Sigma|}{\partial \Sigma} = 2\Sigma^{-1} - D_{\Sigma^{-1}}$$

where $D_{\Sigma^{-1}}$ is a diagonal matrix with i'th diagonal element equal to that of Σ^{-1}. Thus, the derivative of $\log|\Sigma|$ with respect to ψ_i is

$$\begin{aligned} \frac{\partial \log|\Sigma|}{\partial \psi_i} &= \sum_{j,k} \frac{\partial \log|\Sigma|}{\partial \sigma_{jk}} \cdot \frac{\partial \sigma_{jk}}{\partial \psi_i} \\ &= tr[\ \frac{\partial \log|\Sigma|}{\partial \Sigma}\ \frac{\partial \Sigma}{\partial \psi_i}\ ] \\ &= tr[(2\Sigma^{-1} - D_{\Sigma^{-1}})(diag(e_i))] \\ &= \sum_{j,k} (2\sigma^{jk} - \sigma^{jk}\ \delta_{jk}) \cdot (diag(e_i))_{jk} \\ &= 2\sigma^{ii} - \sigma^{ii} \\ &= \sigma^{ii}. \end{aligned} \tag{3.70}$$

Using (3.65) and the same idea as in the last equation, the derivative of $\log|\Sigma|$ with respect to λ_{lm} is

$$\begin{aligned}\frac{\partial \log|\Sigma|}{\partial \lambda_{lm}} &= tr[\ (2\Sigma^{-1} - D_{\Sigma^{-1}})\ (\frac{\partial \Sigma}{\partial \lambda_{lm}})\ ] \\ &= \sum_{j=1}^{p} \sigma^{lj}\ \lambda_{jm}. \end{aligned} \tag{3.71}$$

Then the partial derivative of log-likelihood $l(\theta)$ with respect to ψ_j is

$$\frac{\partial l(\theta)}{\partial \psi_j} = -\frac{R}{2} \cdot (\Sigma^{-1} - \Sigma^{-1} C_{yy} \Sigma^{-1})_{jj},\ 1 \le j \le q \tag{3.72}$$

and the partial derivative of log-likelihood $l(\theta)$ with respect to λ_{mk} is

$$\frac{\partial l(\theta)}{\partial \lambda_{mk}} = -R \cdot (\Sigma^{-1}\ \Lambda - \Sigma^{-1} C_{yy} \Sigma^{-1}\ \Lambda)_{mk},\ 1 \le m \le p, 1 \le k \le q. \tag{3.73}$$

The score statistic

$$S_1(\theta_1, \theta_2) = \frac{\partial l(\theta)}{\partial \psi_j} = -\frac{R}{2} \cdot (\Sigma^{-1} - \Sigma^{-1} C_{yy} \Sigma^{-1})_{jj} \tag{3.74}$$

and

$$S_2(\theta_1, \theta_2) = \frac{\partial l(\theta)}{\partial \theta_2} \tag{3.75}$$

can be obtained through (3.72) and (3.73).

Similarly, we can calculate the second partial derivative of log-likelihood $l(\theta)$ to get

$$\frac{\partial^2 l(\theta)}{\partial \psi_j^2} = \frac{R}{2}[(\Sigma^{-1})_{jj}^2 - (\Sigma^{-1} C_{yy} \Sigma^{-1})_{jj}\ (\Sigma^{-1})_{jj}], \tag{3.76}$$

for $j = 1 \ldots p$,

$$\begin{aligned}\frac{\partial^2 l(\theta)}{\partial \lambda_{mk} \partial \psi_j} &= R[(\Sigma^{-1})_{mj}\ (\Sigma^{-1}\Lambda)_{jk} - (\Sigma^{-1})_{mj}\ (\Sigma^{-1} C_{yy} \Sigma^{-1} \Lambda)_{jk} \\ &\quad -(\Sigma^{-1} C_{yy} \Sigma^{-1})_{mj}\ (\Sigma^{-1}\Lambda)_{jk}], \end{aligned} \tag{3.77}$$

for $m = 1 \ldots p$, $k = 1 \ldots q$, and $j = 1 \ldots p$, and

$$\begin{aligned}
\frac{\partial^2 l(\theta)}{\partial \lambda_{ij} \partial \lambda_{mk}} &= R[(\Sigma^{-1}\Lambda)_{mj}\ (\Sigma^{-1}\Lambda)_{ik} + (\Sigma^{-1})_{im}\ (\Lambda^t\ \Sigma^{-1}\Lambda)_{kj} \\
&- (\Sigma^{-1}\Lambda)_{mj}\ (\Sigma^{-1} C_{yy} \Sigma^{-1}\Lambda)_{ik} - (\Sigma^{-1})_{mi}\ \delta_{k=j} \\
&- (\Sigma^{-1})_{im}\ (\Lambda^t\ \Sigma^{-1} C_{yy} \Sigma^{-1}\Lambda)_{jk} - (\Sigma^{-1} C_{yy} \Sigma^{-1}\Lambda)_{mj}\ (\Sigma^{-1}\Lambda)_{ik} \\
&- (\Sigma^{-1} C_{yy} \Sigma^{-1})_{mi}\ (\Lambda^t \Sigma^{-1}\Lambda)_{kj} \\
&+ (\Sigma^{-1} C_{yy} \Sigma^{-1})_{mi}\ \delta_{\{k=j\}}\ ] \qquad (3.78)
\end{aligned}$$

for $i = 1 \ldots p$, $m = 1 \ldots p$, $k = 1 \ldots q$, and $j = 1 \ldots q$.

Taking the expectation on (3.76), (3.77) and (3.78), we get

$$E[\frac{\partial^2 l(\theta)}{\partial \psi_j^2}] = -\frac{R}{2}[(\Sigma^{-1})_{jj}]^2 \qquad (3.79)$$

for $j = 1 \ldots p$,

$$E[\frac{\partial^2 l(\theta)}{\partial \lambda_{mk} \partial \psi_j}] = -R[(\Sigma^{-1})_{mj}\ (\Sigma^{-1}\Lambda)_{jk}] \qquad (3.80)$$

for $m = 1 \ldots p$, $k = 1 \ldots q$, and $j = 1 \ldots p$.

$$E[\frac{\partial^2 l(\theta)}{\partial \lambda_{ij} \partial \lambda_{mk}}] = -R[(\Sigma^{-1}\Lambda)_{mj}\ (\Sigma^{-1}\Lambda)_{ik} + (\Sigma^{-1})_{im}\ (\Lambda^t\ \Sigma^{-1}\Lambda)_{kj}] \qquad (3.81)$$

for $1 \leq i, m \leq p$ and $1 \leq j, k \leq q$. Therefore, the Fisher information matrix $\mathcal{I} = E[-H]$ can be estimated through (3.79), (3.80) and (3.81), where

$$H(\theta) = \frac{\partial^2 l}{\partial \theta \partial \theta^t} = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}. \qquad (3.82)$$

The score test statistic is $T = S_1^t\ \mathcal{I}_{11.2}^{-1}\ S_1$ with S_1 and $\mathcal{I}_{11.2}$ evaluated at $\theta_2 = \hat{\theta}_2$ and $\theta_1 (= \psi_j) = 0$.

3.6 Test of Fit for the PARAFAC Model

Slud et al. [22] found in a specific data example closely related to that explored in Chapter 5 below that the restricted PARAFAC model (M4) did less well, the more highly cross-classified the data were. Due to the highly constrained form and inadequacy of PARAFAC, a more general model such as 3-mode factor analysis model (T3) is needed for representing cross-classified data. The 3-mode factor analysis model, also called Tucker 3 model (T3), was introduced by Tucker [25] and can be written as

$$y_{ias}^{(r)} = \sum_{l=1}^{L}\sum_{m=1}^{M}\sum_{n=1}^{N} \lambda_{il} w_{am} v_{sn} g_{lmn} + u_{ias}^{(r)} \tag{3.83}$$

where g_{lmn} is the element of a three-mode matrix G which is called the $core\ matrix$. The PARAFAC model (M4a) is a special case of the T3 model when

$$g_{lmn} = \begin{cases} 1 & \text{if } l = m = n \\ 0 & \text{otherwise.} \end{cases}$$

Zheng et al. [27] mentioned in other tongue and speech related data sets that the T3 model fits better than PARAFAC (M4a), but it tends to use excess parameters. Thus, the well-defined model hierarchy we constructed may help to rationalize the choice of models. Model (M3) in the model hierarchy we constructed is a more general model than (T3). Thus, we have (M4) $\subset$ (M4a) $\subset$ (T3) $\subset$ (M3). In this section, we will construct a likelihood ratio test of whether the more general models (M3) or (M4a) better represent a statistical data set.

3.6.1 The Likelihood Ratio Test

We derive the likelihood ratio test (LRT) that the model fits. For a specified q, the covariance matrix can be written as $\Sigma_Y = \Lambda\Lambda^t + \Psi$ for some $p \times q$ matrix Λ and

some $p \times p$ diagonal positive definite matrix Ψ. The general strategy of the LRT is to maximize the likelihood under the null hypothesis H_0, and also to maximize the likelihood under the alternative hypothesis H_1. If the distribution of the random sample $Y = (y_1, \ldots, y_n)$ depends on a parameter vector θ, and if $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ are any two nested hypotheses, then the likelihood ratio (LR) statistic for testing H_0 against H_1 is defined as

$$\lambda(\mathbf{y}) = L_0(\hat{\theta}_0)/L_1(\hat{\theta}_1) \tag{3.84}$$

where $L_i(\hat{\theta}_i)$ is the largest value which the likelihood function takes in the parameter space $\Theta_i, i = 0, 1$. Equivalently, we may use the statistic

$$-2\log\lambda = 2(l_1(\hat{\theta}_1) - l_0(\hat{\theta}_0)), \tag{3.85}$$

where $l_i(\hat{\theta}_i) = \log L_i(\hat{\theta}_i), i = 0, 1$. In general, one tends to favor H_1 when the LR statistic (3.84) is low, and H_0 when it is high. A test procedure based on the LR statistic is as follows:

The LRT of size α for testing H_0 against H_1 has as its rejection region

$$R_c = \{\mathbf{y} : \lambda(\mathbf{y}) < c\} \tag{3.86}$$

where c is determined so that

$$\sup_{\theta \in \Theta_0} Pr_\theta(\mathbf{y} \in R_c) = \alpha. \tag{3.87}$$

However, it may not be possible to obtain exact size α, especially when $\lambda(Y)$ is a discrete random variable. If such a c does not exist, we choose an integer c^* such that

$$Pr_\theta(\mathbf{y} \in R_{c^*}) \leq \alpha \text{ and } Pr_\theta(\mathbf{y} \in R_{c^*-1}) > \alpha. \tag{3.88}$$

The LRT has the following very important asymptotic property [21].

Theorem 3.4. *(Wald Theorem) In the notation of (3.85), if Θ_1 is a region in $\mathbf{R}^d$, and if Θ_0 is an r-dimensional subregion of Θ_1, then under suitable regularity conditions including $\theta \in int(\Theta_0) \cap int(\Theta_1)$, $-2\log\lambda$ has an asymptotic χ^2_{d-r} distribution as $n \to \infty$.*

Remark 3.5. *For degrees of freedom $d > 100$, $\sqrt{2\chi^2_d} \overset{\mathcal{D}}{\approx} N(\sqrt{2d-1}, 1)$.*

3.6.2 The LRT for (M4a) against (M3)

In this section, we shall test whether the more general model (M3) fits better than PARAFAC (M4a). Consider the hypotheses $H_0 : \theta \in \Theta_{M4a}$ against $H_1 : \theta \in \Theta_{M3}$, where Θ_{M3} is defined in (2.56), and Θ_{M4a} is defined in (2.69).

Maximizing the likelihood under $H_1 : \theta \in \Theta_{M3}$

To maximize the likelihood under the alternative H_1, we use Newton-Raphson method to minimize $-2logL \equiv -2l_y(\theta)$, where $\theta \in \Theta_{M3}$. For reducing the parameter dimension, we use the profile likelihood method. We first fix Λ, the (M3) model in (2.50) under the parameter space Θ_{M3}. The model can be written as

$$Y^{(r,a,s)} = \Lambda^* f^{(a,s)*} + U^{(r,a,s)*} \tag{3.89}$$

where $Y^{(r,a,s)}$ is a $p \times 1$ vector, $\Lambda^* = (\mathbf{1}|\Lambda)$ is a $p \times (q+1)$ matrix, $\mathbf{f}^{(a,s)*} = \begin{pmatrix} \mu_{as} \\ \mathbf{f}^{(a,s)} \end{pmatrix}$, and $\mathbf{U}^{(r)*}$ is a $p \times AS$ matrix. Define $F^* \equiv (\mathbf{f}^{(1,1)*}|\mathbf{f}^{(1,2)*}|\cdots|\mathbf{f}^{(A,S)*})$ which is a $(q+1) \times AS$ matrix.

The log-likelihood function $l(\theta)$ is

$$l(\theta) = -\frac{1}{2}\sum_{a=1}^{A}\sum_{s=1}^{S}[\, pR\log(\sigma^2_{as}) + \frac{1}{\sigma^2_{as}}\sum_{r=1}^{R}\|Y^{(r,a,s)} - \Lambda^* f^{(a,s)*}\|^2] \tag{3.90}$$

The second term on the right hand side of (3.90) can be decomposed as

$$\begin{aligned}\sum_{r=1}^{R}\|Y^{(r,a,s)}-\Lambda^* f^{(a,s)*}\|^2 &= \sum_{r=1}^{R}(\|Y^{(r,a,s)}-\bar{Y}^{(\cdot,a,s)}\|^2)\\ &+ R(\|\bar{Y}^{(\cdot,a,s)}-\Lambda^* f^{(a,s)*}\|^2) \qquad (3.91)\end{aligned}$$

where

$$\bar{Y}^{(\cdot,a,s)} \equiv \frac{1}{R}\sum_{r=1}^{R} Y^{(r,a,s)}.$$

Minimizing the second term in (3.91), we get

$$\widehat{f}^{(a,s)*} = (\Lambda^{*\,t}\ \Lambda^*)^{-1}\ \Lambda^{*\,t}\ \bar{Y}^{(\cdot,a,s)} \qquad (3.92)$$

Plugging $\widehat{f}^{(a,s)*}$ into the log-likelihood function $l(\theta)$ and maximizing it with respect to σ^2_{as}, we have

$$\widehat{\sigma}^2_{as} = \frac{1}{pR}\sum_{r=1}^{R}\|Y^{(r,a,s)}-\widehat{\Lambda}^*\widehat{f}^{(a,s)*}\|^2 \qquad (3.93)$$

Therefore, the maximized log-likelihood function $l(\widehat{\theta})$ under (M3) is

$$l(\hat{\theta}) = -\frac{RAS}{2} - \frac{pR}{2}\sum_{a=1}^{A}\sum_{s=1}^{S}\log(\hat{\sigma}^2_{as}). \qquad (3.94)$$

Maximizing the likelihood under $H_0 : \theta \in \Theta_{M4a}$

To get the maximum likelihood estimator in (M4a), we can use MATLAB and the N-way Toolbox which can be downloaded from

http://www.models.kvl.dk/courses/. We will discuss it in the next chapter.

Chapter 4

Computational Methods

4.1 EM Algorithm

4.1.1 Introduction

The EM (expectation-maximization) algorithm was first advocated by Dempster, Laird, and Rubin in 1977 [6] for deriving maximum likelihood estimators from incomplete data. It is a very popular and widely applicable computational tool in various statistical models. The attractive features of EM algorithm are its simplicity and stability (e.g. automatic monotone convergence in likelihood). It is often used as an alternative to the Newton-Raphson method, Fisher-scoring method and other optimization methods when the latter are too expensive to use or too complicated to implement. However, the EM algorithms often suffer from slow convergence. Whether this is a real problem in practice depends on models, data sizes, and programs used. Many acceleration methods have been proposed to speed up the convergence of the EM algorithm since Dempster, Laird, and Rubin (1977). Jamshidian and Jennrich [11] classify the acceleration methods into three groups: pure, hybrid, and EM-type accelerators. For accelerating the slow convergence of EM with stability and global convergence, a line search needs to be employed with any acceleration method, which may ruin the sim-

plicity of the EM algorithm. In fact, the simplicity of the EM algorithm is a much more attractive feature if we consider the operating efficiency from the stage of formulating the likelihood to the stages of deriving and implementing an algorithm.

The idea of the EM algorithm is to treat the unobservable common factors as missing data and the complete data to comprise the observations together with these unobservable factors. Let Y be a p-dimensional random vector corresponding to the observed data and $p_Y(y, \theta)$ be the probability density function, where θ is a vector of unknown parameters within the parameter space Θ. Let Z be the random vector containing the missing data portion. Then $X = (Y, Z)$ denotes the vector containing both the observed and missing data, called the complete data, and $p_X(x, \theta)$ denotes the probability density function of X.

Let $l_X(\theta) = log\ p_X(X, \theta)$, which is the log likelihood function based on the complete data and $l_Y(\theta) = log\ p_Y(Y, \theta)$, which is the log likelihood function based on the incomplete data. The goal of the EM algorithm is to find the maximum likelihood estimate of θ, which is the point achieving the maximum of $l_Y(\theta)$.

The EM algorithm approaches indirectly the problem of maximizing the log likelihood $l_Y(\theta)$ based on incomplete data by proceeding iteratively in terms of the log likelihood based on the complete data, $l_X(\theta)$. Since $l_X(\theta)$ is unobservable, it is replaced by the conditional expectation given the observation and the values of parameters in mth iteration:

$$\theta^{(m+1)} = \arg\max_{\theta \in \Theta} E[l_X(\theta)|Y, \theta^{(m)}]. \tag{4.1}$$

Thus, starting with an initial value $\theta^{(0)} \in \Theta$, one finds θ^*, a stationary point of $l_Y(\theta)$, by alternating between the following two steps ($m = 0, 1, \ldots$):

E-step: impute the complete data log likelihood $l_X(\theta)$ by

$$Q(\theta, \theta^{(m)}) = E[l_X(\theta)|Y, \theta^{(m)}] \tag{4.2}$$

M-step: determine $\theta^{(m+1)}$ by maximizing the imputed log likelihood $Q(\theta, \theta^{(m)})$ regarded as a function of θ with $\theta^{(m)}$ fixed:

$$Q(\theta^{(m+1)}, \theta^{(m)}) \geq Q(\theta, \theta^{(m)}) \text{ for all } \theta \in \Theta. \tag{4.3}$$

The E-step and M-step are repeated by turns until they converge in a specified sense, such as the smallness of changes in $|\theta^{(m+1)} - \theta^{(m)}|$. Dempster, Laird and Rubin (1977) pointed out that the incomplete data log likelihood $l_Y(\theta)$ is non-decreasing on each iteration of an EM algorithm, that is,

$$l_Y(\theta^{(m+1)}) \geq l_Y(\theta^{(m)}) \tag{4.4}$$

for $m = 0, 1, 2, \ldots$ This property is useful for debugging the program code for the EM algorithm. Moreover, if the log likelihood $l_Y(\theta)$ based on incomplete data y is bounded above, the value of the log likelihood in the iteration process $l_Y(\theta^{(m)})$ converges to a stationary value of $l_Y(\theta)$.

Under general conditions, if $\theta^{(m)}$ converges, the limiting value can be proved to be either a local maximum or a saddle point of $l_Y(\theta)$ (Boyles, 1983; Wu, 1983). Therefore, if the likelihood function is unimodal and the first derivative of the function Q defined in equation (3.1.2) is continuous with respect to $\theta^{(m)}$ and θ, the EM algorithm converges to the only local maximum. Generally speaking, however, the likelihood function of the incomplete data is not necessarily unimodal. Therefore, it is necessary to compare the values of the log likelihood of the convergence value, starting with many initial values.

4.1.2 EM algorithm and (M0) model

The random effect factor model (M0) is

$$Y^{(r)} = \underline{\mu} + \Lambda \mathbf{f}^{(r)} + U^{(r)} \tag{$M0$}$$

where all assumptions regarding to $Y^{(r)}, \Lambda, \mathbf{f}^{(r)}$, and $U^{(r)}$ are specified in Chapter 1. We consider the special case of (M0) model when μ is a zero vector. Let X be the complete data, which includes observation vectors $Y^{(r)}$ and unobservable vectors $\mathbf{f}^{(r)}$, $r = 1, 2, \ldots, R$. That is, $X = (Y, \mathbf{f})$. Thus, the complete data X becomes a $(p+q)$-dimensional vector. It is assumed that $X^{(1)}, X^{(2)}, \ldots, X^{(R)}$ are independently and identically distributed, and that the *common factors* $\mathbf{f}^{(1)}, \mathbf{f}^{(2)}, \ldots, \mathbf{f}^{(R)}$ independently and identically normally distributed with zero mean and identity covariance matrix I_q; that is,

$$\mathbf{f}^{(r)} \sim N_q(0, I_q). \tag{4.5}$$

The vectors $\mathbf{f}^{(r)}$ are independent of the errors $U^{(r)}$, which are assumed to be independently and identically distributed as $N_p(0, \Psi)$ where Ψ is a $p \times p$ diagonal matrix. Given the unobservable random effect $\mathbf{f}^{(r)}$, the conditional probability distribution over $Y^{(r)}$ is given by

$$Y^{(r)}|\mathbf{f}^{(r)} \sim N_p(\underline{\mu} + \Lambda\mathbf{f}^{(r)}, \Psi), \tag{4.6}$$

where $\Psi = diag(\psi)$. Unconditionally, $\{Y^{(r)}\}$ is independently and identically distributed with

$$Y^{(r)} \sim N_p(\underline{\mu}, \Lambda\Lambda^t + \Psi). \tag{4.7}$$

In (M0), the log-likelihood function is

$$l(\theta) = -\frac{pR}{2}\log 2\pi - \frac{R}{2}\log|\Sigma_y| - \frac{R}{2}tr\,[\,\Sigma_y^{-1}\,\frac{1}{R}\sum_{r=1}^{R}(y^{(r)} - \underline{\mu})(y^{(r)} - \underline{\mu})^t]. \tag{4.8}$$

Since

$$\begin{aligned}
&\sum_{r=1}^{R}(y^{(r)}-\underline{\mu})(y^{(r)}-\underline{\mu})^t \\
&= \sum_{r=1}^{R}[\ (y^{(r)}-\bar{y})+(\bar{y}-\underline{\mu})]\ [\ (y^{(r)}-\bar{y})+(\bar{y}-\underline{\mu})]^t \\
&= \sum_{r=1}^{R}[\ (y^{(r)}-\bar{y})(y^{(r)}-\bar{y})^t]+2\sum_{r=1}^{R}[(y^{(r)}-\bar{y})(\bar{y}-\underline{\mu})^t]+R\ (\bar{y}-\underline{\mu})(\bar{y}-\underline{\mu})^t \\
&= \sum_{r=1}^{R}[\ (y^{(r)}-\bar{y})(y^{(r)}-\bar{y})^t]+R\ (\bar{y}-\underline{\mu})(\bar{y}-\underline{\mu})^t, \qquad (4.9)
\end{aligned}$$

the MLE of $\underline{\mu}$ is $\widehat{\underline{\mu}} = \bar{y}$ (by Anderson [2] p60-63). Since the probability density function of the complete data X can be written as $p(x) = p(y|f)p(f)$ with

$$p(y|f) = \frac{1}{(2\pi)^{p/2}}|\Psi|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(y-\underline{\mu}-\Lambda f)^t\Psi^{-1}(y-\underline{\mu}-\Lambda f)\} \qquad (4.10)$$

and

$$p(f) = \frac{1}{(2\pi)^{q/2}}\exp\{-\frac{1}{2}(f^t f)\} \qquad (4.11)$$

then

$$\begin{aligned}
p(x) &= \frac{1}{(2\pi)^{p/2}}|\Psi|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(y-\underline{\mu}-\Lambda f)^t\Psi^{-1}(y-\underline{\mu}-\Lambda f)\} \\
&\quad \times\frac{1}{(2\pi)^{q/2}}\exp\{-\frac{1}{2}(f^t f)\} \qquad (4.12)
\end{aligned}$$

The complete data log likelihood function is

$$\begin{aligned} l_X(\theta) &= log\prod_{r=1}^{R} p(x^{(r)}) \\ &= -\frac{R(p+q)}{2}log(2\pi) - \frac{R}{2}log|\Psi| - \frac{1}{2}\sum_{r=1}^{R} tr[(f^{(r)})(f^{(r)})^t] \\ &\quad -\frac{1}{2}\sum_{r=1}^{R}(y^{(r)} - \underline{\mu} - \Lambda f^{(r)})^t\ \Psi^{-1}(y^{(r)} - \underline{\mu} - \Lambda f^{(r)}) \\ &= -\frac{R(p+q)}{2}log(2\pi) - \frac{R}{2}log|\Psi| - \frac{R}{2}\ tr[C_{ff}] \\ &\quad -\frac{R}{2}tr[\Psi^{-1}\ \frac{1}{R}\sum_{r=1}^{R}(y^{(r)} - \underline{\mu} - \Lambda f^{(r)})(y^{(r)} - \underline{\mu} - \Lambda f^{(r)})^t] \end{aligned} \quad (4.13)$$

where $C_{ff} = \frac{1}{R}\sum_{r=1}^{R} f^{(r)}(f^{(r)})^t$.

Plugging $\widehat{\underline{\mu}}$ into (4.13), we have

$$\begin{aligned} l_X(\Lambda,\psi) &= l_X(\widehat{\underline{\mu}},\Lambda,\psi) \\ &= -\frac{R(p+q)}{2}log(2\pi) - \frac{R}{2}log|\Psi| - \frac{R}{2}\ tr[C_{ff}] \\ &\quad -\frac{R}{2}tr[\Psi^{-1}\ \frac{1}{R}\sum_{r=1}^{R}(y^{(r)} - \bar{y} - \Lambda f^{(r)})(y^{(r)} - \bar{y} - \Lambda f^{(r)})^t] \\ &= -\frac{R(p+q)}{2}log(2\pi) - \frac{R}{2}log|\Psi| - \frac{R}{2}\ tr[C_{ff}] - \frac{R}{2}tr[\Psi^{-1}\ C_{yy}] \\ &\quad +R\cdot tr[\Psi^{-1}\ \Lambda C_{fy}] - \frac{R}{2}tr[\Psi^{-1}\ \Lambda C_{ff}\Lambda^t] \end{aligned} \quad (4.14)$$

where $\Psi = diag(\psi)$, $C_{yy} = \frac{1}{R}\sum_{r=1}^{R}(y^{(r)}-\bar{y})(y^{(r)}-\bar{y})^t$, and $C_{fy} = \frac{1}{R}\sum_{r=1}^{R} f^{(r)}(y^{(r)}-\underline{\mu})^t$. Suppose that $\Lambda^{(m)}$ and $\Psi^{(m)}$ denote the current values of Λ and Ψ after m cycles of the algorithm and $\theta^{(m)} \equiv (\underline{\mu}^{(m)}, \Lambda^{(m)}, \Psi^{(m)})$ with $\underline{\mu}^{(m)}$ substituted by $\widehat{\underline{\mu}}$. By Rubin and Thayer [18], the basis of the EM algorithm for maximum likelihood factor analysis is:

E-step: Compute $E[f^{(r)}|y^{(r)}]$ and $E[f^{(r)}(f^{(r)})^t|y^{(r)}]$ for each data point $y^{(r)}$, given $\Lambda^{(m)}$ and $\Psi^{(m)}$.

M-step:

$$\Lambda^{(m+1)} = (\sum_{r=1}^{R}(y^{(r)} - \bar{y})E[\mathbf{f}^{(r)}|y^{(r)}, \theta^{(m)}]^t) \ (\sum_{r=1}^{R} E[\mathbf{f}^{(r)}(\mathbf{f}^{(r)})^t|y^{(r)}, \theta^{(m)}])^{-1} \quad (4.15)$$

and

$$\begin{aligned} \psi^{(m+1)} &\equiv diag(\Psi^{(m+1)}) \\ &= \frac{1}{R} diag\{\sum_{r=1}^{R}[(y^{(r)} - \bar{y})(y^{(r)} - \bar{y})^t \\ &\quad -\Lambda^{(m+1)}E[\mathbf{f}^{(r)}|y^{(r)}, \theta^{(m)}](y^{(r)} - \bar{y})^t \ ] \ \}. \end{aligned} \quad (4.16)$$

For simplifying the notations, we define

$$B \equiv \frac{1}{R}\sum_{r=1}^{R}(y^{(r)} - \bar{y})E[\mathbf{f}^{(r)}|y^{(r)}, \theta^{(m)}]^t \quad (4.17)$$

and

$$C \equiv \frac{1}{R}\sum_{r=1}^{R} E[\mathbf{f}^{(r)}(\mathbf{f}^{(r)})^t|y^{(r)}, \theta^{(m)}]. \quad (4.18)$$

Then the equations (4.15) and (4.16) can be simplified as

$$\Lambda^{(m+1)} = B \ C^{-1} \quad (4.19)$$

and

$$\psi^{(m+1)} \equiv diag(\Psi^{(m+1)}) = diag\{C_{yy} - \Lambda^{(m+1)} \ B^t\}. \quad (4.20)$$

Now we will express B and C in terms of $\Lambda^{(m)}$ and $\Psi^{(m)}$ by calculating $E[\ \mathbf{f}^{(r)} \mid y^{(r)}, \theta^{(m)}]$ and $E[\ \mathbf{f}^{(r)} \ \mathbf{f}^{(r)^t} \ |y^{(r)}, \theta^{(m)}]$. Since

$$\begin{pmatrix} Y^{(r)} \\ \mathbf{f}^{(r)} \end{pmatrix} \sim N\left(\begin{bmatrix} \underline{\mu} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \Lambda\Lambda^t + \Psi & \Lambda \\ \Lambda^t & I_q \end{bmatrix}\right) \quad (4.21)$$

The conditional distribution of $\mathbf{f}^{(r)}$ given $Y^{(r)}$ is

$$\mathbf{f}^{(r)}|Y^{(r)} \sim N_q(\ \mu_f + \Sigma_{21}\Sigma_{11}^{-1}(Y^{(r)} - \mu_Y)\ ,\ \Sigma_{22.1}) \quad (4.22)$$

where $\mu_Y = \underline{\mu}$, $\mu_f = 0$, $\Sigma_{11} = \Lambda\Lambda^t + \Psi$, $\Sigma_{21} = \Lambda^t$, and $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ $= I_q - \Lambda^t\,(\Lambda\Lambda^t + \Psi)^{-1}\Lambda$. Then

$$\begin{aligned} E[\mathbf{f}^{(r)}|y^{(r)}, \theta^{(m)}] &= \Sigma_{21}\Sigma_{11}^{-1}(y^{(r)} - \bar{y}) \\ &= (\Lambda^{(m)})^t\,(\Lambda^{(m)}(\Lambda^{(m)})^t + \Psi^{(m)})^{-1}\,(y^{(r)} - \bar{y}) \\ &= (K^{(m)})^t\,(y^{(r)} - \bar{y}) \end{aligned} \tag{4.23}$$

where

$$K^{(m)} = (\Lambda^{(m)}(\Lambda^{(m)})^t + \Psi^{(m)})^{-1}\Lambda^{(m)} \tag{4.24}$$

Similarly,

$$\begin{aligned} & E[\mathbf{f}^{(r)}(\mathbf{f}^{(r)})^t|y^{(r)}, \theta^{(m)}] \\ &= E[\mathbf{f}^{(r)}|y^{(r)}, \theta^{(m)}]\;E[\mathbf{f}^{(r)}|y^{(r)}, \theta^{(m)}]^t \\ &+ Var(\mathbf{f}^{(r)}|y^{(r)}, \theta^{(m)}) \\ &= \Lambda^{(m)\,t}(\Lambda^{(m)}\Lambda^{(m)\,t} + \Psi^{(m)})^{-1}(y^{(r)} - \bar{y})(y^{(r)} - \bar{y})^t(\Lambda^{(m)}\Lambda^{(m)\,t} + \Psi^{(m)})^{-1}\Lambda^{(m)} \\ &+ I_q - \Lambda^{(m)\,t}(\Lambda^{(m)}(\Lambda^{(m)})^t + \Psi^{(m)})^{-1}\Lambda^{(m)} \\ &= (K^{(m)})^t\,(y^{(r)} - \bar{y})(y^{(r)} - \bar{y})^t\,K^{(m)} + I_q - (\Lambda^{(m)})^t\,K^{(m)} \end{aligned} \tag{4.25}$$

Therefore, from (4.17) and (4.23),

$$B = C_{yy}(\Lambda^{(m)}(\Lambda^{(m)})^t + \Psi^{(m)})^{-1}\Lambda^{(m)} = C_{yy}\,K^{(m)} \tag{4.26}$$

and from (4.18) and (4.25),

$$\begin{aligned} C &= I_q - \Lambda^{(m)\prime}(\Lambda^{(m)}\Lambda^{(m)\prime} + \Psi^{(m)})^{-1}\Lambda^{(m)} \\ &+ \Lambda^{(m)\prime}(\Lambda^{(m)}\Lambda^{(m)\prime} + \Psi^{(m)})^{-1}C_{yy}(\Lambda^{(m)}\Lambda^{(m)\prime} + \Psi^{(m)})^{-1}\Lambda^{(m)} \\ &= I_q - (\Lambda^{(m)})^t\,K^{(m)} + (K^{(m)})^t\,C_{yy}\,K^{(m)} \end{aligned} \tag{4.27}$$

Thus, the new estimated parameter $(\Lambda^{(m+1)}, \Psi^{(m+1)})$ is given in equations (4.19) and (4.20) through B and C as a function of C_{yy}, $\Lambda^{(m)}$, and $\Psi^{(m)}$.

4.2 Newton-Raphson method

For a function $g : \mathbf{R}^p \to \mathbf{R}$, the gradient is the vector

$$\nabla g(\theta) = (\frac{\partial g(\theta)}{\partial \theta_1}, \ldots, \frac{\partial g(\theta)}{\partial \theta_p})^t \tag{4.28}$$

and the Hessian matrix is the matrix of second partial derivatives

$$\nabla^{\otimes 2} g(\theta) \equiv (\frac{\partial^2 g(\theta)}{\partial \theta_i \, \partial \theta_j}), \; 1 \leq i, j \leq p. \tag{4.29}$$

The directional derivative of a function $g : \mathbf{R}^p \to \mathbf{R}$ at x in the direction v is defined by

$$\lim_{\delta \to 0} \frac{g(x + \delta v) - g(x)}{\delta} = \frac{\partial}{\partial \delta} g(x + \delta v) \mid_{\delta = 0} = v^{\,t} \, \nabla g(x) \tag{4.30}$$

For smooth functions, g is convex on a set Θ if $\nabla^{\otimes 2} g(\theta)$ is nonnegative definite for all $\theta \in \Theta$. If $\nabla^{\otimes 2} g(\theta)$ is positive definite for all $\theta \in \Theta$, then g is strictly convex on Θ.

The general unconstrained minimization problem for a smooth function g is to find a $\hat{\theta}$ such that

$$g(\hat{\theta}) = \min_{\theta} g(\theta), \tag{4.31}$$

where the minimum is over all $\theta \in \Theta$. In general such a $\hat{\theta}$ need not exist. Another problem is that there may be multiple local minima. Generally, it is impossible to guarantee convergence of a numerical algorithm to a global minimum, unless the function is convex everywhere in Θ. For this reason, the problem considered will be to find a local minimum. Maximum likelihood estimates for a log likelihood $l(\theta)$ can be found by minimizing $-l(\theta)$.

For a smooth function g, if $\hat{\theta}$ is a local minimum, then $\nabla g(\hat{\theta}) = 0$. If $g(\hat{\theta}) = 0$ and $\nabla^{\otimes 2} g(\theta)$ is nonnegative definite, then $\hat{\theta}$ is a local minimum. Thus the search for a

minimum can be to find points $\hat{\theta}$ satisfying $\nabla g(\hat{\theta}) = 0$. Such points need not be local minima, since they could also be local maxima or saddle points.

Many algorithms for searching for a local minimum are similar to the following outline:

1. Given the current point x_0, choose a direction v in which to move next.

2. Find a point $x_1 = x_0 + sv$ such that $g(x_1) < g(x_0)$.

3. Set $x_0 = x_1$, and repeat the first two steps until convergence.

For getting successful convergence, it is important that the direction v chosen at each stage be a descent direction for g. A direction v is a descent direction for g at x_0 if

$$g(x_0 + sv) < g(x_0) \text{ for } 0 < s < \delta, \tag{4.32}$$

for some $\delta > 0$. It is clear that v is a descent direction for g at x_0 if $v^{\,t}\ \nabla g(x_0) < 0$ for δ small enough. We denote the vector of parameter values after the k'th iteration by $\theta^{(k)}$ and its converged point by θ^*. Therefore, consider the iteration stopping criterion according to

$$(1) \quad \|\nabla g(\theta^{(k)})\| < 10^{-6} \tag{4.33}$$

$$(2) \quad \|\theta^{(k+1)} - \theta^{(k)}\| < 10^{-6}. \tag{4.34}$$

To maximize the log likelihood function $l(\theta; y)$, we take $g(\theta) = -l(\theta; y)$. The Newton-Raphson method approximates the objective function (the incomplete data log likelihood function) by a quadratic function and takes its maximizer as the next parameter value. Its formula is:

$$\theta^{(k+1)} = \theta^{(k)} + I^{-1}(\theta^{(k)}; y)\nabla_\theta l(\theta^{(k)}; y) \tag{4.35}$$

An iterative numerical method is said to converge linearly if it holds that with some constant c ($0 < c < 1$) and positive integer k_0,

$$\|\theta^{(k+1)} - \theta^*\| \leq c\|\theta^{(k)} - \theta^*\| \text{ for any } k \geq k_0. \tag{4.36}$$

The constant c is called the convergence rate. If it holds that with some sequence $\{c_k\}$ converging to 0 and positive integer k_0,

$$\|\theta^{(k+1)} - \theta^*\| \leq c_k\|\theta^{(k)} - \theta^*\| \text{ for any } k \geq k_0, \tag{4.37}$$

then the method is said to converge superlinearly. if it holds that with some constant c ($0 < c < 1$) and positive integer k_0,

$$\|\theta^{(k+1)} - \theta^*\| \leq c\|\theta^{(k)} - \theta^*\|^2 \text{ for any } k \geq k_0, \tag{4.38}$$

then the method is said to converge quadratically. A numerical method with the superlinear or quadratic convergence property converges rapidly after the parameter value comes close to θ^*, while a method with the linear convergence property might take a fairly large number of iterations even after the parameter value comes close to θ^*. The Newton-Raphson method converges quadratically, which is extremely fast and is an attractive feature. On the other hand, the Newton-Raphson method requires the observed information matrix, and calculating the Hessian of the objective function takes much more computational time when the parameter dimension increases.

Lindstrom and Bates [13] employed the better quasi-Newton method which do not require calculation of second derivatives and a approximate Hessian matrix is always non-singular. Its update formula is:

$$\theta^{(k+1)} = \theta^{(k)} + \alpha_k B_k^{-1}(\theta^{(k)}; y)\nabla_\theta l(\theta^{(k)}; y) \tag{4.39}$$

where the matrix B_k is updated using only the change in gradient $q_k = \nabla_\theta l(\theta^{(k)}; y) - \nabla_\theta l(\theta^{(k-1)}; y)$ and the change in parameter value $s_k = \theta^{(k)} - \theta^{(k-1)}$. Quasi-Newton

method is like Newton's method with line search, except that Hessian matrix is approximated by a symmetric positive definite matrix which is updated at each iteration. The convergence speed of quasi-Newton algorithms is superlinear [13].

4.2.1 Newton-Raphson method on the profile likelihood

If the likelihood has a unique local maximum, then the maximum likelihood estimators should be the same no matter which numerical approach is used. Thus, we use the Newton-Raphson method on the profile likelihood to verify the results we got from the EM algorithm on the simulated data.

There is an R function *nlm* which finds a local minimum of a nonlinear function using a general Newton-Raphson method optimizer for an input R function. Based on *nlm*, we wrote another R function *ProfileLik* whose input is a data set, a starting point of θ_2, a few control parameters, and whose output is the MLE $\hat{\theta_2}$, the maximized value of the profile log-likelihood, and the restricted MLE $\hat{\theta_1}(\hat{\theta_2})$.

4.3 Computational results on simulated data

In this section, we implement Splus/R functions on simulated data. In our examples, the dimensions are $p = 6, q = 2$, the sample size is $n = 100$, and the parameter is $\theta = (\Lambda, \psi)$ as described below. Since $Y \sim N_p(0, \Lambda\Lambda^t + diag(\psi))$, we can use the Splus command *rmvnorm* to randomly generate multiple data samples.

First, as true parameters $\theta_0 = (\Lambda_0, \psi_0)$ we chose $\Lambda_0 \in O^+_{pq}$ such that $\Lambda_0^t \mathbf{1} = 0$ and satisfying the condition in Theorem 2.4, and chose the entries of ψ_0 as independent $Unif([0, 0.5])$ variates. Thus, the parameter θ_0 is identifiable from the observed data. Using this (Λ_0, ψ_0), we randomly generated 4 sample data sets, specified to be cases

(A)-(D) in Figure 4.1. The values of (Λ_0, ψ_0) are listed in Appendix B.

Second, consider the true parameters $\theta_1 = (\Lambda_1, \psi_1)$ and choose Λ_1 with $e_j \in col(\Lambda)$. The entries of ψ_1 were chosen in the same way as the entries of ψ_0. Therefore, the parameter (Λ_1, ψ_1) is non-identifiable from the data set, by Lemma 2.12. Using this (Λ_1, ψ_1), we randomly generated 4 sample data sets, specified to be cases (O)-(R) in Figure 4.1. The values of (Λ_1, ψ_1) are listed in Appendix B.

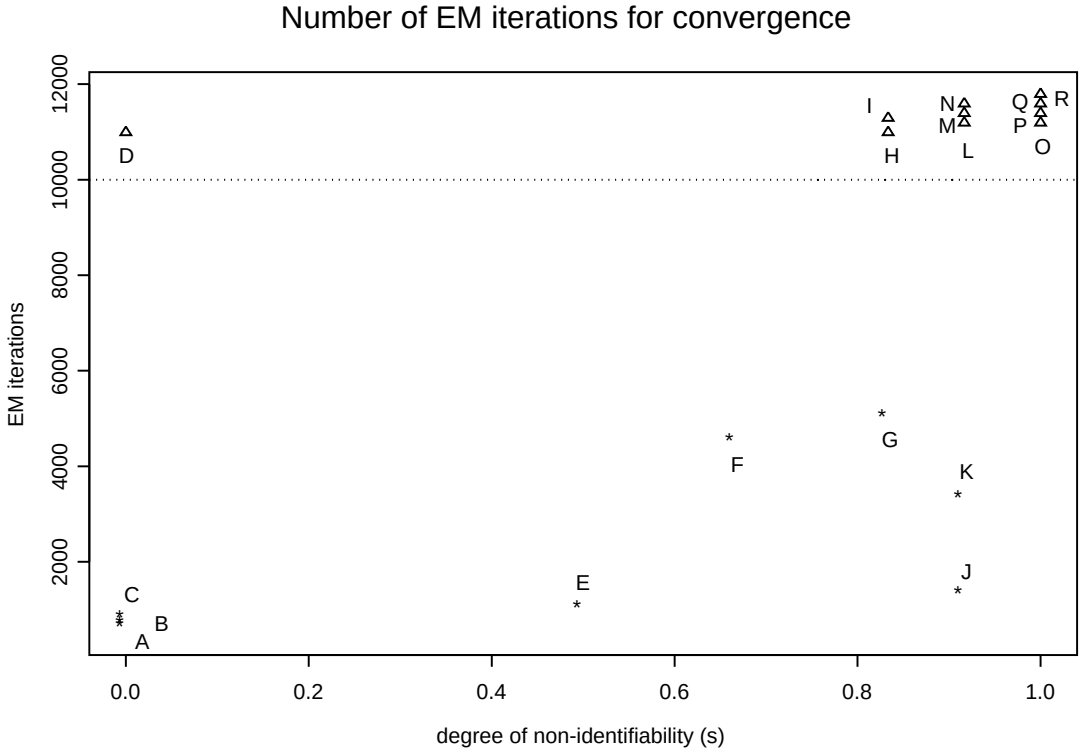

Figure 4.1: Number of iterations needed for EM convergence based on data samples generated by (Λ_s, ψ_s). The x coordinate is the degree of non-identifiability, denoted by s, which is a parameter of convex combination between identifiability and non-identifiability. The points above $10,000$ iterations have y-coordinate plotted arbitrarily, indicating that EM does not converges up to $10,000$ iterations for these data samples.

Finally, consider the convex curves $\Lambda_s = (1-s)\Lambda_0 + s\Lambda_1$ for $s \in (0,1)$ and choose ψ_{s0} such that the entries of ψ_{s0} are independent $Unif([0, 0.5])$ variates. We used this fixed ψ_{s0} and the matrices Λ_s which are different for each different value s, to generate data samples. When s is close to 1, Λ_s is close to Λ_1, and then the parameter (Λ_s, ψ_{s0}) is close to non-identifiable. Since we wanted to explore the behavior of EM algorithm and Newton-Raphson method when the parameter was close to non-identifiable, we specifically chose $s = 1/2, 2/3, 5/6$ and $11/12$. When $s = 1/2$, we generated one data set using (Λ_s, ψ_{s0}) as the true parameter, called case (E). When $s = 2/3$, the data set we generated was called case (F). Choosing $s = 5/6$, we generated three data sets, specified to be cases (G)-(I). Finally, with $s = 11/12$, we used the corresponding (Λ_s, ψ_{s0}) to generate five data samples, called cases (J)-(N) in Figure 4.1.

Using the true parameters as starting points, for each illustrative data set, we iterated 10000 times in the EM algorithm using formulas in (4.15) and (4.16). Applying the profile likelihood method in (3.24), we can express $\hat{\Lambda} = \hat{\Lambda}(\psi)$ as a function of ψ. We substituted it into the log likelihood function, used the Newton-Raphson optimization in R with command *nlm*, and chose various initial parameters from which to find the maximum likelihood estimates. We used the ψ values at the 300'th iteration or 5000'th iteration of the EM algorithm as the initial values of the *nlm* function. Based on the stopping criterion in (4.33), we obtained an MLE $(\hat{\Lambda}, \hat{\psi})$ in each case.

We followed the further steps in each data set:

Step 1. Check if the MLE is in the interior or boundary of the parameter space.

Step 2. Use the Hessian matrix we got from the output of *nlm* and calculate the maximum and minimum eigenvalues of the Hessian at the converged value of the parameter. Observe the *condition number*

$$r = \frac{\alpha_1}{\alpha_2} \tag{4.40}$$

where α_1 and α_2 are maximum and minimum eigenvalues of the negative Hessian matrix, respectively. If the ratio r is too large or α_2 is too small, then the Hessian is close to singular.

Step 3. Use the MLE, denoted by $\hat{\theta}$, obtained from *nlm* to check whether the EM converges and check the quality of convergence of EM algorithm. The iteration stopping criterion for EM algorithm is

$$|\theta^{(k)} - \hat{\theta}| < 10^{-3} \quad \text{for all } k \geq m. \tag{4.41}$$

where $\theta^{(k)}$ denotes the current values of θ after k iteration of the EM algorithm.

Step 4. Observe the convergence of EM to see whether it is approaching the boundary or remains in the interior of the parameter space.

4.3.1 Comparison of EM and Newton-Raphson algorithms

First, we observe the results of using the Newton-Raphson method with the profile likelihood strategy to find the MLE. We found that the Newton-Raphson algorithm converged in all of the 19 cases (cases (A)-(R)). The converged values of the parameters were in the interior of the parameter spaces in cases (A)-(C), (E)-(G), (J)-(K), that is, the $*$ points in Figure 4.1. In each of these cases, the gradient at the estimated maximum of log-likelihood was less than 10^{-6}, so the converged values are the ML estimators. In cases (D), (H), (I), (L)-(R), we found that the converged values of the parameters were very close to the boundary of the parameter space, that is, at least one of the components, say ψ_j, of the estimated entries of ψ was very close to 0. These points are indicated as $\triangle$ points in Figure 4.1. After forcing $\psi_j = 0$ and applying the same Newton-Raphson method with profile likelihood strategy in the reduced model (M1R) and using the condition introduced in Section 3.4, we found that the converged

values are the ML estimators and the MLE $\widehat{\theta} \equiv (\widehat{\Lambda}, \widehat{\psi})$ is in the boundary of the parameter space. Thus, our findings are as follows. The ML estimators were obtained in all of the cases using the Newton-Raphson algorithm. When the parameter is identifiable, the MLE was in the interior of the parameter space except in case (D). The MLE was on the boundary whenever the parameter (Λ_0, ψ_0) was non-identifiable (cases (O)-(R)). When the parameter was close to a non-identifiable value, the MLE had more chance to lie on the boundary. When the parameter was identifiable, the MLE was in the interior of the parameter space. We will discuss the exceptional case (D) later in this section.

Second, we observe the results of using the EM algorithm. Consider the number of iterations needed for EM to be convergent. We found (in Figure 4.1) that the EM algorithm did not converge up to 10000 iterations when the model was non-identifiable (cases (O)-(R)). When the model was identifiable, fewer iterations were needed. When the model was close to a non-identifiable parameter value, more iterations were needed for EM convergence or the EM algorithm had not converged even up to 10000 iterations. However, there are some exceptions. For example, in case (D), the EM algorithm did not converge even though the parameter was identifiable from the data. In case (J), the model is close to non-identifiable, but it only took 1200 iterations to get the EM algorithm to converge.

Now, we explore the reason why the EM algorithm did not converge even though the parameter was identifiable. In each of cases (A)-(D), the parameter (Λ_0, ψ_0) was identifiable. The values of the components of ψ_0 were not close to zero (Appendix B), so neither were their ML estimators. We found that only 500 iterations were needed to get the EM algorithm to converge, and the converged values are very close to the MLE obtained from the Newton-Raphson algorithm. In case (C), the minimum value of

the components of ψ_0 was 0.03667 and the MLE was in the interior of the parameter space. We found that 700 iterations were needed to get EM algorithm to converge and the convergent values were also very close to the MLE. In case (D), there were two components of ψ_0, 0.01616 and 0.02967, close to 0, and the MLE was on the boundary of the parameter space. We found that the EM algorithm did not converge up to $10,000$ iterations. We also found that $\theta^{(k)}$, the values of parameter at k'th EM iteration, approached the same MLE even though the speed of approach was very slow. Thus, the number of iterations needed for EM to converge was associated with whether ψ is close to the boundary when the model is identifiable.

We next compare the estimate we got from the EM algorithm with the MLE from Newton-Raphson method. The Newton-Raphson method on the profile likelihood was shown to give results for each data set that agreed with the EM algorithm. That is, when the MLE we got from *nlm* function was in the interior of the parameter space, then the estimate from the EM algorithm was also in the interior of the parameter space and was close to the MLE.

Now, we explore in cases (E)-(N) the convergence of the EM algorithm when the model was close to non-identifiable. Especially, we are interested in case (J) where the model is close to non-identifiable, but it only took 1200 iterations to get the EM algorithm to converge. Let us observe the condition number r for each case: In Table 4.3.1, we record the condition number r in each case. We found that when the MLE approaches the boundary of the parameter space, the number r is extremely large ($r > 10^7$). That is, the hessian matrix at the estimate maximum of log-likelihood is close to singular. In case (D), the condition number $r > 10^7$ and the EM algorithm does not converge up to 10000 iterations even though the model is identifiable. When a model is nearly non-identifiable, we expect that the EM algorithm will not be able to converge

Case	(A)	(B)	(C)	(D)	(E)
r	20.26	31.89	40.48	390985.4	712.27
EM/nlm				△ ♠	
Case	(F)	(G)	(H)	(I)	(J)
r	126663.8	9747.19	2245182.00	3365352.00	629.05
EM/nlm		△ ♠	△ ♠	△ ♠	
Case	(K)	(L)	(M)	(N)	(O)
r	1253.02	1530232.00	8697203.00	2002751.00	4923976.00
EM/nlm		△ ♠	△♠	△ ♠	△♠

Table 4.1: Table for cases (A)-(O) with the condition number r. The symbol △ indicates the EM algorithm failed to converge and ♠ indicates that the MLE was on the boundary of the parameter space.

up to 10000 iterations and the MLE we get from *nlm* should be in the boundary of the parameter space. However, that is not true in case (J). Observe that the condition number r in case (J) was 629.05 which is small compared to 10^7 and the EM algorithm converges after 1200 iteration. Also, in case (J), both the MLE from *nlm* and EM are close to each other and in the interior of parameter space. Thus, we found that the *condition number* r is strongly associated with the behavior of EM algorithm and Newton Raphson method.

The convergence of the EM algorithm is based on the following criterion:

$$\|\theta^{(k)} - \widehat{\theta}\| < 10^{-3} \tag{4.42}$$

where $\theta^{(k)}$ is the value of the parameter at k'th EM iteration and $\widehat{\theta}$ is the MLE obtained

from the Newton Raphson method. The symbol $\triangle$ in the following table indicates the EM algorithm failed to converge and ♠ indicates that the MLE was on the boundary of the parameter space.

4.4 The LRT for (M4a) against (M3)

We introduced in chapter 3 the general idea of the Likelihood Ratio Test (LRT) and discussed the problem of maximizing the likelihood under $H_1 : \theta \in \Theta_{M3}$. Now we discuss how to maximize the likelihood under $H_0 : \theta \in \Theta_{M4a}$.

4.4.1 Maximize the likelihood under $H_0 : \theta \in \Theta_{M4a}$

To get the maximum likelihood estimator in (M4a), we can use MATLAB and the N-way Toolbox which can be downloaded from http://www.models.kvl.dk/courses/. The N-way Toolbox is compatible with MATLAB 5.x and higher, and can be used to fit "multi-way" models including PARAFAC (M4a) and (M4) and Tucker (T3). The freely downloadable reference is:
R. Bro The N-way on-line course on PARAFAC and PLS
http://www.models.kvl.dk/courses/; 1998-2002.

To fit a PARAFAC model and investigate the model, we use the MATLAB function *parafac* in the N-way Toolbox. The input is a data array, the number of factors sought, and a few optional constraints. The optional constraints can be put on the loadings of the different modes for obtaining orthogonal, nonnegative, or unimodal solutions. If the constraint is not defined, then no constraints are used. In (M4a), $\Lambda_\star$ need not have orthogonal columns, so we can use the default of no constraint. We can also set the optional inputs for the convergence criterion. The PARAFAC model is fit in a least

square sense, that is, by minimizing the norm $\|Y - M\|^2$ where Y is the input data and M is the PARAFAC model. The fit of a model is measured by the sum of squares of residuals. From the data and the model, the fit may thus be obtained.

The algorithm for fitting the PARAFAC model is a so-called alternating least squares algorithm. It is iterative and stops when the relative difference in fit between two successive iterations is below a certain limit. For most types of data this limit can be set to 10^{-6} (default in the algorithm), which will ensure that the model is correct and that not too many iterations are used. For some data, the model is very difficult to fit and a lower convergence criterion may therefore be needed. To assess convergence, the following steps may be used:

(1) Fit the models several times using random initialization.

(2) If all models have the same fit (i.e. loss function value) the models have converged.

(3) If all but a small fraction of the fitted models have the same (and best) fit, the model have converged and the few models with lower fit may be discarded as accidental local minima.

(4) If all models have different fit values, the model is difficult to fit (maybe too many components) and the convergence criterion has to be lowered.

(5) If the models converge to a few different but distinct fit-values, i.e. there are several models with the same fit values, then there are multiple local minima, which is a tricky situation. Likely, it is possible to circumvent this either by using some additional constraints (e.g., non-negativity) or otherwise slightly re-specifying the model.

To convert the output parameters to score and loadings matrices, we use the function *fac2let*. The loading matrices, W and V, are normalized, that is

$$\sum_{a=1}^{A} w_{ak}^2 = 1 = \sum_{s=1}^{S} v_{sk}^2$$

which is expressed by saying that "all variance is kept in the first mode Λ".

The MATLAB function *parafac* is used to fit the restricted PARAFAC model in which each component of the error $U^{(r,a,s)}$, u_{iras}, has equal variance $\sigma_{as}^2 = \sigma^2$. However, since the variances of the error are different in (M4a), we cannot directly apply the function in this toolbox. We should transform our model to a model that has equal variance as follows. If the model for (M4a) is

$$Y^{(r,a,s)} = \Lambda f^{(a,s)} + U^{(r,a,s)} \tag{4.43}$$

and $U^{(r,a,s)} \sim N_p(0, \sigma_{as}^2 I_p)$, then we re-scale the model by α_{as} with

$$\alpha_{as} \equiv \frac{\sigma_{as}^2}{\sum_{b=1}^{A}\sum_{t=1}^{S}\sigma_{bt}^2/AS} \tag{4.44}$$

Then the model in (4.43) can be transformed to

$$\widetilde{Y}^{(r,a,s)} \equiv Y^{(r,a,s)}/\sqrt{\alpha_{as}} = \Lambda \widetilde{f}^{(a,s)} + \widetilde{U}^{(r,a,s)} \tag{4.45}$$

where $\widetilde{f}^{(a,s)} \equiv f^{(a,s)}/\sqrt{\alpha_{as}}$ and $\widetilde{U}^{(r,a,s)} \equiv U^{(r,a,s)}/\sqrt{\alpha_{as}}$. Then $\widetilde{U}^{(r,a,s)} \sim N(0, \sigma^2 I_p)$ with

$$\sigma^2 = \sigma_{as}^2/\alpha_{as} = \sum_{b=1}^{A}\sum_{t=1}^{S}\sigma_{bt}^2/AS. \tag{4.46}$$

The log-likelihood function for (4.43) is

$$l(\theta) = -\frac{1}{2}\sum_{a=1}^{A}\sum_{s=1}^{S}[\, pR\log(\sigma_{as}^2) + \frac{1}{\sigma_{as}^2}\sum_{r=1}^{R}\|Y^{(r,a,s)} - \Lambda f^{(a,s)}\|^2] \tag{4.47}$$

and the log-likelihood function for (4.45) is

$$\begin{aligned} l_r(\theta) &\equiv l_{rescaled}(\theta) \\ &= -\frac{1}{2}\sum_{a=1}^{A}\sum_{s=1}^{S}[\, pR\log(\sigma^2) + \frac{1}{\sigma^2}\sum_{r=1}^{R}\|\widetilde{Y}^{(r,a,s)} - \Lambda\widetilde{f}^{(a,s)}\|^2] \end{aligned} \tag{4.48}$$

Plugging (4.46) into (4.48), we get

$$\begin{aligned} l_r(\theta) &= -\frac{1}{2}\sum_{a=1}^{A}\sum_{s=1}^{S}[\, pR\log(\sigma_{as}^2\,/\alpha_{as}) + \frac{1}{\sigma_{as}^2}\sum_{r=1}^{R}\|Y^{(r,a,s)} - \Lambda f^{(a,s)}\|^2] \\ &= l(\theta) + \frac{1}{2}\sum_{a=1}^{A}\sum_{s=1}^{S}[\, pR\log(\alpha_{as})] \end{aligned} \tag{4.49}$$

Thus, the log-likelihood function for (4.43) is $l(\theta)$, given by

$$l(\theta) = l_r(\theta) - \frac{1}{2}\sum_{a=1}^{A}\sum_{s=1}^{S}[\, pR\log(\alpha_{as})]. \tag{4.50}$$

Now we can apply the *parafac* toolbox to our data in the following steps:

The mean level μ_{as} can be consistently estimated by $\frac{1}{p}\sum_{i=1}^{p} y_{iras}$. Project the data $Y^{(r,a,s)}$ to the space orthogonal to **1**, denoted by

$$Y^{(r,a,s)*} \equiv Y^{(r,a,s)} - \frac{1}{p}\sum_{i=1}^{p} y_{iras}\mathbf{1},$$

and then take the average over the pure replications $r = 1, \ldots, R$ on $Y^{(r,a,s)*}$, to obtain

$$\bar{Y}^{(\cdot,a,s)*} = \Lambda_\star f^{(a,s)} + \bar{U}^{(a,s)} \tag{4.51}$$

Then $\bar{\mathbf{Y}} \equiv (\bar{y}_{ias})$ is a $p \times A \times S$ three-way array.

Initial input: the data array $Y^{(r,a,s)*}$ and the $A \times S$ re-scaling matrix $(\alpha_{as}^{(0)})$ with $\alpha_{as}^{(0)} \equiv 1$, for $1 \le a \le A$ and $1 \le s \le S$.

Step 1: Use the MATLAB function *parafac* in the N-way toolbox, to get estimates $(\hat{\Lambda}_\star, \hat{W}, \hat{V})$ based on $U^{(r,a,s)} \sim N_p(0, \sigma^2 I_p)$.

Step 2: Calculate

$$\hat{\sigma}_{as}^2 = \frac{1}{pR}\sum_{r=1}^{R} \|Y^{(r,a,s)*} - \hat{\Lambda}_\star\, \hat{f}^{(a,s)}\|^2 \tag{4.52}$$

where the k'th component of the vector $\hat{f}^{(a,s)}$ is given by $\hat{f}_{kas} = \hat{w}_{ak}\hat{v}_{ks}$.

Step 3: Calculate the log-likelihood function $l(\hat{\theta})$

$$l(\hat{\theta}) = -\frac{RAS}{2} - \frac{pR}{2}\sum_{a=1}^{A}\sum_{s=1}^{S}\log(\hat{\sigma}_{as}^2) \tag{4.53}$$

Step 4: Calculate the new re-scaling matrix $(\hat{\alpha}_{as}^{(1)})$

$$\hat{\alpha}_{as}^{(1)} \equiv \frac{\hat{\sigma}_{as}^2}{\sum_{b=1}^{A}\sum_{t=1}^{S}\hat{\sigma}_{bt}^2/AS} \tag{4.54}$$

and re-define

$$\hat{\alpha}_{as}^{(1)} \equiv \hat{\alpha}_{as}^{(1)} \cdot \hat{\alpha}_{as}^{(0)} \tag{4.55}$$

Step 5: Re-scale the data $Y^{(r,a,s)*}$ by $\hat{\alpha}_{as}^{(1)}$ to get

$$Y^{(r,a,s)*(1)} \equiv Y^{(r,a,s)*} / \sqrt{\hat{\alpha}_{as}^{(1)}} \tag{4.56}$$

such that $Y^{(r,a,s)*(1)}$ satisfies the following model

$$Y^{(r,a,s)*(1)} = \Lambda_{\star}^{(1)} f^{(a,s)\ (1)} + U^{(r,a,s)\ (1)} \tag{4.57}$$

and $U^{(r,a,s)\ (1)} \sim N_p(0, (\sigma^{(1)})^2\ I_p)$

Step 6: Repeat Steps 1-3 with the new $\hat{\alpha}_{as}^{(1)}$, but calculate the alternative log-likelihood function in step 3 given by

$$l^{(1)}(\hat{\theta}, \hat{\alpha}) = -\frac{RAS}{2} - \frac{pR}{2} \sum_{a=1}^{A} \sum_{s=1}^{S} \log(\hat{\sigma}_{as}^2) - \frac{pR}{2} \sum_{a=1}^{A} \sum_{s=1}^{S} \log(\hat{\alpha}_{as}^{(1)}) \tag{4.58}$$

Repeat the steps until the relative difference of the log-likelihood function on successive iterations is less than 10^{-6} and the differences in estimated parameter values are small, for example,

$$\|\theta^{(k+1)} - \theta^{(k)}\| < 10^{-3}. \tag{4.59}$$

We will apply this algorithm in the real tongue image data in the next chapter to test the hypothesis that the PARAFAC model fits.

4.5 Recommendations based on computational results

Based on our computational experience, we recommend to use the Newton type methods, such as Newton-Raphson method, quasi-Newton method, or the Broyden-Fletcher-Goldfarb-Shanno (BFGS) update, using a profile likelihood strategy. There are Splus

function *nlmin* and R function *nlm* which find a local minimum of a nonlinear function using a general Newton-Raphson method optimizer for an input Splus/R function. Based on *nlmin* or *nlm*, we wrote another Splus/R function *ProfileLik* whose input is a data set, a starting point of θ_2, a few control parameters, and whose output is the MLE $\hat{\theta}_2$, the maximized value of the profile log-likelihood, and the restricted MLE $\hat{\theta}_1(\hat{\theta}_2)$.

An advantage of the profile Newton-Raphson method is the reduction of the dimension of the parameter space. The convergence speed of Newton type methods is very fast. If we compare simply the numbers of iterations until algorithms converge, the Newton type methods would take fewest iterations. An attractive feature of the quasi-Newton method is that it automatically produces the observed information matrix.

It is often objected that the quasi-Newton methods perform poorly at the beginning of iterations. One can use the EM algorithm for the forst several iterations and then switch to quasi-Newton method (Watanable and Yamaguchi [26]). For example, one can use the values at the 300'th EM iteration as the initial input of the quasi-Newton method. Then we can get the converged values of θ_2 in the full model, denoted by $\widehat{\theta}_2^{full}$. If the converged values are in the interior of the parameter spaces and the gradient at the estimated maximum of log-likelihood is less than 10^{-6}, then the converged values are the ML estimators. If we find that the converged values of the parameters are very close to the boundary of the parameter space, that is, at least one of the component, say θ_{2j}, of the estimated values of θ_2 is very close to 0, then we consider to fit the data with the restricted model (M1R). In this situation, we first force θ_{2j} to be 0 and apply the same Newton-Raphson method with the profile likelihood strategy in the reduced model (M1R). Then we can get the converged values of θ_2 in the restricted model, denoted by $\widehat{\theta}_2^r$. Then we can follow the same steps as we previously described.

Chapter 5

Application to 2-D Coronal Tongue Surface

In this chapter, we first introduce a real data set of ultrasound cross-sectional images of the human tongue during speech. Then we apply factor analysis models (M3) and (M4a) to the tongue image data. Finally, we use the Likelihood Ratio Test (LRT) to test whether the more general models (M3) or (M4a) represent the tongue data better.

5.1 Data Set

The cross-sectional tongue surface was recorded and measured for six normal, adult, native speakers of American English (3 Caucasian females, 2 African-American males, 1 Hispanic male) by ultrasound, VCR and the μ-Tongue software package in the Vocal Tract Visualization Laboratory of M. Stone in Baltimore. Each subject attended three recording sessions and repeated the speech materials five times while ultrasound and acoustic recordings were made. Methods for the ultrasound recordings of tongue movement are discussed in detail in Stone et al.(1997). The eleven vowels sounds of English ae, ah, aw, e, eh, ih, iy, o, uh, uu, uuh, with respective phonetic symbols (æ, *a*, ɔ, e, ɛ, ɪ, i, o, ʊ, u, ʌ), were produced in ∂CVC∂ utterances (vowel sounds sandwished between consonants with "shwa" sounds ∂ as break points) using two consonant con-

texts (/s/, /l/). The coronal section was recorded in the region of the palatal vault to support the largest variation of tongue movement and shape. In the vault region there is room for upward tongue motion, and on the palatal contact the tongue will reflect its archlike shape.

The cross-sectional tongue surface for six subjects (MS, MD, SG, CS, GW, and LG) were extracted from recorded ultrasound images. Thus, we obtained 6 subjects $\times 11$ vowels $\times 2$ contexts $\times 5$ replications $\times 3$ sessions, for a total of 1980 cross-sectional tongue images. Each image curve, whatever its length along the x-dimension, is represented by 120 pairs (x, y), and different curves do not necessarily have the same range of x values. Pre-processing strategies were introduced and implemented by Slud et al. (2002), involving translation in the x and y direction, extension, padding or truncation within session, and subtracting a mean level for each speaker and sound. After preprocessing, the number of points per curve was chosen to be 101 based on the degree of padding chosen.

Let (x_{abcdi}, y_{abcdi}), for $a = 1,\ldots, 6,\ b = 1, 2, 3,\ c = 1, 2, \ldots, 22,\ d = 1, \ldots, 5,\ i = 1, \ldots 120,$ be our raw data set, where a indexes subject, b indexes session, c indexes sound/context, d indexes replications within session, and i indexes observations (points) on the image curves. After preprocessing, the final data set on a common (x, y) coordinate system based on five replicated measurements in three sessions for each of the six subjects is (x_i, y_{abcdi}), where subject is indexed by $a = 1,\ldots, 6$, session by $b = 1, 2, 3$, vowel/consonant by $c = 1, 2, \ldots, 22$, replication by $d = 1, \ldots, 5$, and observations (points) along the image curve by $i = 1, \ldots 101$.

We now focus only on the eleven vowels and six subjects and treat the two consonants as pure replications. Then the pure replications are 2 consonant contexts $\times 5$ replications $\times 3$ sessions, for a total of 30 replications. Therefore, the data

can be rewritten as (x_i, y_{iras}), where subject is indexed by $a = 1,\ldots,6$, vowel by $s = 1, 2, \ldots, 11$, pure replication by $r = 1, \ldots, 30$, and observations (points) along the image curve by $i = 1, \ldots 101$. For convenience, let A denote the total number of subjects, S denote the total number of vowels, R denote the number of replications, and p denote the number of points per curve. Then $A = 6$, $S = 11$, $R = 30$, and $p = 101$.

5.2 Application of Factor Analysis Models to Tongue Image Data

The hierarchical family of models (M2), (M3), (M4), (M4a) and (M4′) we constructed can be used on real data involving coronal cross-sectional pictures of the human tongue surface during speech. The PARAFAC model (M4a) has been used previously to analyze tongue images data but with different imaging technology (X-ray instead of ultrasound) and different cross-section (lengthwise instead of transverse to the tongue). Harshman and Lundy [8] reported that the success of a PARAFAC analysis depends on the use of adequate statistical pre-processing. Slud et al. [22] actually found that the PARAFAC (M4) modelling approach did not adequately represent the coronal tongue data. They found that the PARAFAC model did less well, the more highly cross-classified the data were. Due to the highly constrained form and inadequacy of PARAFAC, a more general model such as the 3-mode factor analysis model (T3), defined in (3.83), is needed for representing cross-classified data. The model T3 fits better than PARAFAC on some data, but it tends to use excess parameters (Zheng et al. [27]. Thus, the model hierarchy we constructed in Chapter 2 may help to rationalize the choice of models. In this section, the PARAFAC (M4a) model and a more general

model (M3) extending (T3) are applied to coronal tongue data. The likelihood ratio test (LRT) is used to test whether the more general models (M3) or (M4a) represent the coronal tongue data better. MATLAB and the N-ways toolbox are used to get the MLE in (M4a).

In (M3), the tongue image data satisfies the equation (2.50):

$$Y^{(r,a,s)} = \begin{pmatrix} y_{1ras} \\ \vdots \\ y_{pras} \end{pmatrix} = \mu_{as} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \Lambda \mathbf{f}^{(r,a,s)} + U^{(r,a,s)}$$

That is,

$$y_{iras} = \mu_{as} + \sum_{k=1}^{q} \lambda_{ik} f_{kras} + u_{iras} \tag{5.1}$$

The unknown parameter μ_{as} is the mean level of the surface measurements y_{iras} for the speaker a and vowel s.

In PARAFAC (M4a), the model we consider is

$$Y^{(r,a,s)} = \mu_{as}\, \mathbf{1} + \Lambda_\star\, \mathbf{f}^{(a,s)} + U^{(r,a,s)}$$

where $\Lambda_\star$ is a $p \times q$ matrix with non-orthogonalized columns. The fixed effect $\mathbf{f}^{(a,s)}$ can be written as

$$\mathbf{f}^{(a,s)} = \begin{pmatrix} f_{1as} \\ \vdots \\ f_{qas} \end{pmatrix} \text{ and } f_{kas} = w_{ak} v_{sk} \text{ for } k = 1, \ldots, q$$

The factor weight f_{kas} is represented in PARAFAC as the product of a vowel-independent speaker weight w_{ak} and a speaker-independent vowel weight v_{ks}.

5.2.1 Principal Component Analysis of Tongue Data

Since the coronal tongue data vector $Y^{(r,a,s)}$ is in a high dimensional space $\mathbf{R}^{101}$, it is a good idea to reduce dimension before we analyze the data. Using principal compo-

nent analysis, we can project the coronal tongue data from p dimensions down to m_0 dimensions. That is,

$$Y^{(r,a,s)} \to X^{(r,a,s)} = L^t Y^{(r,a,s)} \tag{5.2}$$

where L is a $p \times m_0$ loading matrix orthogonal to $\mathbf{1}$ and the columns of L are the first m_0 eigenvectors corresponding to the first m_0 largest eigenvalues of the covariance matrix of $Y^{(r,a,s)} - (\mathbf{1}^t \ \frac{1}{p} \ \bar{Y}^{(\cdot,a,s)}) \ \mathbf{1}$. We now determine the number m_0 so that it will retain most of the data information after the dimension reduction by the PCA.

Let us consider the ratio

$$R(m) \equiv \frac{\sum_{k=m+1}^{p} \lambda_k}{\sum_{k=1}^{p} \lambda_k} \tag{5.3}$$

where $\lambda_1, \ldots, \lambda_p$ are eigenvalues of the covariance of the Coronal tongue data $Y^{(r,a,s)}$. The values of $100 \cdot (1 - R(m))$ are the percent of the total sum of squares for ordinate values. It can be used to determine the number q of principal components to retain in describing data as we described in (1.4). The percentage of the cumulated variance accounted for by the successive PC's are: 69.377%, 90.391%, 96.617%, 98.863%, 99.553%, 99.830%, 99.933%, 99.975%, 99.990%, 99.996%. So we simply choose $m_0 = 10$.

We can also determine the minimum number m_0 of m such that $-\log R(m)$ exceeds the threshold 7 to retain 99.9% of the data information. That is,

$$m_0 = \min\{m \in \mathbf{N} : -\log R(m) > 7\} \tag{5.4}$$

Figure 5.1 shows the graph that $-\log R(m)$ against m. We see that $m_0 = 7$ was good enough to retain 99.9% of the data information, but we simply chose $m_0 = 10$ and projected the coronal tongue data $Y^{(r,a,s)}$ from 101 dimensions down to 10 dimensions.

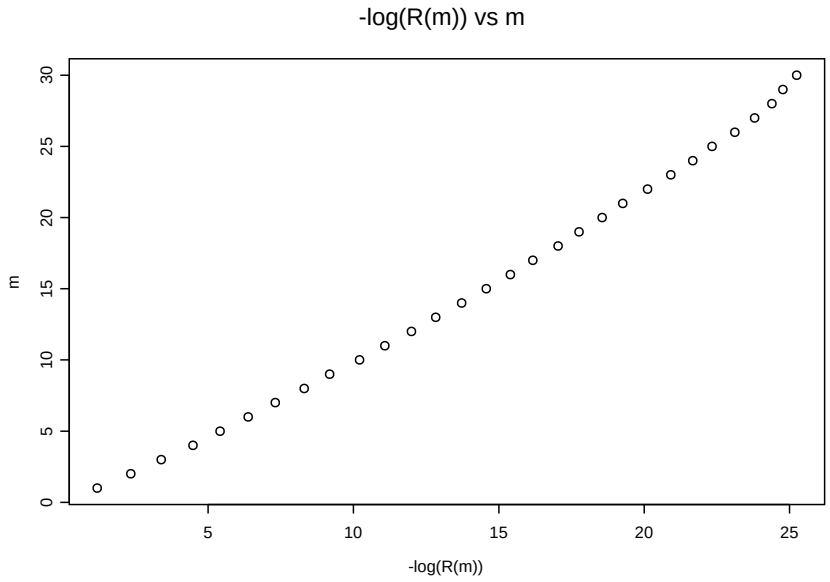

Figure 5.1: Graph of $-\log R(m)$ against m for the coronal tongue data.

5.2.2 Test of the Hypothesis that the PARAFAC Model Fits

We use the Likelihood Ratio Test (LRT) to test which model fits the tongue data better. The null hypothesis is $H_0 : \theta \in \Theta_{M4a}$, against alternative $H_1 : \theta \in \Theta_{M3}$, where Θ_{M3} and Θ_{M4a} are defined in (2.56) and (2.69), respectively.

Using the Newton-Raphson method based on a profile likelihood strategy, we find the maximum log-likelihood in (M3) is $l(\widehat{\theta})_{M3} = -21677.48$. Using the MATLAB function *parafac* in the N-way toolbox together with the algorithm we constructed in Chapter 4, we find the maximized log-likelihood in (M4a) is $l(\widehat{\theta})_{M4a} = -22570$. The likelihood ratio statistic

$$-2\log\lambda = 2(l(\widehat{\theta})_{M3} - l(\widehat{\theta})_{M4a}) = 1785.04 \tag{5.5}$$

Let $dim(\Theta)$ denote the dimension of the parameter space Θ. Since it is impossible for a subject to speak a sound always exactly the same way, always $\sigma^2_{as} > 0$ in the real tongue data. Thus, the true parameter is in the interior of the parameter space in both

PARAFAC and (M3). By Theorem 3.4, under suitable regularity conditions, for each $\theta \in \Theta_0$,

$$-2\log\lambda \to \chi^2_{d-r} \text{ when } R \to \infty$$

where

$$\begin{aligned} d-r &= dim(\Theta_{M3}) - dim(\Theta_{M4a}) \\ &= qAS - (Aq - q + Sq - q) = 102 \end{aligned} \tag{5.6}$$

By Remark 3.5, $\sqrt{2\chi^2_{102}} \overset{\mathcal{D}}{\approx} N(\sqrt{203},\ 1)$. Let $X \equiv -2\log\lambda(y)$ and let $Z \sim N(0,1)$. The rejection region is

$$\begin{aligned} R &= \{\lambda(y) < c\} = \{X > -2\log c\} \\ &= \{\sqrt{2X} > \sqrt{-4\log c}\,\} \\ &= \{\sqrt{2X} - \sqrt{203} > \sqrt{-4\log c} - \sqrt{203}\} \\ &= \{Z > \sqrt{-4\log c} - \sqrt{203}\} \end{aligned} \tag{5.7}$$

Then $\sqrt{-4\log\lambda(y)} - \sqrt{203} = \sqrt{2 \cdot 1785.04} - \sqrt{203} = 45.50219$. Since this is represents a very extreme quantile for $N(0,1)$, we reject the null hypothesis. Therefore, the (M3) model fits the coronal tongue data better than the PARAFAC (M4a) model.

5.2.3 Comparison of fitted loading matrices among (M3), (M4a), and PCA

We chose $q = 2$ and used the Newton-Raphson method based on a profile likelihood strategy and the MATLAB function *parafac* to get the estimated Λ in model (M3)

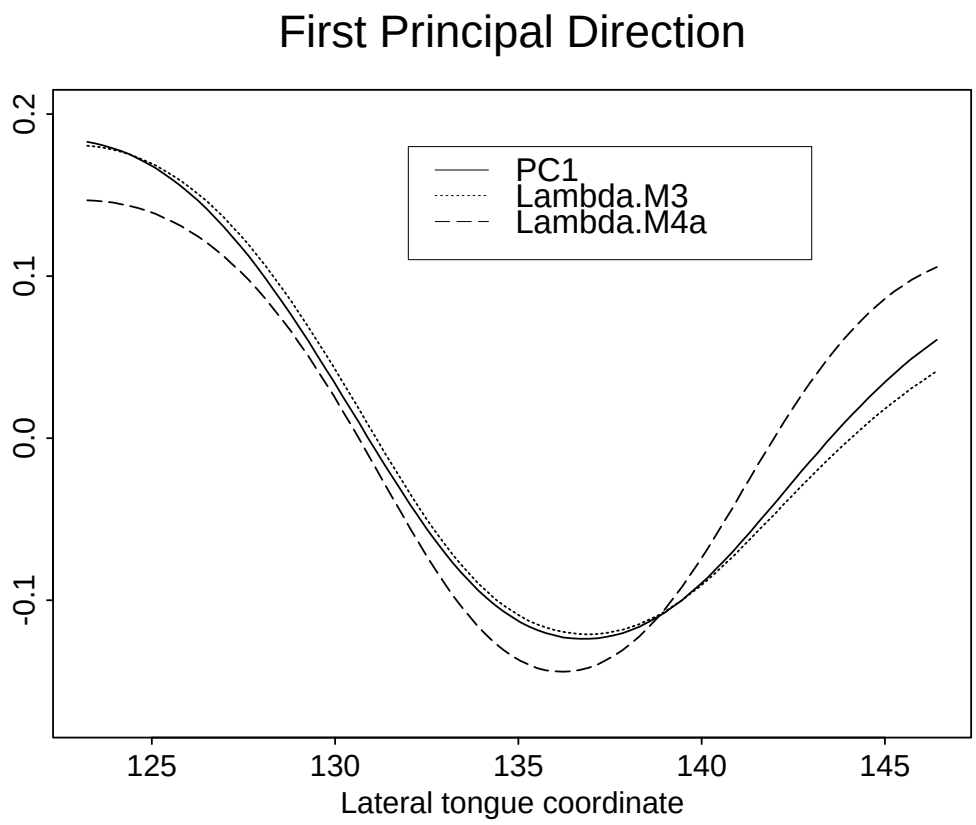

Figure 5.2: First Principal Direction for coronal tongue data based on (PCA), (M3) and (M4a).

and the PARAFAC model (M4a), respectively. The first column of $\widehat{\Lambda}$ is called the first Principal Direction and the second column of $\widehat{\Lambda}$ is called the second Principal Direction. We also get the first two Principal Directions, denoted by PC1 and PC2, based on the Principal Component Analysis (PCA) or equivalently by model (M1). Figure 5.2 shows the curves of the first Principal Directions based on PCA, the model (M3), and the PARAFAC model (M4a). Since we knew from the LRT in Section 5.2.2 that (M3) fits the data better than (M4a), we think that the first principal direction (dotted line) based on the model (M3) in Figure 5.2 should represent the data better than the principal direction (dashed line) based on PARAFAC (M4a). By Slud et al. [22], the percent of variance (after subtraction of curve mean) accounted for by the two

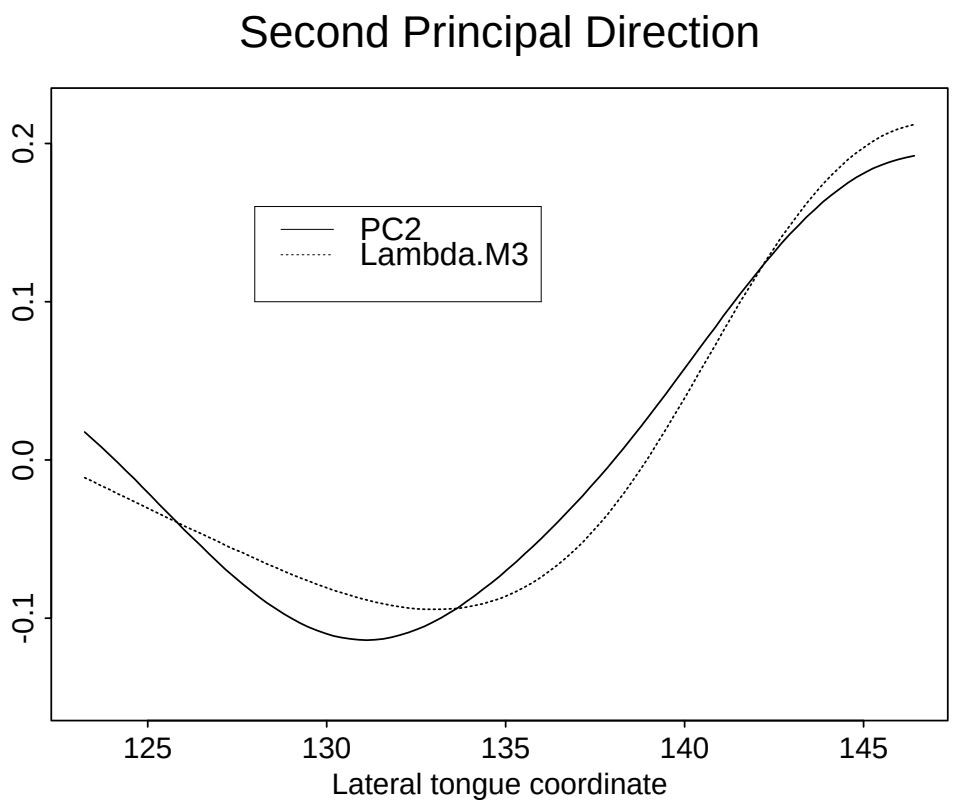

Figure 5.3: Second Principal Direction for coronal tongue data based on (PCA) and (M3).

PCs was 69.4 and 21.0, respectively. Thus, we think that PC1 plays the more important role in describing the data than PC2. In Figure 5.2, the first principal direction (dotted line) based on the model (M3) is very close to PC1 (solid line) and the dash line based on the PARAFAC model is far from the other two curves. So this indicates that the PARAFAC model did not adequately represent the data. Thus, the result in Figure 5.2 agrees with the result of LRT. The second Principal Direction can be compared in Figure 5.3.

5.2.4 Identification of vowels and subjects

The values in Table 5.1 are $\widehat{\alpha}_{as}^{(k)}$, the estimated values of the scaled parameters α_{as} at the 10'th iteration convergence based on the MATLAB function *parafac* and the algorithm we constructed in Section 4.4. The values of $\widehat{\alpha}_{as}^{(k)}$ are very stable up to the 3rd digital place after the 8'th iteration. For a specific subject a and vowel s, the value $\widehat{\alpha}_{as}^{(k)}$ can be viewed as the variance of speaker a and vowel s relative to all of the subjects and vowels. Based on these estimated values of α_{as} in Table 5.1, we can try to distinguish particular vowels or subjects. For example, we found that the vowel "iy" has very large $\widehat{\alpha}_{as}$ values for most of the subjects. The vowel "uu" could be a vowel sound also having larger $\widehat{\alpha}_{as}$ values. Now, let us focus on the subjects. We found that the subject "C.S." tends to speak vowels consistently (ie, with relatively small variance).

The other information such as $\widehat{\Lambda}$ and $\widehat{\sigma}^2$ is listed in Appendix B.

					k=10	
	M.S.	M.D.	S.G.	C.S.	G.W.	L.G.
ae	1.033	0.955	0.793	0.788	1.161	0.443
ah	0.583	0.663	1.073	0.901	0.843	0.996
aw	0.683	0.370	0.922	0.703	0.554	1.312
e	1.421	1.418	0.832	0.400	1.543	1.307
eh	0.738	0.962	1.002	0.650	1.199	0.691
ih	1.828	1.003	0.920	0.735	1.119	0.498
iy	1.147	3.082	2.784	0.304	3.615	2.241
o	0.548	0.530	0.767	0.634	0.804	1.092
uh	0.593	0.372	1.081	0.714	0.539	1.218
uu	2.096	0.976	1.569	0.722	1.110	0.677
uuh	0.480	0.480	1.240	0.655	0.643	1.564

Table 5.1: The estimated values of the scaled parameters α_{as}.

Chapter 6

Summary and Future Work

We have constructed a new model hierarchy related to Factor Analysis, in which vector measurements are linearly decomposed into a relatively small set of hypothetical principal directions, for purposes of dimension reduction. A hierarchical family of cross-classified factor models has been built for the application to a real tongue data set. We unified the mathematical specification of unknown parameters in the models and established that in the right parameterizations, the unknown parameters were uniquely identifiable from the data. We found some new results related to non-identifiable models and parameter values in the boundary of the parameter space: There exists a solution of $\Sigma_Y = \Lambda\Lambda^t + diag(\psi)$ on the boundary of the parameter space Θ^* when the model is non-identifiable.

We found and implemented computationally effective maximum likelihood estimators for the unknown parameters using the Newton-Raphson method with a profile likelihood strategy. This method is much faster computationally since the dimension is sharply reduced. It is also very effective since the MLE can be always obtained in our simulated data samples while the EM algorithm converges extremely slowly or sometimes does not converge. We found the MLE from the profile likelihood method and the converged values of the EM algorithm agree if the EM algorithm converges.

We found a condition combined with the restricted model (M1R) to check whether the converged point on the boundary of the parameter space is the MLE.

We ultimately established statistical tests of goodness of fit of the models to data. In this research, we only focused on testing the fit of the PARAFAC model against (M3) and built the Likelihood Ratio Test (LRT). In (M3), we maximized the log-likelihood using the Newton-Raphson method with profile likelihood strategy. In the PARAFAC model (M4a), we used the MATLAB function *parafac* and established a two-step profile likelihood algorithm to transform our model to be compatible with the *parafac* function. The algorithm we constructed starting from the MATLAB toolbox is extremely efficient. The speed of convergence is very fast. The N-way toolbox can also be used to get the MLE for (M4) or (T3).

We applied the LRT to a real data set involving coronal cross-sectional pictures of the human tongue surface during speech. We found that the PARAFAC model (M4a) is inadequate to represent the data. The more general model (M3) fits the coronal tongue data better than the PARAFAC model.

In the next stage of work, we will focus on the following. First, we will test the inadequacy of (M4) and check whether (T3) is adequate. This part should be easy to test since the N-way toolbox provides the option to add the constraint on Λ to have orthogonal columns. Also the N-way toolbox contains a function to fit the model (T3).

Second, we would like to test the adequacy of (M1) and (M2). We know that the model (M3) is a very general fixed effect cross-classified factor model and we found that (M3) fits a coronal tongue data set better than (M4a), but we don't know whether (M3) is adequate to present the data. It is possible that (M3) is also inadequate for this coronal tongue data and a more general random effect cross-classified factor model, such as (M2), might fit the data better. However, (M3) is not nested in (M2) since

one is fixed effect and the other is random effect. Since we found in Figure 5.2 that the first Principal Direction in (M3) is very close to PC1 based on PCA and the PC1 can be interpreted as the first Principal Direction in (M3) by Lemma 3.1, we can test the goodness of fit for (M1) against (M2) instead of (M3). Third, we did not prove the convergence of the alternating algorithms in Section 4.4.1, and we intend to do so. Finally, we want to apply our research to real sagittal tongue data. Since there are only five replications in the tongue data, we might need to consider bootstrapping strategy to deal with estimation of variability.

Appendix A

Matrix Algebra

Theorem A.1. *(Graybill [7], p 88) Let A be a $p \times q$ matrix. Then the null space of A^t is the orthogonal component of the column space of A. That is,*

$$NS(A) = \{v \in V :< v, w >= 0 \text{ for all } w \in col(A)\}$$

where $col(A)$ denotes the column space of A.

Lemma A.2. *(Singular value decomposition theorem)[15] If A is an $p \times q$ matrix of rank r, then A can be written as*

$$A = UDV^t \tag{A.1}$$

where U ($p \times r$) and V ($q \times r$) are column orthonormal matrices ($U^tU = V^tV = I_r$) and D is a $r \times r$ diagonal matrix with positive elements.

Lemma A.3. *[1] Given a positive definite symmetric matrix $A \in \mathbf{R}^{p \times p}$, there is a uniquely determined orthogonal matrix U (except for possible changes of sign of the columns) such that U^tAU is diagonal with diagonal elements arranged in non-increasing order.*

Lemma A.4. *(Jennrich's Basic Uniqueness Theorem [10])*
If $\sum_l U_{il}V_{jl}W_{kl} = \sum_l U^*_{il}V^*_{jl}W^*_{kl}$ and if the respectively $I \times L$, $J \times L$, and $K \times L$,

matrices U, V, W each have rank $L \leq I, J, K$, then

$$U^* = URD_1, V^* = VRD_2, W^* = WRD_3 \tag{A.2}$$

where R is a permutation matrix and D_1, D_2, and D_3 are diagonal matrices with $D_1 D_2 D_3 = I$.

Lemma A.5. *(Graybill [7], p 266) Let A be a $k \times k$ symmetric matrix of independent real variables (subject only to $a_{ij} = a_{ji}$); then*

$$\frac{\partial |A|}{\partial A} = 2[A_{ij}] - D_{[A_{ij}]}$$

where A_{ij} is the cofactor of a_{ij} and $D_{[A_{ij}]}$ is a diagonal matrix with i'th diagonal element equal to A_{ii}, the cofactor of a_{ii}.

Lemma A.6. *(Graybill [7], p 267) Let A be a $k \times k$ symmetric nonsingular matrix of independent real variables (subject only to $a_{ij} = a_{ji}$); then*

$$\frac{\partial (\log |A|)}{\partial A} = 2A^{-1} - D_{A^{-1}}$$

where $D_{A^{-1}}$ is a diagonal matrix with i'th diagonal element equal to that of A^{-1}.

Appendix B

Technical Appendix

B.1 Computational results on simulated data

In Section 4.3, we discussed computational results based on simulated data. We list the values of (Λ_0, ψ_0) and (Λ_1, ψ_1) for Cases (A)-(D) and (O)-(R) in the following tables. The notation $\Lambda_i^{(k)}$ denotes the k'th column of Λ_i, for $i = 0, 1$.

The reason for choosing the values of Λ_0 listed in Table B.1 was to construct a Λ_0 satisfying the conditions in Θ_{M0a2}: orthogonal columns, column norms in decreasing order, and $\Lambda_0^t \mathbf{1} = 0$. Then the parameter (Λ_0, ψ_0) is identifiable in model (M0a) in the case $\mu = 0$. Starting from a 6×2 matrix Λ_{00} with the first column $\Lambda_{00}^{(1)} = (3, 2, 1, -1, -2, -3)$ and the second column $\Lambda_{00}^{(2)} = (1, 2, -3, -1, 3, -2)$, then $\Lambda_{00}^t \mathbf{1} = 0$, but Λ_{00} does not have orthogonal columns. So we used the Gram-Schmidt orthogonalization process to get Λ_0 in Table B.1 which satisfies the conditions in Θ_{M0a2}.

In Cases (E)-(N), we choose $s \in (0, 1)$, let $\Lambda_s = (1 - s) \cdot \Lambda_0 + s \cdot \Lambda_1$ be the convex combination between Λ_0 and Λ_1 and fix the entries of ψ_{s0} as independent $Unif([0, 0.5])$ variates, simulated as $\psi_{s0} = (0.11654,\ 0.37053,\ 0.05444,\ 0.46252,\ 0.00746,\ 0.44479)$. We generated a data sample, called Case (E), based on the pa-

rameter (Λ_s, ψ_{s0}) with $s = 1/2$; generated a data sample, called Case (F), based on the parameter (Λ_s, ψ_{s0}) with $s = 2/3$; generated 3 data samples, called Cases (G)-(I), based on the parameter (Λ_s, ψ_{s0}) with $s = 5/6$; and generated 5 data samples, called Cases (J)-(N), based on the parameter (Λ_s, ψ_{s0}) with $s = 11/12$.

B.2 Computational result on coronal tongue data

The Sum of Squares of residuals (SSR) at 10'th iteration is 1300.148474. Thus, the σ^2, defined in (4.46), is $\sigma^2 = SSR/(pR) = 1300.148474/(10 * 30) = 4.333828$.

The values in Table B.3 are the ML estimates of Λ in model (M3) and (M4a), and the first two principal Directions from PCA.

(A)-(D)	(A)-(D)	(A)	(B)	(C)	(D)
$\Lambda_0^{(1)}$	$\Lambda_0^{(2)}$	ψ_0	ψ_0	ψ_0	ψ_0
3	0.46429	0.22186	0.46202	0.08430	0.02967
2	1.64286	0.33539	0.11055	0.31064	0.40038
1	-3.17857	0.29213	0.18112	0.49082	0.30020
-1	-0.82143	0.09462	0.21458	0.03677	0.46742
-2	3.35714	0.36865	0.33569	0.24644	0.01616
-3	-1.46429	0.33451	0.24978	0.03667	0.39166

Table B.1: The simulated values of the first two columns of Λ_0 and ψ_0 in cases (A)-(D)

(O)-(R)	(O)-(R)	(O)	(P)	(Q)	(R)
$\Lambda_1^{(1)}$	$\Lambda_1^{(2)}$	ψ_1	ψ_1	ψ_1	ψ_1
2	0	0.29273	0.47350	0.42129	0.29987
0	1	0.03241	0.33326	0.45984	0.04459
0	0	0.27562	0.25244	0.46000	0.11360
0	0	0.46130	0.20707	0.10466	0.09580
0	0	0.34056	0.12609	0.15538	0.30627
0	0	0.14803	0.04231	0.20651	0.35284

Table B.2: The simulated values of the first two columns of Λ_0 and ψ_0 in cases (O)-(R)

PC1	$\widehat{\Lambda}^{(1)}_{M3}$	$\widehat{\Lambda}^{(1)}_{M4a}$	PC2	$\widehat{\Lambda}^{(2)}_{M3}$	$\widehat{\Lambda}^{(2)}_{M4a}$
1.000	0.997	0.959	0.000	0.076	0.902
0.000	-0.076	0.251	1.000	0.981	-0.410
0.000	0.000	0.133	0.000	0.170	-0.131
0.000	-0.019	-0.001	0.000	0.054	-0.037
0.000	-0.002	-0.008	0.000	0.008	0.000
0.000	0.003	0.002	0.000	0.005	0.006
0.000	0.002	0.002	0.000	-0.008	0.001
0.000	0.000	0.000	0.000	-0.003	0.000
0.000	-0.001	-0.001	0.000	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.000

Table B.3: The MLEs of Λ in model (M3) and (M4a), and the first two principal directions from PCA. $\widehat{\Lambda}^{(k)}_{M3}$ denotes the k'th column of the MLE of Λ in model (M3), and $\widehat{\Lambda}^{(k)}_{M4a}$ denotes the k'th column of the MLE of Λ in model (M4a)

BIBLIOGRAPHY

[1] Anderson, T.W. and Rubin, H.: *" Statistical Inference in Factor analysis"*, Proc. of the Third Berkeley Symposium, Vol. 5 (1956), 111-150.

[2] Anderson, T. W., An Introduction to Multivariate Statistical Analysis, John Wiley & Sons, Inc., New York Chichester Brisbane Toronto Singapore, 1984.

[3] Bro, R., The N-way on-line course on PARAFAC and PLS, http://www.models.kvl.dk/courses/; 1998-2002.

[4] Cheng, Y., University of Maryland: Maximum Likelihood Estimation and Computation in a Random Effect Factor Medel, Statistics Program, College Park, 2004.

[5] Cox, D. R., and Hinkley, D. V., Theoretical Statistics, Chapman and Hall: London, 1974.

[6] Dempster, A. P., Laird, N. M. and Rubin, D. B.: *"Maximum likelihood from incomplete data via the EM algorithm"*, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1(1977), 1-38.

[7] Graybill, F. A.:Introduction to Matrices with Applications in Statistics, Wadsworth Publishing Company, Inc., Belmont, California, 1969.

[8] Harshman, R.; Lundy, M.: ”The PARAFAC model for the three-way factor analysis and multidimensional scaling,” in Research Methods of Multimode Data Analysis, edited by H. G. Law, C. W. Snyder, J. A. Hattie, and R. P. MacDonald, Praeger, New York (1984), 122-215.

[9] Harshman, R.; Ladefoged, P.; Goldstein, L.: *“Factor analysis of tongue shapes”*, J. acoust. Soc. Am. 62 (1977), 693-707.

[10] Harshman, R.: *“Foundations of the PARAFAC procedure: Models and conditions for an explanatory multimodal factor analysis”*, UCLA Working Papers in Phonetics, Vol. 16 (1970), 1-84.

[11] Jamshidian, M., and Jennrich, R.I.: *”Acceleration of the EM algorithm by using quasi-Newton methods”*, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 59 (1997), 563-587.

[12] Lehmann, E. L.: Testing Statistical Hypotheses. New York, Wiley, 1959.

[13] Lindstrom, M., Bates, D. M.: *”Newton-Raphson and EM algorithm for linear mixed-effects models for repeated-measures data”*, Journal of American Statistical Association. Vol. 83 (1988), 1014-1022.

[14] Magnus, J.; and Neudecker, H.: Matrix Differential Calculus with Applications in Statistics and Econometrics, 1988.

[15] Mardia, K.V.; Kent, J.T.; Bibby, J.M.: Multivariate Analysis, Academic Press, Harcourt Brace & Company, Publishers, New York Boston Sydney, 1997.

[16] Neyman, J.: *”Optimal asymptotic tests of composite hypotheses”*, In V. Grenander, editor, Probability and Statistics: *The Harold Cramér Volume*, 213-234. Wiley, New York, 1959.

[17] Neyman, J., and Scott, E.: *"On the use of $C(\alpha)$ optimal tests of composite hypotheses"*, Bulletin of the International Statistical Institute, Proceeding of the 35th Session, 41 (1966), 477-497.

[18] Rubin, D., and Thayer, D.: *EM algorithms for ML factor analysis*, Psychometrika, 47 (1982), 69-76.

[19] Stone, M.; Goldstein, M.; and Zhang, Y.: *"Principal component analysis of cross sections of tongue shapes in vowel production"*. Speech Communication 22 (1996), 3728-3737.

[20] Self, S.; and Liang, K.: *"Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions"*. Journal of the American Statistical Association 82 (1987), 605-610.

[21] Silvey, S. D., Statistical Inference. Penguin, Baltimore, 1970.

[22] Slud, E.; Stone, M.; Smith, P.; and Goldstein, M.: *"Principal components representation of the two-dimensional coronal tongue surface"*. Phonetica 59 (2002), 108-133.

[23] Stone, M.; Goldstein, M.; Zhang, Y.: *"Principal components analysis of cross sections of the tongue shapes in vowel production"*. Speech Commun. 22 (1997), 173-184.

[24] Tipping, M.; and Bishop, C.: *"Probabilistic Principal Component Analysis"*. Journal of the Royal Statistical Society, Series B, 61 (1998), 611-622.

[25] Tucker, L.: *"Some mathematical notes on three-mode factor analysis"*. Psychometrika 31 (1966), 279-279.

[26] Watanable, M. and Yamaguchi, K.: The EM Algorithm and Related Statistical Models, Marcel Dekker, Inc., New York, Basel, 2004.

[27] Zheng, Y.; Hasegawa-Johnson, M.; and Pizza, S.: *"Analysis of the three-dimensional tongue shape using a three-index factor analysis model"*. J. Acoust. Soc. Am. 113(1) (2003), 478-486.