ABSTRACT

Title of dissertation: MAXIMUM LIKELIHOOD ESTIMATION AND COMPUTATION IN A RANDOM EFFECT FACTOR MODEL

Yang Cheng, Doctor of Philosophy, 2004

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

We first briefly review some multivariate statistical models such as Principal Component Analysis (PCA), Factor Analysis (FA), and Probabilistic PCA (PPCA). Alternatively, we approach PCA from the least-squares point of view. We introduce a Random Effect Factor Model I (REFM_1), which expresses the observed vectors up to random errors as a linear combination of a relatively small number of axis directions in a new coordinate system with random effect coefficients. Then, we characterize the maximum likelihood estimators (MLE) under REFM_1 by a profile likelihood method, that is, by maximizing the likelihood over mean and variance parameters θ_1 first with the coordinate direction parameters component θ_2 fixed, we have a restricted MLE $\hat{\underline{a}},\ \hat{B},\ \hat{\sigma}^2$ in terms of the factor directions, and substituting the estimates $\hat{\theta}_1(\theta_2)$ into the likelihood, finally maximizing the profile likelihood over the factor directions θ_2. We show that the maximizer of the profile likelihood function $l_p(\theta_2)$ over the factor directions combined with the restricted MLE for other parameters when the factor directions are fixed is the joint MLE of the

likelihood function. Some asymptotic properties of the MLE such as consistency and asymptotic normal distribution are established.

In order to analyze the multivariate data from s groups ($s > 1$), we briefly review the Common Principal Components (CPC) model. Other Random Effect Factor Models are introduced. The model REFM_2 assumes all s groups have a common factor space but differing mean and variance parameters for factor loadings and error terms, and REFM_3 is a new model which has not only a common factor space but also an additional individual space belonging to each group only. We discuss the identifiability of parameters, and again use the profile likelihood method to find the MLE.

We develop an EM algorithm to compute the MLE for REFM_1, and indicate extensions of the algorithm to REFM_2 and REFM_3. The performance of the algorithm on simulated data is described. Quasi-Newton methods are also used to calculate the MLE of the profile likelihood $l_p(\theta_2)$ and they yield the same results as the EM algorithm. Finally, we apply the EM algorithm for REFM_1 estimation to a real data set on ultrasound cross-sectional images of the tongue during speech.

MAXIMUM LIKELIHOOD ESTIMATION AND COMPUTATION IN A RANDOM EFFECT FACTOR MODEL

by

Yang Cheng

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
2004

Advisory Commmittee:

Dr. Eric Slud, Chair/Advisor
Dr. Paul Smith
Dr. Benjamin Kedem
Dr. Jian-Guo Liu
Dr. Patricia Campbell

© Copyright by
Yang Cheng
2004

ACKNOWLEDGMENTS

The completion of this dissertation would not have been possible without the help and support of many wonderful people to whom I am indebted.

I would like to express my deepest gratitude in the strongest possible terms to my advisor, Dr. Eric V. Slud, who is not only a superb advisor but also a thoughtful, sensitive and generous friend; for his overall guidance, support, and encouragement. His numerous creative ideas and constructive suggestions have made this work more interesting and exciting.

Next, I would like to express my sincere thanks to Dr. Paul Smith for much useful advice, many helpful suggestions and casual chats. I was never afraid to knock on his office door. No less appreciation is due to Dr. Patricia Campbell. She supported my early graduate study through the Project IMPACT. Many experiences that I gained while working with her on the Project IMPACT have been beneficial in my current career. Dr. Campbell stepped in at the last minute to serve on this thesis committee. I am extremely indebted to her for everything she has done for me. Also, I want to thank Dr. Benjamin Kedem and Dr. Jian-Guo Liu for serving on my thesis committee, and for the time they have taken out of their lives to assist me with my thesis.

Thanks to my wonderful family for being patient and providing continuous support, especially my parents. I owe them much for their unlimited love, encour-

agement and support. My wife, Dr. Bei Wang, spent a lot of time during the last few months helping with typing and proofreading many chapters of this dissertation. My beautiful daughters, Victoria and Virginia, who once wrote me a poem using the word "Dad - 'Doctor's Goal, Almost There, Don't Give Up' ", were confident and very proud of their Daddy working on his Doctor's goal.

In addition, I would like to thank my cousin Raymond Chin for proofreading this paper. Thanks are also due to my colleagues Ross Bailey and Bob Shaw at USPS for their support and encouragement. And thanks go to my friends Dr. Zhihui Tang and Shuigen Xiao, for providing software support.

TABLE OF CONTENTS

Chapter 1

Introduction

1.1 Background

Much of this research was motivated by the analysis of real data from NIH Project Grant R01 DC 01758, a research project on ultrasound imaging of human tongue during speech, with Dr. Maureen Stone as principal investigator. The primary problem we face is how to construct a statistical model for the coronal tongue contours, two-dimensional cross-sectional curves representing the surface of the tongue during speech. These curves, recorded in discretized form as large vectors, are very noisy, high dimensional and lacking in fixed landmarks.

Our clear mathematical goal is to find a reduced data representation of p-dimensional random curves preserving shapes and the relationships among curves, subjects, and sounds. One initial approach focused on building a smooth model. We explored many methods and models such as curve estimates, spline smoothing, functional data analysis, projection pursuit, nonparametric regression, wavelet analysis, and Principal Component Analysis. In the early studies of this thesis, we have adapted the work done by Silverman (1996), Wahba (1990) and Ramsay, and considered an underlying nonparametric smooth model:

$$y_i = f(x_i) + \epsilon_i, \qquad i = 1,\ 2,\ \ldots,\ n, \tag{1.1}$$

where $\epsilon = (\epsilon_1,\ \epsilon_2,\ \ldots,\ \epsilon_n)^t \sim N_n(0,\ \sigma^2 I_n)$ and f is some smooth function. If f is unknown, but a fixed function, then we can estimate the smooth curve f by

minimizing a penalized sum of squares. If the smooth function f is a stochastic process, then we can predict the value of $f(x)$ by a minimum variance unbiased linear (MVUL) predictor.

The tongue data are heavily cross-classified, by multiple subjects, multiple sessions and multiple sounds, and thus falls unavoidably into multiple groups. Very little work has been done on formal likelihood-based methodology for Principal Components (PC) or factor models falling into several groups. Our study is derived from Flury (1984, 1988). He claims that his Common Principal Components is the first multivariate model which is especially designed for data with multiple subjects. He also claims that his book offers a little bit of everything for the multivariate statistician: elaborate mathematics, interesting applications, and challenging computational problems. Flury's work on numerical methods was particularly extensive. Unfortunately, Flury was killed in a tragic accident on July 6, 1999 when he was hit by a falling boulder while hiking in the Dolomite Mountains near Trento, Italy. Since then, the work on Common Principal Components has not progressed.

Flury's work on Common Principal Components is highly motivating for this thesis. Of particular interest are his multivariate models, regarded as factor models, because Principal Components and Factor Analysis are highly related. After many exploratory studies on multivariate statistical models, we discovered the Random Effect Factor Model, which is well suited for a multiple subject data set. Before we introduce the Random Effect Factor Model, we briefly review some multivariate statistical models such as Principal Component Analysis (PCA), Factor Analysis (FA), and Probabilistic PCA (PPCA) in the following sections of this chapter.

1.2 Principal Component Analysis

1.2.1 Introduction

Principal Component Analysis (PCA) (Jolliffe 1986) is a classical well established multivariate technique for data dimensionality reduction. A chapter on the subject along with the analyses of covariance and correlation structures may be found in numerous textbooks on multivariate analysis. The most common derivation of PCA is in terms of a standardized linear projection which maximizes the variance in the projected space (Hotelling 1933). The mathematical treatment of PCA is based on characteristic roots and vectors of positive definite symmetric matrices. PCA is a one-group method (Flury 1988), and it is somewhat surprising that no generalizations to the case of several groups have appeared in the statistical literature until recently. There are many PCA applications, which include data compression, image processing, visualization, exploratory data analysis, pattern recognition, and time series prediction.

Principal Component Analysis (PCA) can be looked at from three different points of view (Flury 1988):

1. It is a method of transforming correlated variables into uncorrelated ones.

2. It is a method for finding linear combinations with relatively large or relatively small variability.

3. It is a tool for data reduction.

1.2.2 Definition

Suppose the random vector Y of p components has the covariance matrix Σ. Without loss of generality, we assume for simplicity that the mean of Y is 0; otherwise we would subtract from it (an estimate based on data of) the constant mean $E(Y)$. Actually, in doing Principal Component Analysis, we are interested only in variances and covariances. Moreover, in developing the ideas and algebra here, the actual distribution of Y is irrelevant except for the covariance matrix. If Y has a normal distribution, then more meaning can be given to the principal components.

Let c be a p-component unit vector, that is, a vector with $c^t c = 1$. The main operation of PCA is to find $c \in \mathbf{R}^p$ which maximizes the variance of $c^t Y$, i.e.,

$$c = \arg\max \mathrm{Var}(c^t Y). \tag{1.2}$$

Let

$$f(c, \lambda) = \mathrm{Var}(c^t Y) - \lambda(c^t c - 1) \;=\; c^t \Sigma c - \lambda(c^t c - 1), \tag{1.3}$$

where λ is a constant *Lagrange multiplier*. A vector c maximizing $f(c, \lambda)$ must satisfy the following equation:

$$\nabla_c f(c, \lambda) \;=\; 2\Sigma c - 2\lambda c = 0. \tag{1.4}$$

That is,

$$\Sigma c = \lambda c \tag{1.5}$$

which means that λ is an eigenvalue of Σ and c is a corresponding eigenvector

of Σ. Indeed, λ must be the largest eigenvalue, $\lambda_{\max}$, since

$$\max_{c:\, c^t c=1} \mathrm{Var}(c^t Y) \;=\; \max_{c:\, c^t c=1} c^t \Sigma c \;=\; \max_{c:\, c^t c=1} \lambda c^t c \;=\; \max_{c^t c=1} \lambda \;\equiv\; \lambda_{max}. \tag{1.6}$$

Let c_1 be the eigenvector of Σ corresponding to λ_{max}. Then $c_1^t Y$ is called the *first principal component*. Now let us find a normalized linear combination $c^t Y$ with maximum variance among all linear combinations of components of Y which are uncorrelated with the first principal component. Lack of correlation means that

$$0 \;=\; E(c^t Y Y^t c_1) \;=\; c^t \,\Sigma\, c_1 \;=\; \lambda_{max}\, c^t c_1, \tag{1.7}$$

or equivalently, $c^t c_1 = 0$. We now want to maximize

$$f(c, \lambda, \mu) \;=\; c^t \,\Sigma\, c \;-\; \lambda\,(c^t c - 1) - 2\mu c^t c_1, \tag{1.8}$$

subject to the unit-norm and orthogonality constraints, where λ and μ are Lagrange multipliers. The vector of partial derivatives is

$$\nabla_c f(c, \lambda, \mu) \;=\; 2\Sigma c \;-\; 2\lambda c \;-\; 2\mu c_1, \tag{1.9}$$

which we set equal to 0. After multiplying on the left by c_1^t, we obtain

$$0 \;=\; 2c_1^t \,\Sigma\, c \;-\; 2\,\lambda\, c_1^t c \;-\; 2\mu c_1^t c \;=\; -2\mu. \tag{1.10}$$

Therefore, $\mu = 0$, and c and λ must satisfy (1.9). Let λ_2 be the maximum eigenvalue λ, other than λ_1. Then there is a vector c satisfying

$$(\Sigma - \lambda_2 I_p)c \;=\; 0 \quad , \qquad c^t c \;=\; 1 \;\; , \qquad c^t c_1 = 0. \tag{1.11}$$

Call this vector c_2 and the corresponding inner product $c_2^t y$, the *second principal component*.

Similarly, we can find the k^{th} principal component $c_k^t Y$ (for $k \leq p$) by solving $\max_{c_k^t c_k=1,\, c_k \perp c_i,\, i=1,2,\ldots,k-1} \mathrm{Var}(c_k^t Y)$, that is,

$$\max_{c_k} \{c_k^t \Sigma c_k - \lambda c_k^t c_k - 2 \sum_{i=1}^{k-1} \mu_i c_k^t \Sigma c_i\},$$

where $c_i^t Y,\ i = 1, 2, \ldots, k-1$ are the first $k-1$ principal components. In general, principal component analysis finds a new coordinate system for multivariate data such that the projection of Y on the first coordinate has maximal variance, projection on the second coordinate has maximal variance subject to being orthogonal to the first, etc.

In practice, we can use the sample variance, $S = (n-1)^{-1} \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})^t$ to estimate Σ, where $\bar{y} = n^{-1} \sum_{i=1}^n y_i$. After finding the eigenvalues and corresponding eigenvectors of S and sorting them in order of decreasing magnitude of eigenvalue, the first few eigenvectors are retained as the first few principal component directions.

1.2.3 Least-squares interpretation

Alternatively, we can approach our problem from the least-squares or geometric point of view. For a multivariate data set $\{y_i,\ i = 1, 2, \ldots, n\}$, the restated problem is to find a unit vector $v \in \mathbf{R}^p$ minimizing the sum of squared distances between y_i and their projections on v, $i = 1, 2, \ldots, n$, that is, $\min_{v^t v=1} n^{-1} \sum_{i=1}^n \|y_i - (y_i, v)v\|^2$. Now we can write down the Lagrange multiplier objective function,

$$f(v, \lambda) \;=\; \frac{1}{n} \sum_{i=1}^n \|y_i - (y_i, v)v\|^2 \;+\; \lambda(v^t v - 1)$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}(y_i-(y_i,v)v)^t(y_i-(y_i,v)v) \;+\; \lambda(v^tv-1)\\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i^ty_i-(y_i,v)v^ty_i) \;+\; \lambda(v^tv-1)\\
&= \frac{1}{n}\sum_{i=1}^{n}\|y_i\|^2 \;-\; v^tC_{yy}v \;+\; \lambda(v^tv-1), \qquad (1.12)
\end{aligned}
$$

where $C_{yy}=n^{-1}\sum_{i=1}^{n} y_iy_i^t$. Set $\nabla_v f(v,\lambda)=0$. Then

$$
0 \;=\; \nabla_v f(v,\lambda) \;=\; -2C_{yy}v \;+\; 2\lambda v, \qquad (1.13)
$$

which implies that $C_{yy}v=\lambda v$. Since $n^{-1}\sum_{i=1}^{n}\|y_i\|^2$ is a data-dependent constant, not depending upon v, the minimum of $n^{-1}\sum_{i=1}^{n}(\|y_i\|^2-(y_i,v)^2)$ corresponds to the maximum of $n^{-1}\sum_{i=1}^{n}(y_i,v)^2$. Then v is interpreted as the direction in which $y_i,\ i=1,2,\ldots,p,$ has maximum variation. Thus v^ty is an alternative definition for the first principal component, and we call $v^ty_i,\ i=1,2,\ldots,n$ the *loadings* for v.

Remark: The principal component directions v_i defined in this subsection are exactly the same as those of the previous subsection if and only if all of the vectors y_i are orthogonal to the vector of all 1's; that is, if and only if $\bar{y}=0$. □

Denoting the v just defined by v_1, we can find v_2 by minimizing the sum over $i=1,2,\ldots,n$ of the distances between y_i and their projections on v_2, subject to the constraints $\|v_2\|^2=1,\ v_2\perp v_1$. That is, we find $\min_{v_2^tv_2=1,\,v_2^tv_1=0}\ n^{-1}\sum_{i=1}^{n}\|y_i-(y_i,v_2)v_2\|^2$. Then v_2 is the eigenvector of C_{yy} corresponding to the second largest eigenvalue, and v_2^ty is the second alternative principal component, with loadings $v_2^ty_i,\ i=1,2,\ldots,n$. Similarly, we can obtain the third, fourth, and higher alternative principal components. This sequence of optimizations yields unique solutions

when the eigenvalues of C_{yy} are distinct.

Next consider the problem of minimizing, simultaneously with respect to two orthonormal vectors $u,\ v,$ the sum of squared distances between $y_i,\ i = 1, 2, \ldots, n$ and their projections on the span of $u,\ v$. That is,

$$\min_{u^t u=1,\, v^t v=1,\, u \perp v} \frac{1}{n} \sum_{i=1}^{n} \|y_i - (y_i, u)u - (y_i, v)v\|^2. \tag{1.14}$$

The result is the sum of the two largest eigenvalues of $C_{yy} = n^{-1} \sum_{i=1}^{n} y_i y_i^t$. The resulting vectors $u,\ v$ are not unique, even in the case where all eigenvalues are distinct, but the space which they span is unique. Thus all solutions are of the form

$$u \ = \ \alpha_1 v_1 \ + \ \beta_1 v_2 \qquad\qquad v \ = \ \alpha_2 v_1 \ + \ \beta_2 v_2, \tag{1.15}$$

where v_1 and v_2 are eigenvectors of C_{yy} corresponding to the two largest eigenvalues of C_{yy}, $\alpha_i^2 + \beta_i^2 \ = \ 1,\ i = 1, 2,$ and $\alpha_1 \alpha_2 \ + \ \beta_1 \beta_2 \ = \ 0.$

As was mentioned above, the two approaches to constructing principal components will coincide if and only if $\bar{y} = 0$. The sample covariance matrix S is constructed from centered data vectors, that is, the residuals of these vectors from their average $\bar{y}$, while the sum C_{yy} of exterior products in the second approach is defined without subtracting the average vector and perserves the original shape of the data in applications where the vectors y_i are interpreted as curves. If these vectors y_i are replaced by $y_i - \bar{y},\ i = 1, 2, \ldots, n,$ which generally distorts their shape if the vectors represent curves, then C_{yy} and S differ only by a constant multiple, and the unit eigenvectors defined by the two approaches are obviously identical.

If there exists q such that $\mathrm{Var}(y, v_q)$ is much larger than $\mathrm{Var}(y, v_k)$, $k = q+1, \ldots, p$, then we can treat the term $\sum_{k=q+1}^{p} (y_i, v_k) v_k$ as an error term, which contains high frequency noise. If one wishes to select the number q of principal components, the following criterion can be used to determine the number q of principal components to retain in describing data y_i, $i = 1, 2, \ldots, n$:

$$\frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq 1 - \alpha \qquad \text{or} \qquad \frac{\lambda_{q+1}}{\sum_{i=1}^{p} \lambda_i} \leq \beta, \tag{1.16}$$

for suitably defined constants α, β, usually, .05, .01 respectively.

The q principal components of the observed vector y_i are given by the vector $x_i = \pi_1^t (y_i - \bar{y})$, where $\pi_1 = (c_1, c_2, \ldots, c_q)$. The variables x_i are uncorrelated such that the covariance matrix $n^{-1} \sum_{i=1}^{n} x_i x_i^t$ is diagonal with elements λ_i. A complementary property of PCA, and the most closely related to the original discussions of Pearson (1901) is that, of all orthogonal linear projections $x_i = \pi_1^t (y_i - \bar{y})$, the principal component projection minimizes the squared reconstruction error $\sum_{i=1}^{n} \|y_i - \hat{y}_i\|^2$, where the optimal linear reconstruction of y_i is given by $\hat{y}_i = \pi_1 x_i + \bar{y}$.

However, a notable feature of these definitions of PCA is the absence of an associated probability model for the observed data. This will limit the ability to derive PCA within an inferential statistical framework. Also, PCA restricts itself to a linear setting, where high-order statistical information is discarded. Probabilistic Principal Component Analysis (PPCA), proposed by Tipping and Bishop (1999), overcomes the first mentioned disadvantage by using a special factor analysis model.

1.3 Factor Analysis

1.3.1 Introduction

The origin of factor analysis is generally attributed to Charles Spearman (1904). His outstanding work in developing a psychological theory involving a single general factor and a number of specific factors goes back to 1904 when his paper "General intelligence, objectively determined and measured" was published in the *American Journal of Psychology*. He is regarded as the father of factor analysis because he devoted the remaining forty years of his life to the development of factor analysis (Harman 1976). Perhaps a more crucial study of the statistical aspects is the paper by Karl Pearson (1901), in which he establishes "the method of principal axes". Our least-squares interpretation from section 1.2.3 is an extension of the original discussions of Pearson.

Factor analysis is a branch of multivariate analysis that is generally understood to refer to a set of closely related models intended for exploring or establishing correlational structures amongst the observed random variables. The method was developed primarily to provide a mathematical model for the explanation of psychological theories concerning human ability and behavior. It was originally used for the analysis of scores in mental tests. However, the methods are useful in a much wider range of situations.

Applications of factor analysis in fields other than psychology have become very popular since 1950, along with the development of the computer. These fields include such varied disciplines as meteorology, medicine, political science, taxon-

omy, archaeology, economics, and sociology. Factor analysis is used as a tool in the empirical sciences. In order to analyze observed data, one approach is to provide a statistical model, to explain the underlying behavior of the data. Some simple examples include: (1) a linear regression for the prediction of school success from three entrance exams; (2) a mathematical curve, such as the normal distribution or one from the Pearson family of curves, for the study of an observed frequency distribution; (3) a chi-square test of significance for the independence of such classifications as "treated or not treated with a certain serum"; (4) the least-squares interpretation from section 1.2.3 is the mathematical motivation, which extends the original ideas from Pearson (1901).

1.3.2 Model definition

Let the observable vector Y be written as

$$Y = Wf + \mu + \epsilon, \tag{1.17}$$

where Y, μ, and ϵ are column vectors of p components, f is a column vector of q ($q < p$) components, and W is a $p \times q$ matrix with nonrandom constant elements. We assume that ϵ is distributed independently of f and with mean $E\epsilon = 0$ and covariance matrix $E(\epsilon\epsilon^t) = \Psi$ which is diagonal. The $p \times q$ matrix W relates the two sets of variables, while the parameter vector μ permits the model to have non-zero mean. The vector f will be treated alternatively as a random vector and as a vector of parameters that varies from observation to observation. The elements of W are called factor loadings, and the elements of f are called common factors.

Remark: In principle, the model with random factors is appropriate when different samples consist of different individuals; the nonrandom factor model is more suitable when the specific individuals involved and not just the structure are of interest. □

Conventionally, we let $f \sim N_q(0,\ I_q)$, which means the factors are uncorrelated. Let W^tW be diagonal; that is, let W have orthogonal columns. If the error ϵ is multivariate normal, the equation (1.17) leads to a corresponding normal distribution for the observations $Y \sim N_p(\mu,\ \Sigma)$, where $\Sigma = WW^t + \Psi$. If $y_1, y_2, \ldots, y_n$ are a set of n observations of Y, the likelihood for this sample is

$$L = (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \exp\{-\frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^t\Sigma^{-1}(y_i - \mu)\}. \tag{1.18}$$

The model parameters $\mu,\ W$, and Ψ can be determined by maximum likelihood estimation whenever they are functionally determined by $(\mu,\ \Sigma)$. First we find that the maximum likelihood estimator of the mean μ is

$$\hat{\mu} = \overline{y} = \frac{1}{n}\sum_{1}^{n} y_i. \tag{1.19}$$

Secondly, before we find the other parameters W and Ψ, we add the restrictions (Anderson 1984 and Basilevsky 1994) that

$$\Gamma = W^t\Psi^{-1}W \tag{1.20}$$

is diagonal. If the diagonal elements of Γ are ordered and different ($\gamma_{11} > \gamma_{22} > \ldots > \gamma_{qq}$), then W is uniquely determined. When $\Psi = \sigma^2 I_p$, the model $\Sigma = WW^t + \sigma^2 I_p$ is identifiable. Next, we maximize the logarithm of (1.18) with μ replaced by $\hat{\mu}$; this is called the logarithm of the "concentrated likelihood"; that is

$$-\frac{1}{2}np\ \log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{n}{2}tr(S\Sigma^{-1}) \tag{1.21}$$

where $S = n^{-1} \sum_{i=1}^{n} (y_i - \overline{y})^{\otimes 2}$. Usually, there is no closed-form analytic solution for W and Ψ, except in special cases, such as $\Psi = \sigma^2 I_p$. Then W and Ψ can be obtained by an iterative procedure.

1.3.3 Factor analysis and PCA

By constraining the error covariance Ψ to be a diagonal matrix whose elements ψ_i are usually estimated from the data, the key assumption for the factor analysis model is that the observed variables Y_i are conditionally independent when the common factors f are given. These common factors are intended to explain the correlations between observation variables while ϵ_i represents variability unique to a particular Y_i. This is where factor analysis fundamentally differs from standard PCA, which effectively treats covariance and variance identically (Tipping 1999).

Because of the distinction made between variance and covariance in the standard factor analysis model, the subspace defined by the maximum likelihood estimates of the columns of W will generally not correspond to the principal subspace of the observed data. However, certain links between the two methods have been established, such as the connection for the special case of an isotropic error model, where the residual variances $\Psi = \sigma^2$ are constrained to be equal.

The approach was used in the early Young-Whittle factor analysis model (Young 1940), where the residual variance σ^2 was assumed known; that is, the model likelihood was a function of W alone. In that case, maximum likelihood estimation is equivalent to a least-squares method, and a principal component solution appears in a straightforward manner. Also, the common factors f were considered

as parameters to be estimated rather than random variables. However, a stochastic treatment of f recovers a similar result, that the $p-q$ smallest eigenvalues of the sample covariance S are equal to σ^2. It is simple to show that both W and σ^2 are determined analytically through eigen-decomposition of the sample covariance matrix S, without making use of iteration (Anderson 1963 and Basilevsky 1994).

1.4 Probabilistic PCA

Probabilistic PCA (PPCA) is an extension of traditional PCA, proposed by Roweis (1997), Tipping and Bishop (1999). The goal is to define a proper probability model for PCA. Note that in traditional PCA, we project all data from p-dimensional space to a principal subspace, which is spanned by the q principal axes. The components of data "outside" the principal subspace are simply discarded. In PPCA, however, these components are assumed to be i.i.d. Gaussian white noise.

The original p-dimensional observed data vector y_i, $i = 1,\ 2,\ \ldots,\ n$, can be described in terms of a lower q-dimensional data x_i $(q < p)$ and a noise term

$$y_i = Ax_i + \epsilon, \tag{1.22}$$

where A is a $p \times q$ loading matrix $(q < p)$ and ϵ is multivariate i.i.d. Gaussian with a diagonal covariance matrix $\sigma^2 I_p$. This model is also called a latent variable model: the latent variable x_i is related to a p-dimensional observation y_i. The distribution of the latent variable also Gaussian, and conventionally specified as $x_i \sim N_q(0,\ I_q)$. The marginal distribution for the observed data vector y is obtained by integrating

out the latent variable and is also Gaussian:

$$y_i \sim N_p(0,\ \Sigma), \tag{1.23}$$

where the observation covariance is specified by $\Sigma = AA^t + \sigma^2 I_p$. From the definition of PPCA, W is intuitively found as in the original PCA, and σ^2 is found by calculating the average of the variances in the discarded directions:

$$\hat{\sigma}^2 = \frac{1}{p-q}\sum_{i=q+1}^{p} \lambda_i. \tag{1.24}$$

The probabilistic PCA can be utilized as a general Gaussian density model. The benefit of doing so is that the maximum likelihood estimates for the parameters associated with the covariance matrix can be computed from the sample principal components. Tipping and Bishop (1999) show that MLEs of A and σ^2 are given by the following:

$$A_{ML} = U_q(\Lambda_q - \sigma^2 I_q)^{\frac{1}{2}} R, \tag{1.25}$$

and

$$\sigma^2_{ML} = \frac{1}{p-q}\sum_{j=q+1}^{p} \lambda_j = \frac{1}{p-q}[\mathrm{tr}(\Sigma) - \mathrm{tr}(\Lambda_q)], \tag{1.26}$$

where $U_q = (u_1,\ u_2,\ \ldots,\ u_q)$ and $\Gamma_q = \mathrm{Diag}(\lambda_1,\ \lambda_2,\ \ldots,\ \lambda_q)$ contain the top q eigenvectors and eigenvalues of Σ, respectively, and R is any orthogonal matrix with the typical choice being $R = I_q$. Alternatively, using a Gaussian prior (zero mean and unit standard deviation) over the latent variables x_i, we find the latent variable formulation leads naturally to an iterative and computationally efficient expectation-maximization (EM) algorithm by treating the latent variables as 'missing' data. Furthermore, the probability model also can be easily extended to mixture models,

by introducing a mean μ^k for each model k and re-estimating $p(y_i|k)$, and the prior probabilities for each model, $p(k)$, in each step of the EM algorithm. Bishop has also proposed Bayesian methods to automatically determine m, the number of dimensions to retain.

The probabilistic formulation of PCA from a Gaussian latent variable model is closely related to statistical factor analysis. But we note that an important distinction results from the use of the isotropic noise covariance $\sigma^2 I_p$; that is, PPCA is covariant by rotation of the original data axes, as is the standard PCA, while factor analysis is covariant under component-wise rescaling. Another point of contrast is that in factor analysis, neither of the factors found by a two-factor model is necessarily the same as that found by a single-factor model. In probabilistic PCA, we see above that the principal axes may be found incrementally.

1.5 Overview of the Thesis

In Chapter 2, we give a detailed discussion of Random Effect Factor Model I (REFM_1). We first introduce this new model based on a homogeneous group of observed random vectors and discuss the identifiability of all parameters θ. The likelihood method is implemented in two steps, first by assuming that part of the parameter vector, θ_2, is fixed, where θ_2 parameterizes the factor directions. The restricted MLE, $\hat{\theta}_1$, for the other unknown parameter components is a unique continuous function of the factor directions θ_2. Next, the profile likelihood is introduced. Also, we discuss the asymptotic behaviors of the likelihood function and the profile likelihood function. We prove that there is a unique local maximum of the asymp-

totic profile likelihood, leading to the conclusion that the MLE is a locally consistent estimator. We discuss the calculus maximization and the Hessian matrix of the likelihood. We prove the combination of the restricted MLE and MLE of the profile likelihood is the joint MLE based on the original likelihood. Finally, we establish the asymptotic normal distribution of the MLE.

In Chapter 3, we discuss how to analyze the multivariate data from s groups. First, we review the relationship among covariance matrices $\Sigma_1, \ \ldots, \ \Sigma_s$, and the common principal component (CPC) model. Next, we extend the REFM_1 model from a single group to multiple groups. Then we introduce two new models, REFM_2 and REFM_3, to fit s groups of multivariate data. The REFM_2 model assumes all s groups have a common factor space but differing mean and variance parameters for factor loadings and error terms, and REFM_3 is a model which has not only a common factor space for all s groups but also an additional individual factor space belonging to each group. We discuss the identifiability of parameters and use the profile likelihood method to obtain the MLE.

In Chapter 4, we introduce the Newton-Raphson method and the EM algorithm, develop an EM algorithm to compute the MLE for REFM_1, and extend the algorithm to REFM_2. The performance of the algorithms on simulated data is described. The Quasi-Newton methods are also used to calculate the maximum likelihood estimation of the profile likelihood function $l_p(\theta_2)$ and are shown to give results for REFM_1 that agree with the EM algorithm.

In Chapter 5, we introduce a real dataset of ultrasound cross-sectional images of the human tongue during speech. We apply the EM algorithm directly to the

final, preprocessed tongue data set. For computational efficiency, we choose a lower dimensional principal subspace and apply the EM algorithm to the data set projected to that subspace. We compare the numerical results of the EM algorithm on REFM_1 with results of previous analysis of the data.

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

Chapter 2

Random Effect Factor Model I

2.1 The Model

2.1.1 Introduction

Our motivating data set of two-dimensional coronal tongue contours, which we will discuss in greater detail in Chapter 5, is very high dimensional, very noisy, with a high degree of cross-classification. With so many dimensions, it will be difficult to see any pattern in its inter-relationships. In fact, our ability to visualize relationships is limited to two or three dimensions, which places us under extraordinary pressure to reduce the dimensionality of the data in a manner which preserves as much of the structure as possible. Our objective is first to condense the many measured variables into a much smaller number with as little loss of information as possible, and secondly to build a model using the reduced dimensional data to represent a true two-dimensional tongue surface.

In mathematical language, we observe p-dimensional random vectors with p a relatively large number, and we want to find a small number, q, of orthonormal vectors whose linear combinations provide a good fit with high probability to the observed vectors. Our first model will partition the observed vector Y into two parts: an unobserved systematic part and an unobserved error part. The components of the error vector are considered uncorrelated or independent, while the systematic part is taken to be a linear combination of a relatively small number of axis directions in

a new q-dimensional coordinate system. The subspace spanned by these coordinate directions is called the factor space. In terms of this factor space, any observation, a point in p dimensional space, will be projected to a point in the q dimensional factor subspace. If the coordinates of this projection are considered parameters, then unfortunately the number of parameters goes up in proportion to the sample size, and this creates problems with the behavior of maximum likelihood estimators (Neyman and Scott 1948). However, there are circumstances in which such methods are relatively simple, and can be made to yield estimates of the parameters which are virtually the same as those derived from a random effect model. "They thus have a certain practical interest but in spite of a voluminous and often polemical literature they are, from our standpoint, outside the mainstream of theoretical development" (Bartholomew 1987). Thus, we introduce the random effect model here to reduce the number of parameters.

2.1.2 Definition of the Model

Random Effect Factor Model 1 ($\mathbf{REFM}_1$). Assume that the observable random vector Y can be written as

$$Y = \sum_{k=1}^{q} c_k P_k + \epsilon \quad , \tag{2.1}$$

where Y, the nonrandom orthonormal coordinate directions P_k, $k = 1, 2, \ldots, q$, and ϵ lie in $\mathbf{R}^p$. Assume that the random effects $c_k \sim \mathcal{N}(a_k, b_k^2)$, $1 \leq k \leq q$, and the error $\epsilon \sim \mathcal{N}_p(0, \sigma^2 I_p)$ are independent.

Later on, we refer to this model as $\mathbf{REFM}_1$. Now, we calculate the mean and

variance of random vector Y under the model:

$$E(Y) = \sum_{k=1}^{q} a_k P_k = \pi_1 \underline{a} \tag{2.2}$$

and

$$Var(Y) = \sum_{k=1}^{q} b_k^2 P_k^{\otimes 2} + \sigma^2 I_p = \pi_1 B_1 \pi_1^t + \sigma^2 I_p \tag{2.3}$$

where

$$\pi_1 = (P_1, P_2, ..., P_q) \tag{2.4}$$

is a $p \times q$ matrix with orthonormal column vectors, and

$$\underline{a}^t = (a_1, a_2, ..., a_q) \tag{2.5}$$

is a q dimensional mean vector. In terms of π_1, we define the factor space $V_1 \equiv col(\pi_1) = span\{P_1, P_2, ..., P_q\}$, and

$$B_1 = \text{Diag}(b_1^2,\, b_2^2 \,\ldots\, b_q^2) \tag{2.6}$$

is a $q \times q$ matrix with diagonal elements $b_1^2,\, b_2^2 \,\ldots\, b_q^2$ and zeros elsewhere. Based on the model assumption, Y follows a normal distribution with mean $\mu_y = \pi_1 \underline{a}$ and covariance matrix $\Sigma_y = \pi_1 B_1 \pi_1^t + \sigma^2 I_p$.

If there are n independent observations with the distribution of Y, say, y_1, $\ldots$, y_n, our data model under $\mathbf{REFM}_1$ is

$$y_i = \sum_{k=1}^{q} c_{ik} P_k + \epsilon_i \quad , \qquad i = 1, 2, ..., n \tag{2.7}$$

where random effects $c_{ik} \sim \mathcal{N}(a_k, b_k^2)$, $1 \le k \le q$, are independent, $\epsilon_1,\, \epsilon_2,\, \ldots, \epsilon_n$ are i.i.d. with $\epsilon_i \sim \mathcal{N}_p(0, \sigma^2 I_p)$, the series $\{\epsilon_i\}$ and $\{c_{ik}\}$ are independent, and

$P_1, P_2, ..., P_k$ are orthonormal. Thus,

$$y_i \sim N_p(\mu_y, \Sigma_y), \tag{2.8}$$

where

$$\begin{cases} \Sigma_y = \pi_1 B_1 \pi_1^t + \sigma^2 I_p \\ \\ \mu_y = \pi_1 \underline{a} \end{cases} \tag{2.9}$$

If we could regard the unknown parameter as μ_y and Σ_y with only the restriction that Σ_y is a nonsingular covariance matrix, then the maximum likelihood estimates of μ_y and Σ_y would be the sample mean $\bar{y} = n^{-1} \sum_{i=1}^n y_i$, and $(n-1)/n$ times the sample variance S_y, or $n^{-1} \sum_{i=1}^n (y_i - \bar{y})^{\otimes 2}$. Also, $\bar{y}$ and S_y would be **sufficient** statistics for the parameters μ_y and Σ_y. Under $\mathbf{REFM}_1$, the mean parameter μ_y and covariance parameter Σ_y can be expressed in terms of $\pi_1, \underline{a}, B_1$, and σ^2 through equation (2.9). Now, let us define the parameter space for $\mathbf{REFM}_1$.

Let $\theta = (\underline{a},\ \underline{b}^2,\ \sigma^2,\ \pi_1)$, where $\underline{b}^2 \equiv \text{Diag}(B_1)$. We define the parameter space as $\Theta = \{\theta :\ -\infty < a_k < +\infty,\ 0 < b_k^2 < +\infty,\ 0 < \sigma^2 < +\infty,\ \pi_1^t \pi_1 = I_q \text{ for } 1 \leq k \leq q, \}$. Thus, Θ is a subset of $R^q \times (R^+)^{q+1} \times (R^p)^q$. The true value θ will be denoted by θ_0. We first need to show that our parameter θ is identifiable from the observed data in $\mathbf{REFM}_1$.

2.1.3 Identifiability

A parameter θ for a family of distributions $p_\theta, \theta \in \Theta$, is said to be identifiable if distinct values of θ correspond to distinct distribution; that is, θ is identifiable if $\theta \neq \theta'$ implies $p_\theta \neq p_{\theta'}$. The mean μ and covariance matrix Σ_y determine p_θ when p_θ is a family of multivariate normal distributions.

Given a mean μ_y and a covariance matrix Σ_y and a number q of factors, we ask whether there exist $\underline{a}, \underline{b}^2, \pi_1$, and σ^2 to satisfy (2.9), and if so, whether they are unique, and what is the relationship. First, let us count how many equations in (2.9) we have, how many parameters in REFM_1, and what is the relationship between them. There are q equations regarding μ_y, and $\frac{p(p+1)}{2}$ equations regarding Σ_y since the covariance matrix Σ_y is a symmetric matrix. Hence, the total number of equations is $q + \frac{p(p+1)}{2}$. There are q, q, 1, and $(pq - \frac{q(q+1)}{2})$ parameter dimensions for $\underline{a}, B_1, \sigma^2$, and π_1, respectively. Thus, the total of number of parameters under REFM_1 is $2q + 1 + pq - \frac{q(q+1)}{2}$. Subtracting the number of parameters from number of equations and simplifying, we have

$$(\#of\ equations) - (\#of\ parameters) = \frac{1}{2}(p - q + 2)(p - q - 1). \tag{2.10}$$

Since $p > q$, the above expression indicates that the dimension of Θ is less than that of the unrestricted multivariate normal model.

The factor space, $V_1 = col(\pi_1)$ is also a subspace of $V \equiv R^p$. We define V_2 as the orthogonal complement space of V_1 in V, that is, $V_1 \bigoplus V_2 = V$, with V_2 a $(p - q)$ dimensional space. Let $\{P_i,\ i = q + 1,\ q + 2, \ldots,\ p\}$ be any orthonormal basis for V_2, that is, $\|P_i\|^2 = 1$, for $i > q$, and $P_i^t P_j = 0$, for $i,\ j >\ q$ and $i \neq j$. Note that, $u^t v = 0$ if $u \in V_1$ and $v \in V_2$.

The covariance matrix Σ_y is positive definite with rank p, and has all positive eigenvalues. Multiplying both sides of equation (2.9) by P_i, we have

$$\Sigma_y P_i = \begin{cases} (b_i^2 + \sigma^2)P_i & \text{for } i \leq q, \\ \sigma^2 P_i & \text{for } i > q. \end{cases} \tag{2.11}$$

Thus, all of $P_1,\ P_2,\ ...,\ P_q,\ P_{q+1},\ ...,\ P_p$ are eigenvectors of the covariance matrix Σ_y corresponding to eigenvalues $b_1^2+\sigma^2,\ b_2^2+\sigma^2,\ ...,\ b_q^2+\sigma^2,\ \sigma^2,\ ...,\ \sigma^2$. If we sort the eigenvalues from the biggest to the smallest, and assume the condition:

$$b_1^2 > b_2^2 > ... > b_q^2, \tag{2.12}$$

then we can uniquely determine the eigenvalues. Thus, we can uniquely determine the sequence $\{b_k^2, k = 1, 2, \ldots, q\}$, the orthonormal eigenvectors $P_1,\ P_2,\ ...,\ P_q$, and σ^2. This really means that, subject to (2.12), the consistently estimable multivariate-normal parameters μ_y, Σ_y uniquely determine $B_1,\ \pi_1$, and σ^2. Then, in terms of the mean μ_y of random vector Y, the unique π_1 determines a unique $\underline{a}$ because we can solve for $\underline{a}$ from equation (2.9); that is,

$$\underline{a} = ({\pi_1}^t\pi_1)^{-1}{\pi_1}^t\mu_y = {\pi_1}^t\mu_y. \tag{2.13}$$

When $q = 1$, the condition (2.12) is vacuous. Since y_i are i.i.d.,

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \xrightarrow{a.s.} E(Y) = aP_1, \tag{2.14}$$

by the Strong Law of Large Numbers (SLLN). That is,

$$Pr\left\{\frac{1}{n}\sum_{i=1}^{n} y_i = aP_1\right\} = 1. \tag{2.15}$$

Hence, P_1 is determined if $a \neq 0$, since $E(Y)$ and P_1 are in the same or reverse direction, and P_1 is a unit vector. In any case, $a = \|E(Y)\|$ is determined, and P_1 is the unique eigenvector associated with the maximal eigenvalue of Σ_y. Then b^2 and σ^2 are determined from the system of equations (2.11).

2.1.4 Relationship with other multivariate models

The identifiability condition (2.12) for the REFM_1 will be mentioned again in later sections and chapters. It is the key assumption in proving the existence of a local unique maximum for the asymptotic profile likelihood and the consistency of that estimate. Here we want to emphasize that the identifiability condition (2.12) guarantees that V_1 is a principal subspace and that $\{P_i, 1 \leq i \leq q\}$ are principal axes. Thus, the REFM_1 model includes what Principal Component Analysis covers.

We have the parameters $(\underline{a}, \underline{b}^2, \sigma^2, \pi_1)$ where $\pi_1^t \pi_1 = I_q$ in REFM_1, and $(W,\ \sigma^2)$, where W has orthogonal columns in the Factor Analysis model. When $a = 0$, REFM_1 becomes a factor analysis model. The matrix W (equivalent to π_1) only appears in the covariance structure through the factor loadings of the Factor Analysis model, while π_1 appears in both the mean and variance in REFM_1. This is the key difference between REFM_1 and the FA model. REFM_1 models the mean parameter, and FA model does not. Specifically, when we apply the EM algorithm for both models in data simulation, FA struggles to converge, and REFM_1 easily finds the maximum point because π_1 is constrained more by the data in REFM_1 compared to the FA model.

The PPCA is a special case of FA model with $\mu = 0$ and error $\sigma^2 I_p$. This makes PPCA a closer model to REFM_1, compared to PCA and FA models. The REFM_1 model still has the advantage in numerical computation because the parameter π_1 is shared by mean and variance.

2.2 Maximum Likelihood Estimates for REFM_1

2.2.1 Simplifying the probability density function

In this section, we will find the maximum likelihood estimators for the parameters $\underline{a}, \underline{b}^2, \sigma^2$, and π_1 under $\mathbf{REFM}_1$ when the observations are normally distributed, that is, when the factor scores and errors are normal,

$$Y \sim \mathcal{N}_p(\pi_1 \underline{a}, \pi_1 B_1 \pi_1^t + \sigma^2 I_p) \tag{2.16}$$

where π_1, $\underline{a}$, and B_1 satisfy (2.4) - (2.6).

We start with the probability density function of Y under $\mathbf{REFM}_1$:

$$f(y) \;=\; \frac{\exp\{-\frac{1}{2}(y-\pi_1\underline{a})^t(\pi_1 B_1 {\pi_1}^t + \sigma^2 I_p)^{-1}(y-\pi_1\underline{a})\}}{(2\pi)^{p/2}|\pi_1 B_1 {\pi_1}^t + \sigma^2 I_p|^{1/2}}. \tag{2.17}$$

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

We are interested in estimating the parameters $\underline{a},\ \underline{b}^2,\ \sigma^2,\ P_1,\ P_2,\ \ldots,\ P_q$. The number of free parameters in $\underline{a},\ \underline{b}^2,\ \sigma^2$, and π_1 are q, q, 1, and $pq - \frac{q(q+1)}{2}$, since π_1 has orthonormal columns. Since $\pi_1 B_1 \pi_1^t$ is a $p \times p$ matrix with rank q, we can write by the definition of B_1 and π_1

$$\pi_1 B_1 {\pi_1}^t = \sum_{k=1}^{q} b_k^2 P_k P_k^t = \sum_{k=1}^{p} b_k^2 P_k P_k^t = \pi B \pi^t, \tag{2.18}$$

where

$$\pi = (P_1, ..., P_q, P_{q+1}, ..., P_p) \tag{2.19}$$

and B is the $p \times p$ diagonal matrix with $b_k^2 \equiv 0$ for $k > q$:

$$B = \text{Diag}(b_1^2, \ldots, b_q^2, b_{q+1}^2, \ldots, b_p^2). \tag{2.20}$$

Now, we rewrite and simplify the probability density of the random vector Y:

$$\begin{aligned}
f(y) &= (2\pi)^{-\frac{p}{2}}|\pi B\pi^t + \sigma^2\pi\pi^t|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(y-\pi_1\underline{a})^t(\pi B\pi^t+\sigma^2\pi\pi^t)^{-1}(y-\pi_1\underline{a})\} \\
&= (2\pi)^{-\frac{p}{2}}|\pi(B+\sigma^2 I_p)\pi^t|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(y-\pi_1\underline{a})^t[\pi(B+\sigma^2 I_p)\pi^t]^{-1}(y-\pi_1\underline{a})\} \\
&= (2\pi)^{-\frac{p}{2}}|B+\sigma^2 I_p|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(y-\pi_1\underline{a})^t\cdot\pi(B+\sigma^2 I_p)^{-1}\pi^t(y-\pi_1\underline{a})\} \\
&= (2\pi)^{-\frac{p}{2}}\left[\prod_{k=1}^{q}(b_k^2+\sigma^2)^{-\frac{1}{2}}\right](\sigma^2)^{-\frac{p-q}{2}} \\
&\qquad \cdot\exp\{-\frac{1}{2}[\pi^t(y-\pi_1\underline{a})]^t(B+\sigma^2 I_p)^{-1}\pi^t(y-\pi_1\underline{a})\} \\
&= (2\pi)^{-\frac{p}{2}}(\sigma^2)^{-\frac{p-q}{2}}\left[\prod_{k=1}^{q}(b_k^2+\sigma^2)^{-\frac{1}{2}}\right]\cdot\exp\{-\frac{1}{2}[\pi^t(y-\pi_1\underline{a})\right]^t \\
&\qquad \text{Diag}\left(\frac{1}{b_1^2+\sigma^2},\ldots,\frac{1}{b_q^2+\sigma^2},\sigma^{-2},\ldots,\sigma^{-2}\right)[\pi^t(y-\pi_1\underline{a})]\}. \qquad (2.21)
\end{aligned}$$

Since

$$P_k^t(y-\pi_1\underline{a}) = P_k^t y - P_k^t\pi_1\underline{a} = \begin{cases} P_k^t y - a_k & \text{if } k\le q \\ P_k^t y & \text{if } k>q \end{cases} \qquad (2.22)$$

it follows that

$$\pi^t(y-\pi_1\underline{a}) = \begin{pmatrix} P_1^t(y-\pi_1\underline{a}) \\ \ldots \\ P_q^t(y-\pi_1\underline{a}) \\ P_{q+1}^t(y-\pi_1\underline{a}) \\ \ldots \\ P_p^t(y-\pi_1\underline{a}) \end{pmatrix} = \begin{pmatrix} P_1^t y - a_1 \\ \ldots \\ P_q^t y - a_q \\ P_{q+1}^t y \\ \ldots \\ P_p^t y \end{pmatrix}. \qquad (2.23)$$

Finally, we have simplified our probability density function to:

$$\begin{aligned}
f(y) &= (2\pi)^{-\frac{p}{2}}(\sigma^2)^{-\frac{p-q}{2}}\left[\prod_{k=1}^{q}(b_k^2+\sigma^2)^{-\frac{1}{2}}\right] \\
&\qquad \cdot\exp\left\{-\frac{1}{2}\sum_{k=1}^{q}\frac{1}{b_k^2+\sigma^2}(P_k^t y-a_k)^2 - \frac{1}{2\sigma^2}\sum_{k=q+1}^{p}(P_k^t y)^2\right\}. \qquad (2.24)
\end{aligned}$$

The Maximum Likelihood Estimate of θ is defined as that value $\hat{\theta}$ of θ which maximizes the probability density; hence the Maximum Likelihood Estimate of $g(\theta)$ is $g(\hat{\theta})$.

2.2.2 Likelihood function and ML equations

If $y_1,\ ...,\ y_n$ are a sample of n independent observations on Y, the joint probability density function $f(y_1,\theta)\cdots f(y_n,\theta)$, evaluated at $\mathbf{y} = (y_1,\ldots,y_n)$, and considered as a function of θ, is called the likelihood function. Under $\mathbf{REFM}_1$, the likelihood function for this sample $\{y_i,\ i = 1,\ 2,\ ...,\ n\}$ is

$$\begin{aligned} & L(\theta) = \prod_{i=1}^{n} f(y_i,\theta) \\ = \ & (2\pi)^{-\frac{np}{2}}(\sigma^2)^{-\frac{n(p-q)}{2}} \prod_{k=1}^{q}({b_k}^2+\sigma^2)^{-\frac{n}{2}} \\ \cdot \ & \exp\{-\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{q}\frac{1}{{b_k}^2+\sigma^2}(P_k^t y_i - a_k)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{k=q+1}^{p}(P_k^t y_i)^2\} \end{aligned} \tag{2.25}$$

The maximum likelihood estimates of θ are values $\hat{\theta}$ of θ which maximize the likelihood function $L(\theta)$, or equivalently, the logarithm of the likelihood function (since the logarithm function is strictly increasing). The log likelihood is denoted

$$\begin{aligned} l(\theta) \ \equiv \ & \log(L(\theta)) \\ = \ & -\frac{np}{2}\log(2\pi) - \frac{n(p-q)}{2}\log(\sigma^2) - \frac{n}{2}\sum_{k=1}^{q}\log(b_k^2+\sigma^2) \\ & -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{q}\frac{1}{b_k^2+\sigma^2}(P_k^t y_i - a_k)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{k=q+1}^{p}(P_k^t y_i)^2 \end{aligned} \tag{2.26}$$

In order to maximize the log-likelihood function, we first calculate the partial derivative of $l(\theta)$ with regard to a_k for $1 \le k \le q$,

$$\frac{\partial l(\theta)}{\partial a_k} = \sum_{i=1}^{n}\frac{1}{b_k^2+\sigma^2}(P_k^t y_i - a_k) \tag{2.27}$$

Setting this expression equal to 0 yields the first q likelihood equations,

$$\sum_{i=1}^{n}(P_k^t y_i - \hat{a}_k) = 0 \tag{2.28}$$

For $\pi_1 = (P_1,\ P_2,\ ...,\ P_q)$ assumed fixed, we can solve these equations for a_k as a function of π_1. That is,

$$\hat{a}_k = \frac{1}{n}\sum_{i=1}^{n} P_k^t y_i = P_k^t(\frac{1}{n}\sum_{i=1}^{n} y_i) = P_k^t \bar{y}, \tag{2.29}$$

where $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$. Furthermore, we have

$$\underline{\hat{a}} = \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ ... \\ \hat{a}_q \end{pmatrix} = \begin{pmatrix} P_1^t\bar{y} \\ P_2^t\bar{y} \\ ... \\ P_q^t\bar{y} \end{pmatrix} = \pi_1^t\bar{y}. \tag{2.30}$$

Remark: There is an alternative way to obtain the same estimation result. From (2.9), $\underline{a}$ is in one-to-one correspondence with the mean μ_y for given π_1, and is not related to the variance Σ_y; and the maximum likelihood estimator of the mean parameter μ_y is $\hat{\mu}_y = \bar{y} = n^{-1}\sum_{k=1}^{n} y_i$. We have a restricted maximum likelihood estimator $\underline{\hat{a}}$ for fixed (known) π_1 through the following equations:

$$\bar{y} = \hat{\mu}_y = \pi_1\underline{\hat{a}} \qquad \Longrightarrow \qquad \underline{\hat{a}} = \pi_1^t\bar{y}. \qquad \square$$

Similarly, we can differentiate $l(\theta)$ with respect to b_k^2 for all $k \leq q$:

$$\frac{\partial l(\theta)}{\partial b_k^2} = -\frac{n}{2}\frac{1}{b_k^2+\sigma^2} + \frac{1}{2}\sum_{i=1}^{n}\frac{1}{(b_k^2+\sigma^2)^2}(P_k^t y_i - a_k)^2 \tag{2.31}$$

Setting these expressions equal to 0, we have another system of q likelihood equation,

$$\hat{b}_k^2 + \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(P_k^t y_i - \hat{a}_k)^2 \tag{2.32}$$

Again for fixed $P_1,\ P_2, \ldots,\ P_q$, we can solve for $b_k^2 + \sigma^2,\ k \leq q$, using (2.29):

$$\begin{aligned}
\hat{b}_k^2 + \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(P_k^t y_i - P_k \bar{y})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}[P_k^t(y_i - \bar{y})]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} P_k^t(y_i - \bar{y})(y_i - \bar{y})^t P_k \\
&= P_k^t \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^{\otimes 2} P_k \\
&= P_k^t S P_k, \qquad (2.33)
\end{aligned}$$

where $S = n^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^{\otimes 2}$ is $(n-1)/n$ times the sample variance of the random vector Y. The derivative of $l(\theta)$ with respect to σ^2 is

$$\begin{aligned}
\frac{\partial l(\theta)}{\partial \sigma^2} &= -\frac{n(p-q)}{2}\frac{1}{\sigma^2} - \frac{n}{2}\sum_{k=1}^{q}\frac{1}{b_k^2 + \sigma^2} + \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{q}\frac{1}{(b_k^2 + \sigma^2)^2}(P_k^t y_i - a_k)^2 \\
&\quad + \frac{1}{2}\frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}\sum_{k=q+1}^{p}(P_k^t y_i)^2 \qquad (2.34)
\end{aligned}$$

Setting this expression equal to 0 yields another likelihood equation,

$$\begin{aligned}
&-\frac{n(p-q)}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}\sum_{k=q+1}^{p}(P_k^t y_i)^2 - \frac{n}{2}\sum_{k=1}^{q}\frac{1}{\hat{b}_k^2 + \hat{\sigma}^2} \\
&\qquad + \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{q}(\frac{1}{\hat{b}_k^2 + \hat{\sigma}^2})^2(P_k^t y_i - \hat{a}_k)^2 = 0. \qquad (2.35)
\end{aligned}$$

After substituting equation (2.32) in equation (2.35), we can simplify the equation (2.35) to

$$-\frac{n(p-q)}{2}\frac{1}{\hat{\sigma}^2} + \frac{1}{2}\frac{1}{(\hat{\sigma}^2)^2}\sum_{i=1}^{n}\sum_{k=q+1}^{p}(P_k^t y_i)^2 = 0. \qquad (2.36)$$

We can solve for $\hat{\sigma}^2$ as a function of $P_1,\ P_2,\ ...,\ P_q$,

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n(p-q)}\sum_{i=1}^{n}\sum_{k=q+1}^{p} P_k^t y_i y_i^t P_k \\
&= \frac{1}{p-q}\sum_{k=q+1}^{p} P_k^t C_{yy} P_k, \qquad (2.37)
\end{aligned}$$

where $C_{yy} = n^{-1}\sum_{i=1}^{n} y_i y_i^t$ is the sample second moment of the random vector Y.

There is a standard relationship between sample mean, sample variance and sample second moment, which we demonstrate in the following.

Lemma 1 *If $y_1, ..., y_n$ are a set of n observations on Y, with $\bar{y} \equiv n^{-1}\sum_{i=1}^{n} y_i$, $S \equiv n^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^{\otimes 2}$, and $C_{yy} \equiv n^{-1}\sum_{i=1}^{n} y_i^{\otimes 2}$, then*

$$C_{yy} = S + \bar{y}^{\otimes 2} \tag{2.38}$$

Proof.

$$\begin{aligned}
\mathbf{S} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(y_i^t - \bar{y}^t) \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i y_i^t - \bar{y}y_i^t - y_i\bar{y}^t + \bar{y}\bar{y}^t) \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i y_i^t) - \bar{y}\frac{1}{n}\sum_{i=1}^{n} y_i^t - (\frac{1}{n}\sum_{i=1}^{n} y_i)\bar{y}^t + \bar{y}^{\otimes 2}) \\
&= C_{yy} - \bar{y}^{\otimes 2} - \bar{y}^{\otimes 2} + \bar{y}^{\otimes 2} \\
&= C_{yy} - \bar{y}^{\otimes 2}. \qquad \square
\end{aligned}$$

$\square$

2.2.3 The profile log-likelihood

So far, we have the restricted maximum likelihood estimates $\underline{\hat{a}}$, $\hat{B}_1$, and $\hat{\sigma}^2$ in terms of fixed (assumed known) common factors $P_1, P_2, \ldots, P_q$ under $\mathbf{REFM}_1$. The estimators $\underline{\hat{a}}$, $\hat{B}_1$, and $\hat{\sigma}^2$ are functions of π_1, the sample mean $\bar{y}$, and the sample second moment C_{yy}. The idea of the **profile likelihood** is similar to the **concentrated likelihood** from Anderson (1984). We represent the parameter space as

the Cartesian product of two component subspaces and optimize the likelihood on one subspace first with the other parameter component fixed. Thus, we can work on the overall maximum in two separate steps. This can be done for the general log-likelihood case.

Let $l(\theta)$ be the logarithm of the likelihood function on the parameter space Θ. We can decompose Θ into two subspaces Θ_1 and Θ_2, with $\Theta = \Theta_1 \times \Theta_2$. Assume that $\theta_2 \in \Theta_2$ is given, and that there exists a unique maximum likelihood estimate $\hat{\theta}_1(\theta_2)$. Denote $\hat{\theta}_1(\theta_2) = h(\theta_2)$ for fixed data, where h is a well-defined smooth function of θ_2 (and the data). Then, we call $l(\hat{\theta}_1(\theta_2), \theta_2)$ the **profile log-likelihood**.

For our case, we let $\theta_1 = (\underline{a},\ \underline{b}^2,\ \sigma^2)$, and $\theta_2 = (P_1,\ P_2,\ ...,\ P_q)$. The parameter space Θ_1 is $\{(\underline{a}, \underline{b}^2, \sigma^2) \in R^q \times (R^+)^q \times R^+ : for\ 1 \leq k \leq q,\ b_k^2 > 0,\ \sigma^2 > 0\}$, and Θ_2 is $\{\theta_2 \in \mathcal{M} : \theta_2^t \theta_2 = I_q\}$, where $\mathcal{M}$ is the set of real $p \times q$ matrices. Rewrite expression (2.37) as

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{p-q} \sum_{i=q+1}^{p} \mathrm{tr}(P_k^t C_{yy} P_k) \\ &= \frac{1}{p-q} \mathrm{tr}((\sum_{i=q+1}^{p} P_k P_k^t) C_{yy}) \\ &= \frac{1}{p-q} \{\mathrm{tr}(C_{yy}) - \mathrm{tr}(\pi_1 \pi_1^t C_{yy})\}, \end{aligned} \tag{2.39}$$

and rewrite the expression (2.33) as

$$\hat{\underline{b}}^2 = \mathrm{Diag}(\pi_1^t S \pi_1) - \hat{\sigma}^2 1_q. \tag{2.40}$$

Here 1_q means a q-dimensional column vector where all elements are 1, that is, $1_q = (1, 1, \ldots, 1)^t$, and $\mathrm{Diag}(A)$ denotes the column vector of diagonal elements of a matrix; that is, if A $= (a_{i,j})$ is a $q \times q$ matrix, then $\mathrm{Diag}(A) = (a_{11}, a_{22}, \ldots, a_{qq})^t$.

Under **REFM**$_1$, we can exhibit the function $\hat{\theta}_1(\theta_2) \equiv h(\theta_2)$ explicitly from the equations (2.30), (2.40), and (2.39) as follows:

$$\hat{\theta}_1(\pi_1) = \begin{pmatrix} \hat{\underline{a}} \\ \hat{\underline{b}}^2 \\ \hat{\sigma}^2 \end{pmatrix} = \begin{pmatrix} \pi_1^t \bar{y} \\ \text{Diag}(\pi_1^t S \pi_1) - \hat{\sigma}^2 1_q \\ \frac{1}{p-q}\{\text{tr}(C_{yy}) - \text{tr}(\pi_1 \pi_1^t C_{yy})\}. \end{pmatrix} \tag{2.41}$$

The restricted joint maximum likelihood estimate $\hat{\theta}_1(\theta_2)$ is a function of θ_2, $\bar{y}$, and C_{yy}. Next, we will replace θ_1 in expression (2.26) by $\hat{\theta}_1(\theta_2)$. Hence, the profile likelihood function $l_p(\theta_2) = l(\hat{\theta}_1(\theta_2),\ \theta_2)$ is a function of θ_2, $\bar{y}$, and C_{yy} alone.

Remark: There is a close relationship among the factor space $V_1 = span\{P_1, P_2, \ldots, P_q\}$, the parameter space $\Theta_2 = \{\theta_2 \in \mathcal{M} : \theta_2^t \theta_2 = I_q\}$, and the $p \times q$ matrix $\pi_1 = (P_1, P_2, \ldots, P_q)$. By combining all q coordinate directions in V_1 into a $p \times q$ matrix π_1, we have an element of Θ_2. For any given element $\theta_2 \in \Theta_2$, the q column vectors of θ_2 are q coordinate directions in V_1. For any matrix π_1 with orthogonal columns, the q column vectors of π_1 are also q coordinate directions in V_1, and $\pi_1 \in \Theta_2$. Hence, the notations θ_2 and π_1 are interchangeable.

In a mathematical notation, our approach is as follows:

$$\begin{aligned} \sup_{\theta \in \Theta} l(\theta) &= \sup_{\theta_2 \in \Theta_2} \{\max_{\theta_1 \in \Theta_1} l(\theta|\theta_2)\} \\ &= \sup_{\theta_2 \in \Theta_2} l(\hat{\theta}_1(\theta_2), \theta_2) \\ &= \sup_{\theta_2 \in \Theta_2} l_p(\theta_2; \bar{y}, C_{yy}) \end{aligned} \tag{2.42}$$

An evaluation of the left hand side of equation (2.42) is the standard problem of maximum likelihood estimation. The right hand side of the equation (2.42) is

evaluated in our profile likelihood approach. The following Lemma will show that a sufficient condition for the equation (2.42) to hold is that a unique maximum likelihood estimate $\hat{\theta}_1(\theta_2)$ exists when θ_2 is given.

Lemma 2 *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*

$$\max_{\theta_1 \in \Theta_1} l(\theta_1; \theta_2) = l(\hat{\theta}_1(\theta_2); \theta_2) \equiv l_p(\theta_2), \tag{2.43}$$

then we have

$$\sup_{\theta \in \Theta} l(\theta) = \sup_{\theta_2 \in \Theta_2} l_p(\theta_2). \tag{2.44}$$

Furthermore, if $l_p(\theta_2)$ is continuous, and Θ_2 is compact, then the right hand side of equation (2.44) is a maximum, that is,

$$\sup_{\theta_2 \in \Theta_2} l_p(\theta_2) = \max_{\theta_2 \in \Theta_2} l_p(\theta_2). \tag{2.45}$$

Proof. Let $\sup_{\theta \in \Theta} l(\theta) = l_*$. By the definition of the supremum, there exists a sequence $\theta_m = (\theta_{1m}, \theta_{2m})$ such that $l(\theta_m) \to l_*$ as $m \to \infty$. But

$$l(\theta_{1m}, \theta_{2m}) \leq l(\hat{\theta}_1(\theta_{2m}), \theta_{2m}) = l_p(\theta_{2m}) \leq l_* \tag{2.46}$$

Hence, $l_p(\theta_{2m}) \to l_*$ as $m \to \infty$, that is,

$$\sup_{\theta \in \Theta} l(\theta) = \sup_{\theta_2 \in \Theta_2} l_p(\theta_2). \tag{2.47}$$

If $l_p(\theta_2)$ is continuous, and Θ_2 is compact, then by the extreme value theorem, there exists a $\hat{\theta}_2$ such that

$$\max_{\theta_2 \in \Theta_2} l_p(\theta_2) = l_p(\hat{\theta}_2). \tag{2.48}$$

Since

$$l_p(\hat{\theta}_2) \leq \sup_{\theta_2 \in \Theta_2} l_p(\theta_2) = l_* \leq \max_{\theta_2 \in \Theta_2} l_p(\theta_2) = l_p(\hat{\theta}_2), \tag{2.49}$$

therefore,

$$\sup_{\theta_2 \in \Theta_2} l_p(\theta_2) = \max_{\theta_2 \in \Theta_2} l_p(\theta_2). \tag{2.50}$$

□

Under **REFM**$_1$ and the assumption of a given set of common factors $P_1,\ P_2,\ \ldots,\ P_q$, the restricted maximum likelihood estimate $\hat{\theta}_1(\theta_2) = (\underline{\hat{a}},\ \hat{\underline{b}}^2,\ \hat{\sigma}^2)$ is the unique solution of the likelihood equations given by the closed form equation (2.41). This means that the profile log-likelihood can be used to find the maximum likelihood estimates as long as we can find the maximum of the profile likelihood. Now, let us simplify the profile likelihood $l(\hat{\theta}_1(\theta_2), \theta_2)$ after substitution of $\hat{\theta}_1(\theta_2)$ into expression (2.26):

$$\begin{aligned}
l_p(\theta_2) &= -\frac{np}{2}\log(2\pi) - \frac{n(p-q)}{2}\log(\frac{1}{p-q}\sum_{k=q+1}^{p} P_k^t C_{yy} P_k) \\
&\quad -\frac{n}{2}\sum_{k=1}^{q}\log(P_k^t S P_k) - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{q}\frac{1}{P_k^t S P_k}(P_k^t y_i - P_k^t \bar{y})^2 \\
&\quad -\frac{1}{2}\frac{1}{(p-q)^{-1}\sum_{k=q+1}^{p}(P_k^t C_{yy} P_k)}\sum_{i=1}^{n}\sum_{k=q+1}^{p}(P_k^t y_i)^2 \\
&= -\frac{np}{2}\log(2\pi) + \frac{n(p-q)}{2}\log(p-q) - \frac{n(p-q)}{2}\log(\sum_{k=q+1}^{p} P_k^t C_{yy} P_k) \\
&\quad -\frac{n}{2}\sum_{k=1}^{q}\log(P_k^t S P_k) - \frac{1}{2}\sum_{k=1}^{q}\frac{1}{P_k^t S P_k}\sum_{i=1}^{n} P_k^t (y_i - \bar{y})^{\otimes 2} P_k \\
&\quad -\frac{p-q}{2}\frac{1}{\sum_{k=q+1}^{p}(P_k^t C_{yy} P_k)}\sum_{k=q+1}^{p}(P_k^t \sum_{i=1}^{n} y_i^{\otimes 2} P_k) \\
&= -\frac{np}{2}\log(2\pi) + \frac{n(p-q)}{2}\log(p-q) - \frac{n(p-q)}{2}\log(\sum_{k=q+1}^{p} P_k^t C_{yy} P_k) \\
&\quad -\frac{n}{2}\sum_{k=1}^{q}\log(P_k^t S P_k) - \frac{1}{2}\sum_{k=1}^{q}\frac{1}{P_k^t S P_k} n P_k^t S P_k
\end{aligned}$$

$$
\begin{aligned}
& -\frac{p-q}{2}\frac{1}{\sum_{k=q+1}^{p} P_k^t C_{yy} P_k} \sum_{k=q+1}^{p} (P_k^t n C_{yy} P_k) \\
= & -\frac{np}{2}\log(2\pi) + \frac{n(p-q)}{2}\log(p-q) - \frac{n(p-q)}{2}\log(\sum_{k=q+1}^{p} P_k^t C_{yy} P_k) \\
& -\frac{n}{2}\sum_{k=1}^{q}\log(P_k^t S P_k) - \frac{nq}{2} - \frac{n(p-q)}{2} \\
= & \frac{n}{2}\{C - (p-q)\log(\sum_{k=q+1}^{p} P_k^t C_{yy} P_k) - \sum_{k=1}^{q}\log(P_k^t S P_k)\} \qquad (2.51)
\end{aligned}
$$

where $C = -p\log(2\pi) + (p-q)\log(p-q) - p$. Since the maximizer of $l_p(\theta_2)$ is the same as the maximizer of $(2/n)l_p(\theta_2)$, we can multiply both sides of equation(2.51) by $2/n$, and find

$$
\frac{2}{n} l_p(\theta_2) = C - (p-q)\log(\sum_{k=q+1}^{p} P_k^t C_{yy} P_k) - \sum_{k=1}^{q}\log(P_k^t S P_k) \qquad (2.52)
$$

Since

$$
\begin{aligned}
\pi^t C_{yy} \pi &= (P_1, P_2, ..., P_p)^t C_{yy} (P_1, P_2, ..., P_p), \\
&= \begin{pmatrix}
P_1^t C_{yy} P_1 & P_1^t C_{yy} P_2 & ... & P_1^t C_{yy} P_q \\
P_2^t C_{yy} P_1 & P_2^t C_{yy} P_2 & ... & P_2^t C_{yy} P_q \\
... & ... & ... & ... \\
P_q^t C_{yy} P_1 & P_q^t C_{yy} P_2 & ... & P_q^t C_{yy} P_q
\end{pmatrix}, \qquad (2.53)
\end{aligned}
$$

we have

$$
\begin{aligned}
\mathrm{tr}(\pi^t C_{yy} \pi) &= P_1^t C_{yy} P_1 + P_2^t C_{yy} P_2 + ... + P_p^t C_{yy} P_p \\
&= \sum_{k=1}^{q} P_k^t C_{yy} P_k + \sum_{k=q+1}^{p} P_k^t C_{yy} P_k. \qquad (2.54)
\end{aligned}
$$

Then, since $\mathrm{tr}(\pi^t C_{yy}\pi) = \mathrm{tr}(C_{yy}\pi\pi^t) = \mathrm{tr}(C_{yy})$,

$$
\sum_{k=q+1}^{p} P_k^t C_{yy} P_k = \mathrm{tr}(C_{yy}) - \sum_{k=1}^{q} P_k^t C_{yy} P_k \qquad (2.55)
$$

Therefore, the simplified expression for profile log-likelihood is

$$\frac{2}{n}l_p(\theta_2) = C - (p-q)\log[\mathrm{tr}(C_{yy}) - \mathrm{tr}(\pi_1\pi_1^t C_{yy})] - \sum_{k=1}^{q}\log(P_k^t S P_k). \tag{2.56}$$

According to Lemma 2, the maximum of the profile likelihood function exists, that is,

$$\sup_{\theta_2\in\Theta_2} l_p(\theta_2) = \max_{\theta_2\in\Theta_2} l_p(\theta_2) \tag{2.57}$$

Observe that $l_p(\theta_2)$ given by (2.56) is a continuous function and $\Theta_2 = \{\theta_2 \in \mathcal{M} : \theta_2^t\theta_2 = I_q\}$ is a bounded closed set in $(\mathbf{R^p})^{\mathbf{q}}$. The profile likelihood does not have a closed-form analytic solution for $P_1, P_2, ..., P_q$, which is not unusual for multivariate analysis problems. We will consider numerical procedures to compute the maximum likelihood estimates in a later chapter.

2.3 Asymptotic Properties of Estimates

2.3.1 Asymptotic profile likelihood function

In the previous section we have shown that there exists a maximum likelihood estimate $\hat{\theta} = (\hat{\theta}_1,\ \hat{\theta}_2)$ under $\mathbf{REFM}_1$. We will further investigate some asymptotic properties of $\hat{\theta}$ such as consistency and asymptotic distribution. First, let us rewrite the log-likelihood function from (2.26) multiplied by $\frac{2}{n}$, as follows:

$$\begin{aligned} l_n(\theta) &= \frac{2}{n}l(\theta) \\ &= -p\log(2\pi) - (p-q)\log(\sigma^2) - \sum_{k=1}^{q}\log(b_k^2+\sigma^2) \\ &\quad -\frac{1}{\sigma^2}\sum_{k=q+1}^{p} P_k^t C_{yy} P_k - \sum_{k=1}^{q}\frac{1}{b_k^2+\sigma^2}[P_k^t S P_k + (P_k^t\bar{y} - a_k)^2] \qquad (2.58) \\ &= l_n(\theta, \bar{y}, S). \end{aligned}$$

Remark: From $l(\theta) = \frac{n}{2} l_n(\theta, \bar{y}, S)$, we have seen that the maximum likelihood estimate, $\hat{\theta}$, is a function of sufficient statistics, which was also guaranteed by the factorization criterion. □

Let $\theta_0 = (\underline{a}_0, \underline{b}_0^2, \sigma_0^2, \pi_{10})$ denote the true parameter value in Θ. When $n \to \infty$, under REFM_1 the sample mean $\bar{y}$, sample covariance matrix S, and sample second moment matrix C_{yy} converge almost surely to EY, $Var(Y)$, and $EY^{\otimes 2}$, respectively, by the Strong Law of Large Numbers (SLLN). That is, we have

$$\bar{y} \xrightarrow{a.s.} EY = \underline{\pi}_{10} \underline{a}_0, \tag{2.59}$$

$$S = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^{\otimes 2} \xrightarrow{a.s.} Var(Y) = \pi_{10} B_{10} \pi_{10}^t + \sigma_0^2 I_p, \tag{2.60}$$

and

$$\begin{aligned} C_{yy} &= \frac{1}{n} \sum_{i=1}^{n} y_i^{\otimes 2} \xrightarrow{a.s.} EY^{\otimes 2} = Var(Y) + (EY)^{\otimes 2}, \\ &= \pi_{10} B_{10} \pi_{10}^t + \sigma_0^2 I_p + (\pi_{10} \underline{a}_0)^{\otimes 2}. \end{aligned} \tag{2.61}$$

The limiting form of the normalized log-likelihood function is

$$\begin{aligned} g(\theta; \theta_0) &= \lim_{n \to \infty} l_n(\theta) \\ &= -p \log(2\pi) - (p-q) \log(\sigma^2) - \sum_{k=1}^{q} \log(b_k^2 + \sigma^2) \\ &\quad - \sum_{k=1}^{q} \frac{1}{b_k^2 + \sigma^2} [P_k^t \pi_{10} B_{10} \pi_{10}^t P_k + \sigma_0^2 + (P_k^t \pi_{10} \underline{a}_0 - a_k)^2] \\ &\quad - \frac{1}{\sigma^2} \sum_{k=q+1}^{p} (P_k^t \pi_{10} B_{10} \pi_{10}^t P_k + \sigma_0^2 + P_k^t \pi_{10} \underline{a}_0 \underline{a}_0^t \pi_{10}^t P_k). \end{aligned} \tag{2.62}$$

We next maximize the limiting form of the log-likelihood function (2.62) over $\theta_1 = (\underline{a}, \underline{b}^2, \sigma^2)$ for fixed $P_1, P_2, \ldots, P_q$. Assume that θ_2 is given, that is, we know the coordinate directions of the factor space V_1: P_1, P_2, $\ldots$, P_q. By steps similar to

those shown above in (2.27), (2.31), and (2.34), we solve the likelihood equations: $\nabla_{\theta_1} g((\theta_1, \theta_2); (\theta_{10}, \theta_{20})) = 0$, and get $\tilde{\theta}_1$ as a function of θ_2, θ_{10}, and θ_{20}. For any $k \leq q$, we find $\tilde{\theta}_1 = (\tilde{\underline{a}}, \tilde{\underline{b}}^2, \tilde{\sigma}^2)$ as a function of $\theta_2 = \pi_1$, defined by

$$\begin{cases} \tilde{a}_k = P_k^t \pi_{10} \underline{a}_0 \\ \tilde{\sigma}^2 = \frac{1}{p-q} \sum_{k=q+1}^{p} [P_k^t \pi_{10} (B_{10} + a_0 a_0^t) \pi_{10}^t P_k + \sigma_0^2] \\ \tilde{b}_k^2 = P_k^t \pi_{10} B_{10} \pi_{10}^t P_k + \sigma_0^2 - \tilde{\sigma}^2. \end{cases} \tag{2.63}$$

Substituting $\tilde{\theta}_1(\theta_2)$ for θ_1 , the asymptotic profile likelihood function is

$$\begin{aligned} g_p(\theta_2; \theta_0) &= g((\tilde{\theta}_1(\theta_2), \theta_2); (\theta_{10}, \theta_{20})) \\ &= -p \log(2\pi) - \sum_{k=1}^{q} \log(P_k^t \pi_{10} B_{10} \pi_{10}^t P_k + \sigma_0^2) \\ &\quad -(p-q) \log\{\frac{1}{p-q} \sum_{k=q+1}^{p} [P_k^t \pi_{10} (B_{10} + \underline{a}_0 \underline{a}_0{}^t) \pi_{10}^t P_k] + \sigma_0^2\} \\ &\quad - \sum_{k=1}^{q} \frac{1}{P_k^t \pi_{10} B_{10} \pi_{10}^t P_k + \sigma_0^2} (P_k^t \pi_{10} B_{10} \pi_{10}^t P_k + \sigma_0^2) \\ &\quad - \frac{p-q}{\sum_{k=q+1}^{p} [P_k^t \pi_{10} (B_{10} + \underline{a}_0 \underline{a}_0^t) \pi_{10}^t P_k + \sigma_0^2]} \\ &\quad \times \sum_{k=q+1}^{p} [P_k^t \pi_{10} (B_{10} + \underline{a}_0 \underline{a}_0^t) \pi_{10}^t P_k + \sigma_0^2] \\ &= -p \log(2\pi) + (p-q) \log(p-q) - q - (p-q) \\ &\quad -(p-q) \log\{ \sum_{k=q+1}^{p} [P_k^t \pi_{10} (B_{10} + \underline{a}_0 \underline{a}_0^t) \pi_{10}^t P_k] + (p-q)\sigma_0^2\} \\ &\quad - \sum_{k=1}^{q} \log(P_k^t \pi_{10} B_{10} \pi_{10}^t P_k + \sigma_0^2) \\ &= C - (p-q) \log\{ \sum_{k=q+1}^{p} [P_k^t \pi_{10} (B_{10} + \underline{a}_0 \underline{a}_0^t) \pi_{10}^t P_k] + (p-q)\sigma_0^2\} \\ &\quad - \sum_{k=1}^{q} \log(P_k^t \pi_{10} B_{10} \pi_{10}^t P_k + \sigma_0^2), \end{aligned} \tag{2.64}$$

where the constant $C = -p \log(2\pi) + (p-q) \log(p-q) - p$.

If we first apply the Strong Law of Large Numbers to the log-likelihood function

$l(\theta)$ under **REFM**$_1$, and then maximize θ_1 in the limiting form of normalized log-likelihood function for fixed $P_1, P_2, \ldots, P_q$, then the resulting expression (2.64) is the asymptotic profile likelihood function. Alternatively, assuming that $\theta_2 = \pi_1$ is given, we have a restricted maximum likelihood estimate $\hat{\theta}_1(\theta_2)$,and can apply the Strong Law of Large Numbers to the profile likelihood function $l_p(\theta_2)$. The following Lemma shows that these two approaches have the same result, that is, **maximization over** θ_1 and the **Strong Law of Large Numbers** are interchangeable:

$$\lim_{n\to\infty} \frac{2}{n} l_p(\theta_2) = \sup_{\theta_1 \in \Theta_1} \{ \lim_{n\to\infty} l_n(\theta_1; \theta_2) \} \tag{2.65}$$

Lemma 3 *Under* **REFM**$_1$*, and when* $n \to \infty$*,* $\bar{y}$*,* C_{yy}*, and* S *satisfy (2.59), (2.60), and (2.61). Let* $l_p(\theta_2,\ \bar{y},\ S)$ *be the profile likelihood function, and let* $\tilde{\theta}_1(\theta_2)$ *maximize the restricted asymptotic likelihood for fixed* θ_2*. That is,*

$$g(\tilde{\theta}_1(\theta_2),\ \theta_2;\ \theta_0) = \sup_{\theta_1 \in \Theta_1} \{ \lim_{n\to\infty} l_n(\theta_1; \theta_2) \} \tag{2.66}$$

Then, when $n \to \infty$*,*

$$\frac{2}{n} l_p(\theta_2) \xrightarrow{a.s.} g(\tilde{\theta}_1(\theta_2), \theta_2, \theta_0). \tag{2.67}$$

To prove the Lemma, we apply the Strong Law of Large Numbers under **REFM**$_1$, using expressions (2.59), (2.60), and (2.61). When n goes to ∞, we have

$$\begin{aligned} P_k^t S P_k &\xrightarrow{a.s.} P_k^t(\pi_{10} B_{10} \pi_{10}^t + \sigma_0^2 I_p) P_k \\ &= P_k^t \pi_{10} B_{10} \pi_{10}^t P_k + \sigma_0^2, \end{aligned} \tag{2.68}$$

$$
\begin{aligned}
\mathrm{tr}(C_{yy}) &\xrightarrow{a.s.} \mathrm{tr}(\pi_{10}B_{10}\pi_{10}^t + \sigma_0^2 I_p + (\pi_{10}\underline{a}_0)^{\otimes 2}) \\
&= \mathrm{tr}(\pi_{10}B_{10}\pi_{10}^t) + p\sigma_0^2 + \mathrm{tr}(\pi_{10}\underline{a}_0\underline{a}_0^t\pi_{10}^t) \\
&= \mathrm{tr}(B_{10}\pi_{10}^t\pi_{10}) + p\sigma_0^2 + \mathrm{tr}(\underline{a}_0\underline{a}_0^t\pi_{10}^t\pi_{10}) \\
&= \mathrm{tr}(B_{10}) + p\sigma_0^2 + \mathrm{tr}(\underline{a}_0\underline{a}_0^t) \\
&= \sum_{i=1}^{q} b_{i0}^2 + p\sigma_0^2 + \mathrm{tr}(\underline{a}_0^t\underline{a}_0) \\
&= \sum_{i=1}^{q} b_{i0}^2 + p\sigma_0^2 + \sum_{i=1}^{q} a_{i0}^2 \\
&= \sum_{i=1}^{q} (a_{i0}^2 + b_{i0}^2) + p\sigma_0^2, \qquad (2.69)
\end{aligned}
$$

and

$$
\begin{aligned}
P_k^t C_{yy} P_k &\xrightarrow{a.s.} P_k^t(\pi_{10}B_{10}\pi_{10}^t + \sigma_0^2 I_p + (\pi_{10}\underline{a}_0)^{\otimes 2})P_k \\
&= P_k^t\pi_{10}B_{10}\pi_{10}^t P_k + \sigma_0^2 + P_k^t\pi_{10}\underline{a}_0\underline{a}_0^t\pi_{10}^t P_k \\
&= P_k^t\pi_{10}(B_{10} + \underline{a}_0\underline{a}_0^t)\pi_{10}P_k + \sigma_0^2. \qquad (2.70)
\end{aligned}
$$

After substituting (2.68), (2.69), and (2.70) into the profile likelihood function (2.51), and simplifying the expression (2.56), we immediately have the expression (2.64). This completes the proof. □

Our next objective is to prove that there exists a unique local maximum for the asymptotic profile log-likelihood function $g_p(\theta_2, \theta_0)$. The unique local maximum means that there exists a point $\theta_2^* \in \Theta_2$ and a sufficiently small neighborhood $\mathcal{N}$ of θ_2^* such that $g_p(\theta_2^*, \theta_0) > g_p(\theta_2, \theta_0)$, where θ_2 is any given point in $\mathcal{N}$. Let $\{P_{i0} : i = 1,\ 2,\ \ldots,\ q\}$ be the true orthonormal factor directions, and let $\pi_{10} = (P_{10},\ P_{20},\ \ldots,\ P_{q0})$ be the corresponding $p \times q$ matrix. We will change our variables from π_1 to a matrix T by mapping: $\pi_1 \longrightarrow T = \pi_1^t\pi_{10}$. The entries $t_{ij} : 1 \le i,\ j \le q$

of T are inner products of the columns P_i of π_1 and the jth direction P_{j0} of the true factor space $V_{10} = span\{P_{10},\ P_{20},\ \ldots,\ P_{q0}\}$, that is, $t_{ij} \equiv P_i^t P_{j0}$, for all $1 \leq i, j \leq q$. Obviously, for given i and j, the value of t_{ij} must be between -1 and +1, that is, $-1 \leq t_{ij} \leq 1$. Now

$$\begin{aligned}
\sum_{k=1}^{q} t_{ki}^2 &= \sum_{k=1}^{q} (t_{ij})^2 \\
&= \sum_{k=1}^{q} P_{i0}^t (P_k P_k^t) P_{i0} \\
&= P_{i0}^t (I_p - \sum_{k=q+1}^{p} (P_k P_k^t)) P_{i0} \\
&= 1 - \sum_{k=q+1}^{p} (t_{ki})^2 \\
&\leq 1, \qquad (2.71)
\end{aligned}$$

and the equality holds if and only if $col(\pi_1) = span\{P_{10}, P_{20}, \ldots, P_{q0}\}$. Similarly, we can show that $\sum_{i=1}^{q} t_{ki}^2 \leq 1$. Thus, $(t_{ki}^2)_{q\times q}$, is a doubly substochastic matrix since the sum of elements in each row and column is less than or equal to 1.

Remark: When $q = p$, the factor space V_1 is the space $V = R^p$. Thus, $\sum_{i=1}^{q} P_i P_i^t = \sum_{i=1}^{q} P_{i0} P_{i0}^t = I_q$. Therefore, $\sum_{k=1}^{q} t_{ki}^2 = 1$, and $\sum_{i=1}^{q} t_{ki}^2 = 1$. The matrix $(t_{ki}^2)_{q\times q}$, is a doubly stochastic matrix in this case. □

We see that the parameter θ_2 enters expression (2.64) only through $\theta_2^t \theta_{20}$. Hence we can write the asymptotic profile log-likelihood function $g_p(\theta_2; \theta_{10}, \theta_{20})$ as $\tilde{g}_p(\theta_2^t \theta_{20}, \theta_{10})$. After changing the variables, the asymptotic profile log-likelihood function $\tilde{g}_p(T, \theta_{10})$ by (2.64) is

$$\tilde{g}_p(T, \theta_{10}) = \lim_{n\to\infty} \frac{2}{n} l_p(\theta_2)$$

$$
\begin{aligned}
&= C-(p-q)\log[\sum_{i=1}^{q}(a_{i0}^2+b_{i0}^2)+(p-q)\sigma_0^2-\sum_{k=1}^{q}\sum_{i=1}^{q}b_{i0}^2t_{ki}^2 \\
&\quad -\sum_{k=1}^{q}(\sum_{i=1}^{q}a_{i0}t_{ki})^2]-\sum_{k=1}^{q}\log(\sum_{i=1}^{q}b_{i0}^2t_{ki}^2+\sigma_0^2), \qquad (2.72)
\end{aligned}
$$

since

$$
\begin{aligned}
P_k^t\pi_{10}B_{10}\pi_{10}^tP_k &= (t_{k1},t_{k2},...,t_{kq})B_{10}\begin{pmatrix} t_{k1} \\ t_{k2} \\ ... \\ t_{kq} \end{pmatrix} \\
&= \sum_{i=1}^{q}b_{i0}^2t_{ki}^2, \qquad (2.73)
\end{aligned}
$$

and

$$
\begin{aligned}
&\sum_{k=q+1}^{p}P_k^t[\pi_{10}(B_{10}+\underline{a}_0\underline{a}_0^t)\pi_{10}^t]P_k \\
&= \mathrm{tr}[\pi_{10}(B_{10}+\underline{a}_0\underline{a}_0^t)\pi_{10}^t]-\sum_{k=1}^{q}P_k^t[\pi_{10}(B_{10}+\underline{a}_0\underline{a}_0^t)\pi_{10}^t]P_k \\
&= \mathrm{tr}[(B_{10}+\underline{a}_0\underline{a}_0^t)\pi_{10}^t\pi_{10}]-\sum_{k=1}^{q}P_k^t\pi_{10}B_{10}\pi_{10}^tP_k \\
&\quad -\sum_{k=1}^{q}P_k^t\pi_{10}\underline{a}_0\underline{a}_0^t\pi_{10}^tP_k \\
&= \mathrm{tr}(B_{10})+\mathrm{tr}(\underline{a}_0\underline{a}_0^t)-\sum_{k=1}^{q}\sum_{i=1}^{q}b_{i0}^2t_{ki}^2-\sum_{k=1}^{q}(\underline{a}_0{}^t\pi_{10}^tP_k)^2 \\
&= \sum_{i=1}^{q}(a_{i0}^2+b_{i0}^2)-\sum_{k=1}^{q}\sum_{i=1}^{q}b_{i0}^2t_{ki}^2-\sum_{k=1}^{q}(\sum_{i=1}^{q}a_{i0}t_{ki})^2. \qquad (2.74)
\end{aligned}
$$

2.3.2 Special case q=1

Before we prove that in general there exists a locally unique maximum of the asymptotic profile log-likelihood function $\tilde{g}_p(T,\theta_{10})$, we explore $g_p(T,\theta_{10})$ in the special case when q=1. Now, T is just a scalar variable t_{11}. The asymptotic profile log-likelihood

function is, from (2.72)

$$\begin{aligned}\tilde{g}_p(T,\theta_{10}) &= -p\log(2\pi)+(p-q)\log(p-q)-p-\log(b_{10}^2t_{11}^2+\sigma_0^2)\\ &\quad -(p-1)\log\{(a_{10}^2+b_{10}^2)+(p-1)\sigma_0^2-b_{10}^2t_{11}^2-a_{10}^2t_{11}^2\}\\ &= -p\log(2\pi)+(p-q)\log(p-q)-p-\log(b_{10}^2t_{11}^2+\sigma_0^2)\\ &\quad -(p-1)\log\{(a_{10}^2+b_{10}^2)(1-t_{11}^2)+(p-1)\sigma_0^2\}. \qquad (2.75)\end{aligned}$$

Lemma 4 *Let the asymptotic profile log-likelihood function* $\tilde{g}_p(s) \equiv C - \log(b_{10}^2 s + \sigma_0^2) - (p-1)\log\{(a_{10}^2+b_{10}^2)(1-s)+(p-1)\sigma_0^2\}, 0 \le s \le 1$. *The point* $s=1$ *is the unique maximizer of* $\tilde{g}_p(s)$.

Proof: Taking the derivative of $\tilde{g}_p(s)$ with respective to s, and simplifying the expression, we have

$$\begin{aligned}\frac{d\tilde{g}_p(s)}{ds} &= -\frac{b_{10}^2}{b_{10}^2 s+\sigma_0^2}-(p-1)\frac{-(a_{10}^2+b_{10}^2)}{(a_{10}^2+b_{10}^2)(1-s)+(p-1)\sigma_0^2}\\ &= \frac{p(a_{10}^2+b_{10}^2)b_{10}^2 s+(p-1)a_{10}^2\sigma_0^2-b_{10}^2(a_{10}^2+b_{10}^2)}{(b_{10}^2 s+\sigma_0^2)[(a_{10}^2+b_{10}^2)(1-s)+(p-1)\sigma_0^2]}. \qquad (2.76)\end{aligned}$$

Setting the above expression equal to 0 yields

$$p(a_{10}^2+b_{10}^2)b_{10}^2 s = b_{10}^2(a_{10}^2+b_{10}^2)-(p-1)a_{10}^2\sigma_0^2. \qquad (2.77)$$

The root of this equation is

$$\begin{aligned}\hat{s} &= \frac{b_{10}^2(a_{10}^2+b_{10}^2)-(p-1)a_{10}^2\sigma_0^2}{p(a_{10}^2+b_{10}^2)b_{10}^2}\\ &= \frac{1}{p}-\frac{(p-1)a_{10}^2\sigma_0^2}{p(a_{10}^2+b_{10}^2)b_{10}^2}. \qquad (2.78)\end{aligned}$$

Clearly, $\hat{s} \le \frac{1}{p}$. If $\hat{s} < 0$, then $d\tilde{g}_p(s)/ds > 0$ for all $0 \le s \le 1$. This means that the asymptotic profile log-likelihood function $\tilde{g}_p(s)$ is strictly increasing over $s \in [0,\ 1]$. Hence, the maximum of $\tilde{g}_p(s)$ will be on the boundary, at $s = 1$.

If $\hat{s} \in [0,\ 1]$, then $\hat{s} \leq \frac{1}{p}$ from (2.78) , with equality holding only if $a_{10} = 0$. Normally, p is a large number, thus $\hat{s}$ is close to 0 rather than to 1. If $a_{10} \neq 0$, that is, $\hat{s} < \frac{1}{p}$, we can check the sign of the first derivative of $\tilde{g}_p(s)$ at the point $s = \frac{1}{p}$ as follows:

$$\frac{d\tilde{g}_p(s)}{ds}|_{s=\frac{1}{p}} = \frac{(p-1)a_{10}^2\sigma_0^2}{2(b_{10}^2\frac{1}{p} + \sigma_0^2)[(a_{10}^2 + b_{10}^2)(1-\frac{1}{p}) + (p-1)\sigma_0^2]} > 0. \tag{2.79}$$

The asymptotic profile log-likelihood function $\tilde{g}_p(s)$ is a linear combination of two logarithm functions. It is a smooth function. Also, it has only one extreme point because of the unique solution from the first derivative equation. Hence, the sign of $d\tilde{g}_p(s)/ds$ is the same for any point s with $s > \hat{s}$. The expression (2.79) indicates that $d\tilde{g}_p(s)/ds > 0$ if $s > \hat{s}$, that is, $\tilde{g}_p(s)$ is strictly increasing if $s > \hat{s}$. The extreme point $\hat{s}$ minimizes the asymptotic profile log-likelihood function $\tilde{g}_p(s)$. Hence, the maximum point of $\tilde{g}_p(s)$ must occur at the boundary, either $s = 1$ or $s = 0$. At $s = 0$, $\tilde{g}_p(0) = -\log\sigma_0^2 - (p-1)\log(a_{10}^2 + b_{10}^2 + (p-1)\sigma_0^2)$, and at $s = 1$, $\tilde{g}_p(1) = -\log(b_{10}^2 + \sigma_0^2 - (p-1)\log((p-1)\sigma_0^2)$. Then

$$\tilde{g}_p(1) - \tilde{g}_p(0) = (p-1)\log\left(1 + \frac{a_{10}^2 + b_{10}^2}{(p-1)\sigma_0^2}\right) - \log(1 + \frac{b_{10}^2}{\sigma_0^2}).$$

Let $x = b_{10}^2/\sigma_{10}^2$. Then since $p \geq 2$,

$$\begin{aligned}
\tilde{g}_p(1) - \tilde{g}_p(0) &= \int_1^{p-1} \frac{\partial}{\partial\alpha}\left\{\alpha\log(1 + \frac{x}{\alpha})\right\} d\alpha \\
&= \int_1^{p-1} \left\{-\log(1 - \frac{x}{x+\alpha}) - \frac{x}{x+\alpha}\right\} d\alpha \\
&> 0.
\end{aligned}$$

Thus, we have $\tilde{g}_p(1) > \tilde{g}_p(0)$, that is, $\tilde{g}_p(s)$ has its maximum value when $s = 1$.

If $a_{10} = 0$, the extreme point $\hat{s}$ of $\tilde{g}_p(s)$ is $1/p$. Again, we check the sign of the first derivative of $\tilde{g}_p(s)$ at a point to the right of $\hat{s}$, that is, at $s = 1$. Since

$$\frac{d\tilde{g}_p(s)}{ds}|_{s=1} = \frac{(p-1)[a_{10}^2 b_{10}^2 + \sigma_0^2(a_{10}^2 + b_{10}^2)}{2(b_{10}^2 \frac{1}{p} + \sigma_0^2)[(a_{10}^2 + b_{10}^2)(1 - \frac{1}{p}) + (p-1)\sigma_0^2]} > 0. \tag{2.80}$$

We can again conclude that the extreme point $\hat{s}$ is the minimum, and the maximum of $\tilde{g}_p(s)$ occurs when $s = 1$. This completes the proof of the Lemma. □

By Lemma 4, in this special case, $q = 1$, the asymptotic profile likelihood under REFM_1 has not only a unique local maximum when $P_1 = P_{10}$ ($\pi_1 = \pi_{10}$) but also a global maximum. This Lemma also suggests that the maximum may occur at $P_k = P_{k0},\ k = 1, 2, \ldots, q$, in the general case.

2.3.3 Unique local maximum of asymptotic profile log-likelihood

In order to show that the maximum likelihood estimate $\hat{\theta}$ is a consistent estimate of the true parameter θ_0 under $\mathbf{REFM}_1$, we will first prove that there exists a unique local maximum of $\tilde{g}_p(T, \theta_{10})$. Define $\mathbf{T} = \{T = (t_{ki})_{q\times q} : \sum_{k=1}^{q} t_{ki}^2 \leq 1$ and $\sum_{i=1}^{q} t_{ki}^2 \leq 1\}$. The existence of a unique local maximum means there exists a $T^* \in \mathbf{T}$ with $\tilde{g}_p(T^*) > \tilde{g}_p(T)$, where $T \in \mathbf{T}$ is any point sufficiently near T^*, but not T^*. The following Lemma proves that the identity matrix I_q is a unique local maximum point of the asymptotic profile log-likelihood function $\tilde{g}_p(T, \theta_{10})$.

Lemma 5 *Let $\tilde{g}_p(T, \theta_{10})$ in the expression (2.72) be the asymptotic profile log-likelihood function under $\mathbf{REFM}_1$ with $t_{ij} = P_i^t P_{j0}$, for all $1 \leq i, j \leq q$. The identity matrix I_q is a unique local maximizer in $\mathbf{T}$ of $\tilde{g}_p(T, \theta_{10})$.*

Proof: Let $T \in \mathbf{T}$ be any point sufficiently near the identity matrix I_q, so that with $\xi_{ij} \equiv t_{ij}^2 - \delta_{ij}, \max_{i,j} |\xi_{ij}| \leq c$ for a small constant c to be chosen below. We need to show

$$\tilde{g}_p(I_q, \theta_{10}) - \tilde{g}_p(T, \theta_{10}) > 0. \tag{2.81}$$

We have ξ_{ij} close to zero, positive when $i \neq j$, and negative when $i = j$. Now we write $\tilde{g}_p(T, \theta_{10})$ from the expression (2.72) at points $I_q,\ T$:

$$\tilde{g}_p(I_q, \theta_{10}) \;=\; C - \sum_{k=1}^{q} \log\left(b_{k0}^2 + \sigma_0^2\right) - (p-q)\log[(p-q)\sigma_0^2] \tag{2.82}$$

and

$$\begin{aligned}\tilde{g}_p(T, \theta_{10}) \;=\; & C - \sum_{k=1}^{q} \log\left(\sum_{i=1}^{q} b_{i0}^2 t_{ki}^2 + \sigma_0^2\right) - (p-q)\log\{\sum_{i=1}^{q}(a_{i0}^2 + b_{i0}^2) \\ & +(p-q)\sigma_0^2 - \sum_{k=1}^{q}\sum_{i=1}^{q} b_{i0}^2 t_{ki}^2 - \sum_{k=1}^{q}(\sum_{i=1}^{q} a_{i0} t_{ki})^2\}. \end{aligned} \tag{2.83}$$

The next four steps will simplify expressions (2.82) and (2.83).

Step (I): Applying the inequality $|\log(1+x) - x| \leq x^2$ valid for all $|x| < 1$, we have

$$\begin{aligned}\log(\sum_{i=1}^{q} b_{i0}^2 t_{ki}^2 + \sigma_0^2) \;&=\; \log(b_{k0}^2 + \sigma_0^2 + \sum_{i=1}^{q} b_{i0}^2 \xi_{ki}) \\ &=\; \log(b_{k0}^2 + \sigma_0^2) + \frac{\sum_{i=1}^{q} b_{i0}^2 \xi_{ki}}{b_{k0}^2 + \sigma_0^2} + \alpha_1 \max_{k,i} |\xi_{ki}|^2 / c^2 \end{aligned}$$

for some number $\alpha_1 \in (-1, 1)$, whenever $\max_{k,i} |\xi_{ki}| \leq c \leq (\sigma_0^2 + \min_k b_{k0}^2) / \sum_{i=1}^{q} b_{i0}^2$.

Step (II): By the Cauchy-Schwarz inequality, for $T \in \mathbf{T}$,

$$\begin{aligned}\sum_{k=1}^{q}\left(\sum_{i=1}^{q} a_{i0} t_{ki}\right)^2 \;&=\; \sum_{k=1}^{q}\sum_{i=1}^{q}\sum_{j=1}^{q} a_{i0} t_{ki} a_{j0} t_{kj} \\ &\leq\; \sum_{i=1}^{q}\sum_{j=1}^{q}(\sum_{k=1}^{q} t_{ki}^2 \sum_{k=1}^{q} t_{kj}^2)^{\frac{1}{2}} a_{i0} a_{j0} \\ &=\; \left(\sum_{i=1}^{q} a_{i0}(\sum_{k=1}^{q} t_{ki}^2)^{\frac{1}{2}}\right)^2 \end{aligned}$$

$$\leq \sum_{i=1}^{q} a_{i0}^2 \sum_{k=1}^{q} t_{ki}^2$$
$$\leq \sum_{i=1}^{q} a_{i0}^2$$

Therefore

$$\sum_{i=1}^{q} a_{i0}^2 \geq \sum_{k=1}^{q} (\sum_{i=1}^{q} a_{i0} t_{ki})^2. \tag{2.84}$$

Step (III): Since $y = \log(x)$ is an increasing function on x, by the result in step (II), we have

$$\begin{aligned}
&\log\{\sum_{i=1}^{q}(a_{i0}^2 + b_{i0}^2) + (p-q)\sigma_0^2 - \sum_{k=1}^{q}\sum_{i=1}^{q} b_{i0}^2 t_{ki}^2 - \sum_{k=1}^{q}(\sum_{i=1}^{q} a_{i0} t_{ki})^2\} \\
&= \log\{\sum_{i=1}^{q} b_{i0}^2 + (p-q)\sigma_0^2 - \sum_{k=1}^{q}\sum_{i=1}^{q} b_{i0}^2 t_{ki}^2 + \sum_{i=1}^{q} a_{i0}^2 - \sum_{k=1}^{q}(\sum_{i=1}^{q} a_{i0} t_{ki})^2\} \\
&\geq \log\{\sum_{i=1}^{q} b_{i0}^2 + (p-q)\sigma_0^2 - \sum_{k=1}^{q}\sum_{i=1}^{q} b_{i0}^2 t_{ki}^2\}.
\end{aligned} \tag{2.85}$$

Step (IV): Again, applying the result $|\log(1+x) - x| \leq x^2$ when $|x| < 1$, we have

$$\begin{aligned}
&\log\{\sum_{i=1}^{q} b_{i0}^2 + (p-q)\sigma_0^2 - \sum_{k=1}^{q}\sum_{i=1}^{q} b_{i0}^2 t_{ki}^2\} \\
&= \log\{(p-q)\sigma_0^2 + \sum_{i=1}^{q}\sum_{k=1}^{q} b_{i0}^2 \xi_{ki}\} \\
&= \log[(p-q)\sigma_0^2] + \frac{\sum_{i=1}^{q}\sum_{k=1}^{q} b_{i0}^2 \xi_{ki}}{(p-q)\sigma_0^2} + \alpha_2 \max_{k,i} |\xi_{ki}|^2 / c^2
\end{aligned}$$

for some number $\alpha_2 \in (-1, 1)$, whenever $\max_{k,i} |\xi_{ki}| \leq c \leq \frac{p-q}{q}\sigma_0^2 / \sum_{i=1}^{q} b_{i0}^2$.

Combining steps (I), (II), (III), and (IV), the left-hand side of expression (2.81) becomes, for $\max_{k,i} |\xi_{ki}| \leq c \leq \min(1, \frac{p-q}{q})\sigma_0^2 / \sum_{i=1}^{n} b_{i0}^2$,

$$\begin{aligned}
&\tilde{g}_p(I_q, \theta_{10}) - \tilde{g}_p(T, \theta_{10}) \\
&= -(p-q)\log[(p-q)\sigma_0^2] - \sum_{k=1}^{q} \log(b_{k0}^2 + \sigma_0^2)
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{k=1}^{q}\left\{\log\left(b_{k0}^2+\sigma_0^2\right)+\frac{\sum_{i=1}^{q}b_{i0}^2\xi_{ki}}{b_{k0}^2+\sigma_0^2}-\alpha_1\max_{k,i}|\frac{\xi_{ki}}{c}|^2\right\}\\
&+(p-q)\log\left\{\sum_{i=1}^{q}b_{i0}^2+(p-q)\sigma_0^2-\sum_{k=1}^{q}\sum_{i=1}^{q}b_{i0}^2t_{ki}^2+\sum_{i=1}^{q}a_{i0}^2-\sum_{k=1}^{q}(\sum_{i=1}^{q}a_{i0}t_{ki})^2\right\}\\
&\geq -(p-q)\log[(p-q)\sigma_0^2]+\sum_{k=1}^{q}\frac{\sum_{i=1}^{q}b_{i0}^2\xi_{ki}}{b_{k0}^2+\sigma_0^2}\\
&+(p-q)\log\{(p-q)\sigma_0^2-\sum_{k=1}^{q}\sum_{i=1}^{q}b_{i0}^2\xi_{ki}\}-\alpha_1 q\max_{k,i}|\frac{\xi_{ki}}{c}|^2\\
&=-(p-q)\log[(p-q)\sigma_0^2]+\sum_{k=1}^{q}\frac{\sum_{i=1}^{q}b_{i0}^2\xi_{ki}}{b_{k0}^2+\sigma_0^2}\\
&+(p-q)\log[(p-q)\sigma_0^2]-(p-q)\frac{\sum_{i=1}^{q}\sum_{k=1}^{q}b_{i0}^2\xi_{ki}}{(p-q)\sigma_0^2}\\
&+(\alpha_2(p-q)-q\alpha_1)\max_{k,i}|\frac{\xi_{ki}}{c}|^2\\
&=\sum_{k=1}^{q}\sum_{i=1}^{q}[\frac{1}{b_{k0}^2+\sigma_0^2}-\frac{1}{\sigma_0^2}]b_{i0}^2\xi_{ki}+(\alpha_2(p-q)-q\alpha_1)\max_{k,i}|\frac{\xi_{ki}}{c}|^2\\
&=-\frac{1}{\sigma_0^2}\sum_{k=1}^{q}\sum_{i=1}^{q}\frac{b_{k0}^2b_{i0}^2}{b_{k0}^2+\sigma_0^2}\xi_{ki}+(\alpha_2(p-q)-q\alpha_1)\max_{k,i}|\frac{\xi_{ki}}{c}|^2. \qquad (2.86)
\end{aligned}
$$

Define $J_1=\sum_{k=1}^{q}\sum_{i=1}^{q}\frac{b_{k0}^2b_{i0}^2}{b_{k0}^2+\sigma_0^2}\xi_{ki}$. Recall the identifiability condition of ordered $\{b_{i0}^2\}_{i=1}^{q}$. If we can show that $J_1<0$ for all sufficiently small c and $0<\max_{k,i}|\xi_{ki}|\leq c$, then we will have established (2.81). Rewrite J_1 as follows:

$$
\begin{aligned}
J_1 &= \sum_{k=1}^{q}\sum_{i=1}^{q}\frac{b_{k0}^2b_{i0}^2}{b_{k0}^2+\sigma_0^2}(t_{ki}^2-\delta_{ki}) \qquad (2.87)\\
&= \sum_{k=1}^{q}\sum_{i=1}^{q}\frac{b_{k0}^2b_{i0}^2}{b_{k0}^2+\sigma_0^2}(P_k^tP_{i0})^2-\sum_{k=1}^{q}\frac{b_{k0}^4}{b_{k0}^2+\sigma_0^2}.
\end{aligned}
$$

Consider now some special cases in expression (2.87). If $b_{i0}=b,\quad i=1,2,...,q$, then we have

$$
\begin{aligned}
J_1 &= \sum_{i=1}^{q}\sum_{k=1}^{q}\frac{b^2b^2}{b^2+\sigma_0^2}P_{i0}^tP_kP_k^tP_{i0}-\sum_{k=1}^{q}\frac{b^4}{b^2+\sigma_0^2}\\
&= \frac{b^4}{b^2+\sigma_0^2}\{\sum_{i=1}^{q}P_{i0}^t(\sum_{k=1}^{q}P_kP_k^t)P_{i0}-q\}\\
&= \frac{b^4}{b^2+\sigma_0^2}\{\sum_{i=1}^{q}P_{i0}^t[I_p-\sum_{k=q+1}^{p}P_kP_k^t]P_{i0}-q\}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{b^4}{b^2+\sigma_0^2}\{\sum_{i=1}^{q} 1 - \sum_{i=1}^{q} P_{i0}^t(\sum_{k=q+1}^{p} P_k P_k^t)P_{i0} - q\} \\
&= -\frac{b^4}{b^2+\sigma_0^2}\cdot\sum_{i=1}^{q}\sum_{k=q+1}^{p}(P_k^t P_{i0})^2 \\
&\leq 0 \qquad (2.88)
\end{aligned}
$$

In this case, $\tilde{g}_p(I_q,\theta_{10}) - \tilde{g}_p(T,\theta_{10}) \geq 0$ for all T, and $\tilde{g}_p(\cdot,\theta_{10})$ has a maximum point at I_q. The equality holds only if $V_1 = span\{P_{10}, P_{20}, \ldots, P_{q0}\}$. We use another Lemma to show the expression (2.87) is negative more generally.

Lemma 6 *For all sequences* $\{b_{i0}^2, i = 1,2,...,q : b_{10}^2 > b_{20}^2 > \ldots > b_{q0}^2 > 0\}$, *and* $\sigma_0^2 > 0$, *the expression* J_1 *defined in (2.87) is negative, that is,*

$$
\sum_{i=1}^{q}\sum_{k=1}^{q}\frac{b_{k0}^2 b_{i0}^2}{b_{k0}^2+\sigma_0^2}(t_{ki})^2 < \sum_{k=1}^{q}\frac{b_{k0}^4}{b_{k0}^2+\sigma_0^2}. \qquad (2.89)
$$

Proof. Let $b_{(q+1)0}^2 = ... = b_{p0}^2 = 0$. The expression (2.89) will not change if we replace q with p. So it is actually the same to prove (2.89) with $q = p$, where without loss of generality $\sum_{i=1}^{q} t_{ki}^2 = 1$ for all k, and $\sum_{i=1}^{q} t_{ki}^2 = 1$ for all i. This relationship implies that $T = (t_{ij})_{q\times q}$ is an orthogonal matrix, which is the same thing as saying that the matrix C with entries $C_{ij} = t_{ij}^2$ is a doubly stochastic matrix. Let us start at J_1 equal to the left-hand side of (2.89) minus the right-hand side of (2.89):

$$
\begin{aligned}
J_1 &= \sum_{k=1}^{q}\sum_{i=1}^{q}\left(1 - \frac{\sigma_0^2}{b_{k0}^2+\sigma_0^2}\right) b_{i0}^2(t_{ki})^2 - \sum_{k=1}^{q}\left[b_{k0}^2 - \sigma_0^2 + \frac{\sigma_0^4}{b_{k0}^2+\sigma_0^2}\right] \\
&= \sum_{k=1}^{q}\sum_{i=1}^{q} b_{i0}^2(t_{ki})^2 - \sigma_0^2\sum_{k=1}^{q}\sum_{i=1}^{q}\frac{b_{i0}^2+\sigma_0^2-\sigma_0^2}{b_{k0}^2+\sigma_0^2}t_{ki}^2
\end{aligned}
$$

$$
\begin{aligned}
& -\sum_{k=1}^{q} b_{k0}^2 + q\sigma_0^2 - \sigma_0^4 \sum_{k=1}^{q} \frac{1}{b_{k0}^2+\sigma_0^2} \\
= & \sum_{k=1}^{q}\sum_{i=1}^{q} b_{i0}^2 (t_{ki})^2 - \sigma_0^2 \sum_{k=1}^{q}\sum_{i=1}^{q} \frac{b_{i0}^2+\sigma_0^2}{b_{k0}^2+\sigma_0^2} t_{ki}^2 \\
& +\sigma_0^4 \sum_{k=1}^{q}\sum_{i=1}^{q} \frac{(t_{ki})^2}{b_{k0}^2+\sigma_0^2} - \sum_{k=1}^{q}\sum_{i=1}^{q} b_{i0}^2 \delta_{ki} + q\sigma_0^2 - \sigma_0^4 \sum_{k=1}^{q} \frac{1}{b_{k0}^2+\sigma_0^2} \\
= & \sigma_0^2 \left(q - \sum_{k=1}^{q}\sum_{i=1}^{q} \frac{b_{i0}^2+\sigma_0^2}{b_{k0}^2+\sigma_0^2} t_{ki}^2 \right) \\
& -\sum_{k=1}^{q}\sum_{i=1}^{q} \left(b_{i0}^2 + \frac{\sigma_0^4}{b_{k0}^2+\sigma_0^2} \right) (\delta_{ki} - t_{ki}^2). \qquad (2.90)
\end{aligned}
$$

Using the fact that $q = \sum_{k=1}^{q} 1 = \sum_{k=1}^{q}\sum_{i=1}^{q} \delta_{ki}$, we can further simplify:

$$
\begin{aligned}
J_1 = & \sigma_0^2 (\sum_{k=1}^{q}\sum_{i=1}^{q} \delta_{ki} - \sum_{k=1}^{q}\sum_{i=1}^{q} \frac{b_{i0}^2+\sigma_0^2}{b_{k0}^2+\sigma_0^2} t_{ki}^2) \\
& -\sum_{k=1}^{q}\sum_{i=1}^{q} (b_{i0}^2 + \frac{\sigma_0^4}{b_{k0}^2+\sigma_0^2})(\delta_{ki} - t_{ki}^2) \\
= & \sigma_0^2 \sum_{k=1}^{q}\sum_{i=1}^{q} \frac{b_{i0}^2+\sigma_0^2}{b_{k0}^2+\sigma_0^2} (\delta_{ki} - t_{ki}^2) \\
& -\sum_{k=1}^{q} \frac{\sigma_0^4}{b_{k0}^2+\sigma_0^2} \sum_{i=1}^{q} (\delta_{ki} - t_{ki}^2) - \sum_{i=1}^{q} b_{i0}^2 \sum_{k=1}^{q} (\delta_{ki} - t_{ki}^2) \\
= & \sigma_0^2 (q - \sum_{k=1}^{q}\sum_{i=1}^{q} \frac{b_{i0}^2+\sigma_0^2}{b_{k0}^2+\sigma_0^2} t_{ki}^2) \qquad (2.91)
\end{aligned}
$$

where in last line of (2.91) we have used $\sum_i t_{ki}^2 = 1$, and $\sum_k t_{ki}^2 = 1$ (valid when $q = p$). Therefore, the expression (2.89) holds if

$$
\sum_{k=1}^{q}\sum_{i=1}^{q} \frac{b_{i0}^2+\sigma_0^2}{b_{k0}^2+\sigma_0^2} t_{ki}^2 > q. \qquad (2.92)
$$

Define $w_i = b_{i0}^2 + \sigma_0^2$. Obviously, $w_i > 0$ for all $1 \le i \le q$. The next Lemma will prove inequality (2.92) with $q = p$, and complete all Lemmas. The proof of Lemma 7 is based on a decomposition into closed irreducible classes of a Markov chain. First, we need to eliminate the possibility of transient states. Luckily, the doubly stochastic matrix guarantees there are no transient states.

Remark: Finite state Markov chains with doubly stochastic transition matrices have no transient states.

Proof: Let P be the transition probability matrix with finite state space $S = \{1,\ 2,\ \ldots,\ a\}$. If P is doubly stochastic, then P^n is also a doubly stochastic matrix for all $n > 1$. If there exists any transient state, say j, then $p_{ij}^n \to 0$ as $n \to \infty$ for all $i \in S$ (Karlin and Taylor 1975, Theorem 7.1, p72). Since $\sum_{i=1}^a p_{ij}^n = 1$, then $1 = \sum_{i=1}^a p_{ij}^n \to 0$. This is a contradiction. Therefore, there are no transient states. □

Lemma 7 *Suppose that C is any doubly stochastic matrix, which could therefore serve as the transition matrix for a Markov chain. The states $S = \{1, \ldots, q\}$ for such a Markov chain can be partitioned uniquely into closed irreducible classes $S_1, \ldots, S_r$ of states (with $S_j \cap S_k = \emptyset$ for $j \neq k$, and $\cup_{j=1}^r S_j = S$). Then*

$$\inf_{w_1, \ldots, w_q > 0} \sum_{k=1}^{q} \sum_{i=1}^{q} \frac{w_i}{w_k} C_{ki} \ = \ q \tag{2.93}$$

and the infimum is attained precisely for the set of $\mathbf{w} \equiv (w_1, \ldots, w_q) \in (\mathbf{R}^+)^q$ such that $w_i = w_k$ whenever $i, k \in S_a$ for some value a.

Proof. Since

$$\sum_{k=1}^{q} \sum_{i=1}^{q} \frac{w_i}{w_k} C_{ki} \ = \ \sum_{a=1}^{r} \sum_{k \in S_a} \sum_{i \in S_a} \frac{w_i}{w_k} C_{ki} \tag{2.94}$$

by definition of the decomposition of the states of a finite Markov chain into closed irreducible classes, there is no loss of generality in restricting attention to the transition submatrices of C corresponding to single closed classes, or equivalently in assuming that C itself is irreducible.

Next, since the double summation $G(\mathbf{w}) = \sum_{k,i} (w_i/w_k) C_{ki}$ is unaffected by replacing w_i with $w_i / \sum_{k=1}^{q} w_k$, we can restrict attention, without loss of generality, to probability vectors $\mathbf{w}$ in the bounded (but not closed) region $\Delta \equiv \{\mathbf{w} : w_i > 0, \sum_i w_i = 1\}$. Moreover, it is easily verified that $G(\mathbf{w})$ approaches $+\infty$ as $\mathbf{w}$ approaches the boundary of Δ, or equivalently, as at least one component of $\mathbf{w}$ approaches 0. Thus, for small $\epsilon > 0$, the continuous function G restricted to the compact region $\{\mathbf{w} \in \Delta : \min_i w_i \geq \epsilon\}$ has no minimizing values located on the boundary, but must have at least one minimizing value.

Note that if $C = I_q$, then

$$\sum_{k=1}^{q} \sum_{i=1}^{q} \frac{w_i}{w_k} C_{ki} = \sum_{k=1}^{q} \sum_{i=1}^{q} \frac{w_i}{w_k} \delta_{ki} = \sum_{k=1}^{q} \frac{w_k}{w_k} = q. \tag{2.95}$$

In this case, (2.93) holds with equality for all $w \in (R^+)^q$. Taking the first derivative with respect to w_j, $j = 1, ..., q$, we have

$$\frac{\partial G(w)}{\partial w_j} = \sum_{k=1}^{q} (\frac{1}{w_k} C_{kj} - \frac{w_k}{w_j^2} C_{jk}). \tag{2.96}$$

Setting $\partial G(w)/\partial w_j = 0$ for all j, we have

$$\sum_{k=1}^{q} \frac{1}{w_k} C_{kj} = \sum_{k=1}^{q} \frac{w_k}{w_j^2} C_{jk}, \quad j = 1, 2, ..., q. \tag{2.97}$$

Multiplying both sides of (2.97) by w_j, leads to

$$\sum_{k=1}^{q} \frac{w_j}{w_k} C_{kj} = \sum_{k=1}^{q} \frac{w_k}{w_j} C_{jk} \tag{2.98}$$

If w_j is the smallest of all of the w's, then

$$\sum_{k=1}^{q} \frac{w_j}{w_k} C_{kj} \leq \sum_{k=1}^{q} 1 \cdot C_{kj} = 1, \tag{2.99}$$

and

$$\sum_{k=1}^{q} \frac{w_k}{w_j} C_{jk} \geq \sum_{k=1}^{q} 1 \cdot C_{jk} = 1 \tag{2.100}$$

Then the left hand side of (2.98) is less than the right hand side of (2.98) unless $w_j = w_k$ for all (j,k) such that $C_{jk} > 0$. Thus, the irreducibility of C and the calculations already performed say that the unique point $\mathbf{w} \in \Delta$ where $\nabla_w G(w) = 0$ is the point $\mathbf{w} = (1/q, 1/q, \ldots, 1/q)$. This point must be the unique minimizer on Δ of G, as was to be shown. □

Remark: In case the matrix C can be taken symmetric, C is a doubly stochastic matrix, and the doubly stochastic matrix guarantees there are no transient states, in the Markov chain of Lemma 7. The proof could be made much simpler. Indeed, by the Cauchy-Schwarz inequality, whether or not C is symmetric,

$$q^2 \;=\; \left(\sum_{i=1}^{q} \sum_{k=1}^{q} C_{ik} \right)^2 \;\leq\; \left(\sum_{i=1}^{q} \sum_{k=1}^{q} \frac{w_k}{w_i} C_{ik} \right) \left(\sum_{i=1}^{q} \sum_{k=1}^{q} \frac{w_i}{w_k} C_{ik} \right) \tag{2.101}$$

with equality if and only if the doubly-indexed arrays $(w_i/w_k \;:\; C_{ik} > 0)$ and $(w_k/w_i \;:\; C_{ik} > 0)$ are proportional, and (using the irreducibility of the Markov chain with transition matrix C) this happens if and only if $\mathbf{w} = (1/q, 1/q, \ldots, 1/q)$. Moreover, when C is symmetric, the right-hand side of (2.101) is precisely $G(\mathbf{w})^2$. In the case of symmetric C, this short argument is the complete proof. □

Next, in the setting of Lemma 6, we apply Lemma 7 with q replaced by p, to the matrix $C : C_{ki} = t_{ki}^2$ (now, doubly stochastic), with $w_i = b_{i0}^2 + \sigma_0^2$, and $b_{i0}^2 \equiv 0$ for $q+1 \leq i \leq p$, $b_{10}^2 > b_{20}^2 > \ldots > b_{q0}^2 > 0$. First, we use the fact that the

minimizer is at $w_i \equiv w$ to conclude (2.92) with $\geq$. Now, if $T \neq I_q$, but $T \approx I_q$, then there exists a closed irreducible class of C, $C_{ki} \equiv t^2_{ki}$, which contains a non-singleton set of states including at one state i of $\{1, 2, \ldots, q\}$. If $j \in \{1,\ 2,\ \ldots,\ q\}\backslash\{i\}$ is the other state, then $b^2_{i0} \neq b^2_{j0}$, and $w_i \neq w_j$. By Lemma 7, this implies strict inequality in (2.92). This completes the proof of Lemma 5, that is, the asymptotic profile log-likelihood function $\tilde{g}_p(T, \theta_{10})$ has a unique local maximizer $\hat{T} = I_q$; equivalently, $g_p(\theta_2; \theta_{10}, \theta_{20})$ attains a locally unique maximum when $P_i = P_{i0}$ for $i = 1,\ 2,\ \ldots,\ q$. □

2.3.4 Consistent estimator

The idea to prove the consistent estimator of the parameter comes from the article by Andersen and Gill (1982). We restate their result as follows.

Lemma 8 *(Andersen and Gill 1982, Corollary II.2., p1116) Let E be an open subset of R^p and let $F_1, F_2, \ldots$ be a sequence of random functions on E such that $\forall x \in E, F_n(x) \xrightarrow{P} f(x)$ as $n \to \infty$ where f is some real function on E. Suppose f has a unique maximum at $\hat{x} \in E$. Let $\hat{X}_n$ maximize F_n. Then $\hat{X}_n \xrightarrow{P} \hat{x}$ as $n \to \infty$.*

The profile likelihood $l_p \equiv F_n$ is a random concave function (for large n) of parameter $\theta_2 = x$ based on a data sample of size n, since $2n^{-1}\ \nabla^{\otimes 2}_{\theta_2} l_p(\theta_2) \xrightarrow{a.s} \nabla^{\otimes 2}_{\theta_2} g_p(\theta_2; \theta_0)$ uniformly in a small neighborhood of θ_{20} as $n \to \infty$, and $\nabla^{\otimes 2}_{\theta_2} g_p(\theta_2; \theta_0)$ is negative definite (refer forward to section 2.4.1). And the asymptotic profile likelihood $g_p = f$ is a nonrandom function. Here E must be the local neighborhood in Θ_2 on which the maximum is unique. Now, from the existence of a unique local

maximizer of the asymptotic profile likelihood function we have proved and Lemma 8, we have $\hat{\theta}_2 \to \theta_{20}$. That is, $\hat{\theta}_2$ is a consistent estimator of θ_{20}. The next Lemma will show that there exists a unique maximizer of the asymptotic log-likelihood function under $\mathbf{REFM}_1$. Again, we have $\hat{\theta} \to \theta_0$. That is, $\hat{\theta}$ is a consistent estimator of θ_0.

Lemma 9 *Let $g(\theta;\theta_0)$ be the limiting form of the log-likelihood function in the expression (2.62), and let $g_p(\theta_2;\theta_0)$ the asymptotic profile likelihood function in the expression (2.64). If $g_p(\theta_2;\theta_0)$ has a locally unique maximum, then $g(\theta;\theta_0)$ also has a locally unique maximum.*

Proof: From equations (2.63), we can see that $\tilde{\theta}_1(\theta_2,\theta_{10},\theta_{20})$ is a unique maximizer for $g(\theta_1,\theta_2;\theta_{10},\theta_{20})$. Lemma 5 shows that the asymptotic profile likelihood function $g_p(\theta_2;\theta_0)$ has a locally unique maximum at $\theta_2 = \theta_{20}$, and $g(\tilde{\theta}_1(\theta_2),\theta_2;\theta_{10},\theta_{20}) = g_p(\theta_2;\theta_{10},\theta_{20})$. Also, we can verify that $\tilde{\theta}_1(\theta_{20}) = \theta_{10}$ holds by substituting θ_{20} into the right hand side of equations (2.63). Therefore, we have

$$\begin{aligned} g(\theta;\theta_0) &= g(\theta_1,\theta_2;\theta_{10},\theta_{20}) \\ &\leq \max_{\theta_1\in\Theta_1} g(\theta_1,\theta_2;\theta_{10},\theta_{20}) \\ &= g(\tilde{\theta}_1(\theta_2),\theta_2;\theta_{10},\theta_{20}) \\ &= g_p(\theta_2;\theta_{10},\theta_{20}) \\ &\leq g_p(\theta_{20};\theta_{10},\theta_{20}) \\ &= g(\theta_{10},\theta_{20};\theta_{10},\theta_{20}) \\ &= g(\theta_0;\theta_0). \end{aligned} \tag{2.102}$$

The equality holds only if $\theta = \theta_0$. We conclude that $g(\theta; \theta_0)$ also has a unique local maximum at $\theta = \theta_0$. □

2.4 Calculus Maximization

2.4.1 Calculus maximization

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. So far, we have proved there exists a maximizer $\hat{\theta}_2$ for the profile log-likelihood function $l_p(\theta_2,\ \bar{y},\ C_{yy})$, and the maximum value of $l_p(\theta_2,\ \bar{y},\ C_{yy})$ is the same as the maximum of log-likelihood function $l(\theta)$. Moreover, in the following Theorem we establish that the combination of the maximizer $\hat{\theta}_2$ and the restricted maximum likelihood estimate $\hat{\theta}_1(\theta_2)$ provides the calculus maximum of log-likelihood function $l(\theta)$. As in most multivariate analysis problems, the profile log-likelihood does not have a closed-form analytic solution for $P_1, P_2, ..., P_q$. That is, the profile log-likelihood equations can not be solved directly. We will look at a numerical procedure to compute the maximum likelihood estimates iteratively. There are various iterative procedures for finding a maximum of the likelihood function, including the steepest descent method, the Newton-Raphson method, and the EM (expectation-maximization) algorithm. First, let us define what is meant by a **calculus maximum**.

Definition 1 *Let $l(\theta)$ be a smooth function with continuous second derivatives. If $\nabla_\theta l(\theta) = 0$ when $\theta = \hat{\theta}$, and the Hessian matrix $\nabla_\theta^{\otimes 2} l(\theta)|_{\theta=\hat{\theta}}$ is negative definite,*

then we call $\hat{\theta}$ *a* **calculus maximum** *of* $l(\theta)$.

Theorem 1 *Let* $\theta_1 \in R^a, \theta_2 \in R^b$, *and let* $l(\theta_1, \theta_2)$ *be a function with continuous second partial derivatives. Assume that* $\nabla_{\theta_1} l(\hat{\theta}_1(\theta_2), \theta_2) = 0$, *where* $\hat{\theta}_1(\theta_2)$ *is continuously differentiable, and the Hessian matrix* $\nabla^{\otimes 2}_{\theta_1} l(\hat{\theta}_1(\theta_2), \theta_2)$ *is negative definite. Also assume that* $D_{\theta_2} l(\hat{\theta}_1(\hat{\theta}_2), \hat{\theta}_2) = 0$, *and the Hessian matrix* $D^{\otimes 2}_{\theta_2} l(\hat{\theta}_1(\hat{\theta}_2), \hat{\theta}_2)$ *is a negative definite matrix, where* D_{θ_2} *denotes total differentiation with respect to* θ_2. *Then the point* $(\hat{\theta}_1(\hat{\theta}_2), \hat{\theta}_2)$ *is a calculus maximum of the function* $l(\theta_1, \theta_2)$.

Proof: Write the Hessian matrix of $l(\hat{\theta}_1, \hat{\theta}_2)$:

$$H = \nabla^{\otimes 2}_{\theta} l(\hat{\theta}_1, \hat{\theta}_2) = \begin{pmatrix} H_{11} & H_{12} \\ \\ H_{21} & H_{22} \end{pmatrix}. \tag{2.103}$$

Then H is a symmetric matrix, and $H^t_{21} = H_{12}, H^t_{11} = H_{11}$, and $H^t_{22} = H_{22}$. By hypothesis $\nabla_{\theta_1} l(\hat{\theta}_1(\theta_2), \theta_2) = 0$, and taking the total derivative in this equation with respect to θ_2, we have

$$\nabla_{\theta_2} \nabla^t_{\theta_1} l(\hat{\theta}_1(\theta_2), \theta_2) + \nabla_{\theta_2} (\hat{\theta}^t_1(\theta_2))^t \nabla^{\otimes 2}_{\theta_1} l(\hat{\theta}_1(\theta_2), \theta_2) = 0. \tag{2.104}$$

Letting $\theta_2 = \hat{\theta}_2$, we can re-write (2.104) as

$$H_{21} + \nabla_{\theta_2} \hat{\theta}^t_1(\theta_2) H_{11} = 0. \tag{2.105}$$

Thus,

$$\nabla_{\theta_2} \hat{\theta}^t_1(\theta_2) = -H_{21} H^{-1}_{11}. \tag{2.106}$$

Let $h(\theta_2) = l(\hat{\theta}_1(\theta_2), \theta_2)$, and take the first order derivative with respect to θ_2,

$$\nabla_{\theta_2} h(\theta_2) = D_{\theta_2} l(\hat{\theta}_1(\theta_2), \theta_2). \tag{2.107}$$

When $\theta_2 = \hat{\theta}_2$, (2.107) yields $\nabla_{\theta_2} h(\theta_2) = 0$. Taking the derivative of $\nabla_{\theta_2} h(\theta_2)$ with respect to θ_2, we have the following equations,

$$\nabla_{\theta_2}^{\otimes 2} h(\theta_2) = \nabla_{\theta_2} \hat{\theta}_1^t(\theta_2) \cdot \nabla_{\theta_1} \nabla_{\theta_2}^t l(\hat{\theta}_1(\theta_2), \theta_2) + \nabla_{\theta_2}^{\otimes 2} l(\hat{\theta}_1(\theta_2), \theta_2) \tag{2.108}$$

Replace θ_2 with $\hat{\theta}_2$ in (2.108), also using (2.106), to obtain

$$\begin{aligned} \nabla_{\theta_2}^{\otimes 2} h(\hat{\theta}_2) &= \nabla_{\theta_2} \hat{\theta}_1^t(\hat{\theta}_2) H_{12} + H_{22} \\ &= -H_{21} H_{11}^{-1} H_{12} + H_{22} \end{aligned} \tag{2.109}$$

To prove that H is negative definite, we need to show that for any given $v = \binom{v_1}{v_2}$, $v^t H v < 0$. Now, for given any v, we have

$$\begin{aligned} v^t H v &= (v_1^t \; v_2^t) \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \\ &= v_1^t H_{11} v_1 + v_2^t (H_{22} - H_{21} H_{11}^{-1} H_{12}) v_2 \\ &\quad + 2 v_1^t H_{12} v_2 + v_2^t H_{21} H_{11}^{-1} H_{12} v_2 \\ &= (H_{12} v_2 + H_{11} v_1)^t H_{11}^{-1} (H_{12} v_2 + H_{11} v_2) \\ &\quad + v_2^t \{H_{22} - H_{21} H_{11}^{-1} H_{12}\} v_2. \end{aligned} \tag{2.110}$$

Since the Hessian matrix $\nabla_{\theta_1}^{\otimes 2} l(\hat{\theta}_1(\theta_2), \theta_2)$ is assumed negative definite, so is $\nabla_{\theta_1}^{\otimes 2} l(\hat{\theta}_1(\hat{\theta}_2), \hat{\theta}_2)$, that is, H_{11} is negative definite. Thus, H_{11}^{-1} is a negative definite matrix, that is, for any given $u \in R^a$, $u^t H_{11}^{-1} u < 0$. Also, by hypothesis $\nabla_{\theta_2}^{\otimes 2} h(\theta_2)$ is a negative definite matrix, that is, $-H_{21} H_{11} H_{12} + H_{22}$ is negative definite from equation (2.109), so that for any given $v \in R^b$, $v^t(-H_{21} H_{11}^{-1} H_{12} + H_{22}) v < 0$. Therefore, $v^t H v < 0$, that is, H is negative definite, and $l(\theta_1, \theta_2)$ has a calculus maximum at $(\hat{\theta}_1(\hat{\theta}_2), \hat{\theta}_2)$. □

From (2.41), we have shown that the restricted maximum likelihood estimate $\hat{\theta}_1(\theta_2)$ is a smooth function of π_1, the sample mean $\bar{y}$, and the sample variance S under the assumption that the common factor directions $P_1,\ P_2,\ \ldots,\ P_q$ are given. From Lemma 2, we know there is a maximizer $\hat{\theta}_2$ of the profile log-likelihood function $l_p(\theta_2)$. In order to conclude that $\hat{\theta} = (\hat{\theta}_1(\hat{\theta}_2),\ \hat{\theta}_2)$ is the calculus maximum likelihood estimate of $l(\theta)$ from Theorem 1, we must verify the following conditions:

1. The Hessian matrix $\nabla^{\otimes 2}_{\theta_1} l(\theta)|_{\theta_1=\hat{\theta}_1(\theta_2)}$ is negative definite.

2. The Hessian matrix $\nabla^{\otimes 2}_{\theta_2} l_p(\theta_2)|_{\theta_2=\hat{\theta}_2}$ is negative definite.

First, we verify 1. From the expression (2.27), by taking the derivative of $\partial l(\theta)/\partial a_k$ with respect to $a_i,\ b_i^2$ and σ^2, respectively. For $1 \le i \le q$, we have

$$\frac{\partial l^2(\theta)}{\partial a_i \partial a_k} = -\frac{n}{b_k^2+\sigma^2}\delta_{ik}, \tag{2.111}$$

$$\frac{\partial l^2(\theta)}{\partial b_i^2 \partial a_k} = \left(-\frac{n}{(b_k^2+\sigma^2)^2}(P_k^t\bar{y} - a_k)\right)\delta_{ik} \tag{2.112}$$

and

$$\frac{\partial l^2(\theta)}{\partial \sigma^2 \partial a_k} = -\frac{n}{(b_k^2+\sigma^2)^2}(P_k^t\bar{y} - a_k). \tag{2.113}$$

From expression (2.31), we take the derivative of $\partial l(\theta)/\partial b_k^2$ with respect to $a_i,\ b_i^2$ and σ^2, respectively, for $1 \le i \le q$, yielding (2.112) along with

$$\frac{\partial l^2(\theta)}{\partial b_i^2 \partial b_k^2} = \left(\frac{n}{2}\frac{1}{(b_k^2+\sigma^2)^2} - \frac{1}{(b_k^2+\sigma^2)^3}\ \sum_{i=1}^{n}(P_k^t y_i - a_k)^2\right)\delta_{ik} \tag{2.114}$$

and

$$\frac{\partial l^2(\theta)}{\partial \sigma^2 \partial b_k^2} = \frac{n}{2}\frac{1}{(b_k^2+\sigma^2)^2} - \frac{1}{(b_k^2+\sigma^2)^3}\sum_{i=1}^{n}(P_k^t y_i - a_k)^2. \tag{2.115}$$

From expression (2.34), we take the derivative of $\partial l(\theta)/\partial\sigma^2$ with respect to a_i, b_i^2 and σ^2, respectively, yielding (2.113), (2.115), and

$$\begin{aligned}\frac{\partial l^2(\theta)}{\partial\sigma^2\partial\sigma^2} &= \frac{n(p-q)}{2}\frac{1}{(\sigma^2)^2} - \sum_{k=1}^{q}\frac{1}{(b_k^2+\sigma^2)^3}\sum_{i=1}^{n}(P_k^t y_i - a_k)^2 \\ &\quad + \frac{n}{2}\sum_{k=1}^{q}\frac{1}{(b_k^2+\sigma^2)^2} - \frac{1}{(\sigma^2)^3}\sum_{i=1}^{n}\sum_{k=q+1}^{p}(P_k^t y_i)^2. \end{aligned} \tag{2.116}$$

By substituting (2.29), (2.33), and (2.37) into all elements in the second order derivative matrix H_{11}, and simplifying, the Hessian matrix becomes

$$\nabla_{\theta_1}^{\otimes 2} l(\theta)|_{\theta_1=\hat{\theta}_1(\theta_2)} = \begin{pmatrix} h_{11} & 0 & 0 \\ 0 & h_{22} & h_{23} \\ 0 & h_{32} & h_{33} \end{pmatrix}, \tag{2.117}$$

where $h_{11} = \nabla_{\underline{a}}^{\otimes 2} l(\theta)|_{\theta_1=\hat{\theta}_1(\theta_2)} = \text{Diag}\{ -\frac{n}{P_1^t S P_1}, -\frac{n}{P_2^t S P_2}, \ldots, -\frac{n}{P_q^t S P_q}\}$, $h_{22} = \nabla_{\underline{b}^2}^{\otimes 2} l(\theta)|_{\theta_1=\hat{\theta}_1(\theta_2)} = \text{Diag}\{ -\frac{n}{2(P_1^t S P_1)^2}, -\frac{n}{2(P_2^t S P_2)^2}, \ldots, -\frac{n}{2(P_q^t S P_q)^2}\}$, $h_{32} = (-\frac{n}{2(P_1^t S P_1)^2}, -\frac{n}{2(P_2^t S P_2)^2}, \ldots, -\frac{n}{2(P_q^t S P_q)^2})$, $h_{23} = h_{32}^t$, and $h_{33} = \nabla_{\sigma^2}^{\otimes 2} l(\theta)|_{\theta_1=\hat{\theta}_1(\theta_2)} = -\frac{n}{2}\sum_{k=1}^{q}\frac{1}{(P_k^t S P_k)^2} - \frac{n(p-q)^3}{2(\sum_{k=q+1}^{p} P_k^t C_{yy} P_k)^2}$.

It is easy to prove that $-\nabla_{\theta_1}^{\otimes 2} l(\theta)|_{\theta_1=\hat{\theta}_1(\theta_2)}$ is positive definite by checking the positivity of all leading minor determinants of $-\nabla_{\theta_1}^{\otimes 2} l(\theta)|_{\theta_1=\hat{\theta}_1(\theta_2)}$. Hence, we conclude that the Hessian matrix $\nabla_{\underline{\theta}_1}^{\otimes 2} l(\theta)|_{\theta_1=\hat{\theta}_1(\theta_2)}$ is a negative definite matrix, by noting that (2.117) is equivalent by row- and column- operations to the diagonal matrix with h_{23} and h_{32} replaced by 0, and with h_{33} replaced by $h_{33}^* = -n(p-q)^3/(2(\sum_{k=q+1}^{p} P_k^t C_{yy} P_k)^2)$. □

The Hessian matrix $\nabla_{\theta_2}^{\otimes 2} l_p(\theta_2)|_{\theta_2=\hat{\theta}_2}$ is too complicated to derive analytically because the $p \times q$ orthonormal matrix θ_2 contains only $pq - \frac{q(q+1)}{2}$ functionally independent parameters. We try to prove that $\nabla_{\theta_2}^{\otimes 2} l_p(\theta_2)|_{\theta_2=\hat{\theta}_2}$ is negative definite

without computing all first and second order derivatives of the profile likelihood function $l_p(\theta_2)$. The idea here is to find a strictly concave quadratic function, which bounds $\tilde{g}_p(T, \theta_{10})$ from above on a small neighborhood of θ_{20}, that is,

$$\tilde{g}_p(I_q, \theta_{10}) - \tilde{g}_p(T, \theta_{10}) \geq \gamma \|I_q - T\|^2, \tag{2.118}$$

where θ_2 lies in a small neighborhood of θ_{20} in $\mathcal{M}$, and $T \equiv \theta_2^t \theta_{20}$, with $\gamma > 0$.

Lemma 10 *Let $\tilde{g}_p(T, \theta_{10})$ from (2.72) be the asymptotic profile log-likelihood function. If $T \neq I_q$, but T lies in a sufficiently small neighborhood of I_q with respect to the matrix norm $\|M\| = (\sum_i \sum_j M_{ij}^2)^{\frac{1}{2}}$, then we have*

$$\tilde{g}_p(I_q, \theta_{10}) - \tilde{g}_p(T, \theta_{10}) \geq \gamma \|I_q - T\|^2, \tag{2.119}$$

where γ is a positive number.

Proof: From (2.86), we have for $\xi_{ki} \equiv t_{ki}^2 - \delta_{ki}$ satisfying $\max_{k,i} |\xi_{k,i}| \leq c \leq \frac{p-q}{q} \sigma_0^2 / \sum_{i=1}^q b_{i0}^2$,

$$\tilde{g}_p(I_q, \theta_{10}) - \tilde{g}_p(T, \theta_{10}) \geq -\frac{1}{\sigma_0^2} \sum_{k=1}^q \sum_{i=1}^q \frac{b_{k0}^2 b_{i0}^2}{b_{k0}^2 + \sigma_0^2} (t_{ki}^2 - \delta_{ki}) + \rho \alpha_3 \max_{k,i} \left| \frac{\xi_{ki}}{c} \right|^2, \tag{2.120}$$

where α_3 is some number in $(-1,\ 1)$. Define J as equal to the first term in the right hand side of (2.120), that is ,

$$J = -\frac{1}{\sigma_0^2} \sum_{k=1}^q \sum_{i=1}^q \frac{b_{k0}^2 b_{i0}^2}{b_{k0}^2 + \sigma_0^2} (t_{ki}^2 - \delta_{ki}). \tag{2.121}$$

By Lemma 6, $J_1 < 0$. Since $J = -J_1/\sigma_0^2$, we have $J > 0$. Now, rewrite J in terms of ξ_{ki}:,

$$J = -\frac{1}{\sigma_0^2} \sum_{k=1}^q \sum_{i=1}^q \frac{b_{k0}^2 b_{i0}^2}{b_{k0}^2 + \sigma_0^2} \xi_{ki},$$

that is, J is a linear combination of $\{\xi_{ki}, 1 \leq i, k \leq q\}$. Dividing both sides of (2.120) by $\sum_{k=1}^{q}\sum_{i=1}^{q}|\xi_{ki}|$, we have

$$\frac{\tilde{g}_p(I_q,\theta_{10}) - \tilde{g}_p(T,\theta_{10})}{\sum_{k=1}^{q}\sum_{i=1}^{q}|\xi_{ki}|} \geq \frac{J}{\sum_{k=1}^{q}\sum_{i=1}^{q}|\xi_{ki}|} + \frac{c^{'}\max_{k,i}|\xi_{ki}|^2}{\sum_{k=1}^{q}\sum_{i=1}^{q}|\xi_{ki}|}. \tag{2.122}$$

As $\max_{k,i}|\xi_{ki}| \to 0$, the second term on the right hand of (2.122) goes to zero since the numerator is quadratic in $\underline{\xi} = (\xi_{ki})$ while the denominator is linear. Since $J/\sum_{k=1}^{q}\sum_{i=1}^{q}|\xi_{ki}|$ is continuous on the surface of a small $\underline{\xi}$ ball and strictly positive, we conclude that there exists $\gamma > 0$ and a sufficiently small positive number c such that for all ξ_{ki} satisfying $\max|\xi_{k,i}| \leq c$,

$$J/\sum_{k=1}^{q}\sum_{i=1}^{q}|\xi_{ki}| \geq \gamma$$

Therefore, (2.120) becomes

$$\tilde{g}_p(I_q,\theta_{10}) - \tilde{g}_p(T,\theta_{10}) \geq \gamma\sum_{k=1}^{q}\sum_{i=1}^{q}|\xi_{ki}|. \tag{2.123}$$

Using the fact $\sqrt{1+\xi_{ki}} - 1 \leq |\xi_{ki}|$ when $|\xi_{ki}|$ is sufficiently small, we find

$$\begin{aligned}
\|T-I\|^2 &= \sum_{k=1}^{q}\sum_{i=1}^{q}\{t_{kj}^2(1-\delta_{ki}) + (t_{ki}-1)^2\delta_{ki}\} \\
&= \sum_{k=1}^{q}\sum_{i=1}^{q}\{|\xi_{ki}|(1-\delta_{ki}) + (\sqrt{1+\xi_{ki}}-1)^2\delta_{ki}\} \\
&\leq \sum_{k=1}^{q}\sum_{i=1}^{q}\{|\xi_{ki}|(1-\delta_{ki}) + |\xi_{ki}|^2\delta_{ki}\} \\
&\leq \sum_{k=1}^{q}\sum_{i=1}^{q}|\xi_{ki}|,
\end{aligned} \tag{2.124}$$

for small $\max_{k,i}|\xi_{k,i}|$. Combining (2.123) and (2.124), we have

$$\tilde{g}_p(I_q,\theta_{10}) - \tilde{g}_p(T,\theta_{10}) \geq \gamma\|I_q - T\|^2$$

where γ is a positive number. □

The Θ_2 can be smoothly coordinatized (locally, near I_q) with $pq - q(q+1)/2$ unconstrained real parameters and Lemma 10 shows that $g_p(\theta_2 \ ; \ \theta_0)$ is bounded above by a strictly concave quadratic function of these parameters. Therefore, the Hessian matrix $\nabla^{\otimes 2}_{\theta_2} g_p(\theta_2 \ ; \ \theta_0)$ is negative definite. Moreover, $I_p(\hat{\theta}_2)$ is positive definite.

2.4.2 Asymptotic distributions of the estimators

We now develop the asymptotic distribution theory for the maximum likelihood estimates $\hat{\theta} = (\hat{\theta}_1(\hat{\theta}_2), \ \hat{\theta}_2)$, $\hat{\theta}_2$, and $\hat{\theta}_1(\theta_2)$ under REFM_1. Let θ_2^* is a b-dimensional unconstrained parameter of θ_2, where $b = pq - q(q+1)/2$ since there are $pq - q(q+1)/2$ functionally independent elements in θ_2. Thus, the overall unconstrained parameter $\theta^* = (\theta_1, \theta_2^*)$. We express that $\hat{\theta}^*$ is the maximum likelihood estimate of $l(\theta^*)$ and θ_0^* is the true parameter of θ^*.

From Lemma 10, we can conclude that the Hessian matrix $\nabla^{\otimes 2}_{\theta_2} g_p(\theta_2 \ ; \ \theta_0)$ is negative definite since $g_p(\theta_2 \ ; \ \theta_0)$ is bounded above by a strictly concave quadratic function of these parameters . Thus the Hessian matrix $\nabla^{\otimes 2}_{\theta_2} g_p(\theta_2 \ ; \ \theta_0)$ is negative definite and the information matrix $I(\theta_2) = -\nabla^{\otimes 2}_{\theta_2} g_p(\theta_2 \ ; \ \theta_0)|_{\theta_2 = \theta_{20}}$ is positive negative. Also, $I(\theta^*)$ is positive definite when the information matrix expressed in terms of free unconstrained real parameters. From Section 2.3.4, $\hat{\theta}^*$ is the locally unique maximum likelihood estimate of the likelihood function $l(\theta^*)$ and $\hat{\theta}^* \xrightarrow{P} \theta_0^*$ as $n \to \infty$. Now we can apply finite dimensional MLE theory since the regularity conditions are clearly satisfied here. Therefore, $\sqrt{n}(\hat{\theta}^* - \theta_0^*)$ has the limiting distribution $N_s(0, \ \Sigma)$, where $s = 2q + 1 + pq - q(q+1)/2$ (Lehmann 1991, Theorem

4.1, and Cox & Hinkley 1974, Section 9.2). The covariance matrix of the limiting distribution of $\sqrt{n}(\hat{\theta}^* - \theta_0^*)$ is given by the inverse of the Fisher information matrix (Cox & Hinkley 1974), that is, $\Sigma = -\left(\nabla_{\theta^*}^{\otimes 2} g_p(\theta^* \ ; \ \theta_0^*)\right)^{-1} |_{\theta=\theta_0^*}$ and

$$\Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \\ \Sigma_{21} & \Sigma_2 \end{pmatrix}, \tag{2.125}$$

where $\Sigma_{21} = \Sigma_{12}^t$. But, Σ is very complicated to derive. The unknown parameters are replaced by their consistent estimators when we compute MLE's and their standard errors from data.

The asymptotic normal distribution of $\hat{\theta}^*$ is equivalent to the joint asymptotic normality of $\hat{\theta}_1$ and $\hat{\theta}_2^*$. Then the marginal distribution also is a normal distribution. That is, $\sqrt{n}(\hat{\theta}_1 - \hat{\theta}_{10})$ has the limiting normal distribution $N_a(0, \ \Sigma_1)$ where $a = 2q+1$, $\Sigma_1 = -(H_{11} - H_{12}H_{22}^{-1}H_{21})^{-1}$, and $\sqrt{n}(\hat{\theta}_2^* - \hat{\theta}_{20}^*)$ has the limiting normal distribution $N_b(0, \ \Sigma_2)$ where $b = pq - q(q+1)/2$, $\Sigma_2 = -(H_{22} - H_{21}H_{11}^{-1}H_{12})^{-1}$, where $H_{11} = \nabla_{\theta_1}^{\otimes 2} g(\theta^* \ ; \ \theta_0)|_{\theta_1(\theta_2^*)=\theta_{10}}$, $H_{22} = \nabla_{\theta_2^*}^{\otimes 2} g_p(\theta_2^* \ ; \ \theta_0)|_{\hat{\theta}_2^*=\theta_{20}^*}$ and $H_{12} = \nabla_{\theta_1}\nabla_{\theta_2^*}^t g_p(\tilde{\theta}_1(\theta_2^*) \ ; \theta_0^*)|_{\theta_2^*=\theta_{20}^*}$.

Chapter 3

Other Random Effect Factor Models

3.1 Common Principal Components

3.1.1 Relationships among several covariance matrices

Before we discuss the generalization of Random Effect Factor Model (REFM) to several groups, we first introduce Flury 1984 generalization of principal component analysis (PCA) to s groups, called common principal component analysis (CPCA). In section 1.2, we have seen that the most important parameter to find in Principal Components (PCA) is the covariance matrix: Principal Components allow a simplified description of the covariance and correlation structures. So, we now investigate some basic ideas of relationship among covariance matrices $\Sigma_1,\ \Sigma_2,\ \ldots,\ \Sigma_s$ of dimension $p \times p$, assuming that all the covariance matrices Σ_i for different groups i are positive definite. By the eigen-decomposition (or singular value decomposition) theorem, for any group i, $Var(Y_i) = \Sigma_i \equiv \beta^{(i)}\Lambda_i\beta^{(i)t}$ with $\beta^{(i)}$ a $p \times p$ orthonormal matrix and diagonal Λ_i.

Case 1: All Σ_i are equal.

This is the previous one-group case. All repeated observations are from a single distribution. Then, the principal components can be found from the pooled data as in Chapter 2. The number of variance parameters is $p(p+1)/2$ in this case, which decreases to $pq - q(q-1)/2$ when q principal components are used.

Case 2: Proportionality of all Σ_i, that is,

$$\Sigma_i = \rho_i \Sigma_1, \quad i = 2, \ldots, s \tag{3.1}$$

for some positive constants $\rho_1, \rho_2, \ldots, \rho_s$.

This model is called the proportional model. It is also an offshoot of the CPC model, by imposing the constraints on $\Lambda_i \equiv Diag(\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{ip})$ that

$$\lambda_{ij} = \rho_i \lambda_{1j}, \quad i = 2, \ldots, s; \quad j = 1, 2, \ldots, p. \tag{3.2}$$

For simplicity we omit the first index of diagonal elements of Λ_1, that is, we let

$$\Lambda = \Lambda_1 = Diag(\lambda_1, \ldots, \lambda_p) \tag{3.3}$$

and the constraints in (3.2) are $\lambda_{ij} = \rho_i \lambda_j$. The number of parameters is $p(p+1)/2$ for Σ_1 plus $(s-1)$ for these ρ_i. There are $pq - q(q-1)/2 + s - 1$ number of parameters when we use q principal components.

Case 3: The CPC model

$$\Sigma_i = \beta \Lambda_i \beta^t, \quad i = 1, 2, \ldots, s, \tag{3.4}$$

where β is an orthonormal $p \times p$ matrix not depending on i, and

$$\Lambda_i = Diag(\lambda_{i1}, \ldots, \lambda_{ip}). \tag{3.5}$$

The number of parameters in this case is $p(p-1)/2$ (for the orthonormal matrix β) plus sp (for the diagonal matrices Λ_i). The CPC model can also be written as

$$\Sigma_i = \lambda_{i1} \beta_1 \beta_1^t + \lambda_{i2} \beta_2 \beta_2^t + \ldots + \lambda_{ip} \beta_p \beta_p^t \quad i = 1, 2, \ldots, s, \tag{3.6}$$

where the β_j are the columns of β.

The representation (3.6) of the Σ_i suggests a further modification, which has mainly been motivated by practical examples. Frequently in applications of principal component analysis the investigator is interested primarily in the first few components and discards the others. Similarly, we may wish to estimate only q common principal components, the remaining $p-q$ ones being possibly different from group to group. An appropriate model could be defined as follows.

Case 4: The partial CPC model. For a fixed positive integer $q < p-1$, let

$$\begin{aligned} \Sigma_i &= \lambda_{i1}\beta_1\beta_1^t + \ldots + \lambda_{iq}\beta_q\beta_q^t \\ &\quad + \lambda_{i,q+1}^{(i)}\beta_{q+1}^{(i)}\beta_{q+1}^{(i)t} + \ldots + \lambda_{ip}\beta_p^{(i)}\beta_p^{(i)t}, \quad i = 1, 2, \ldots, s \end{aligned} \tag{3.7}$$

where $\beta_1, \beta_2, \ldots, \beta_q$ are the common eigenvectors of all Σ_i and $\beta_{q+1}^{(i)}, \ldots, \beta_p^{(i)}$ may be specific to each covariance matrix Σ_i.

Assume that the first q common eigenvectors of all Σ_i are ordered, and labeled 1 to q. If we let

$$\beta^{(i)} = (\beta_1, \ldots, \beta_q, \beta_{q+1}^{(i)}, \ldots, \beta_p^{(i)}) \tag{3.8}$$

then (3.7) can also be written as

$$\Sigma_i = \beta^{(i)}\Lambda_i\beta^{(i)t}, \tag{3.9}$$

but (3.7) exhibits the basic idea underlying the partial CPC model more clearly.

Remark: By the orthogonality of all $\beta^{(i)}$, the model with $q = p-1$ common components implies the ordinary CPC model of **Case 3**. There is not just one

partial CPC model, but a family of models, some of which are nested hierarchically.

□

The number of parameters in the partial CPC model is as follows: sp parameters for the diagonal matrices Λ_i, $pq - q(q+1)/2$ parameters for the common eigenvectors β_1 to β_q, and $s(p-q)(p-q-1)/2$ parameters for the specific vectors $\beta_{q+1}^{(i)}$ to $\beta_p^{(i)}$. (To see this, note that there are $p-(q+1),\ p-(q+2),\ \ldots,\ p-p$ parameters for $\beta_{q+1}^{(i)},\ \beta_{q+2}^{(i)},\ \ldots,\ \beta_p^{(i)}$, respectively, so that the total number of parameters defining $\beta_{q+1}^{(i)}$ to $\beta_p^{(i)}$ is $\sum_{k=1}^{p-q-1} k = (p-q-1)(p-q)/2$). Thus, the total number of parameters is $p(p-1)/2 + sp + (s-1)(p-q)(p-q-1)/2$.

As stated above, if we set $q = p-1$ or $q = p$, the partial CPC model is the ordinary CPC model. The other extreme case, $q = 0$, leads to **Case 5**.

Case 5: $\Sigma_1, \ldots, \Sigma_s$ are arbitrary positive definite covariance matrices.

Here there is no assumed relationship amongst the s groups. Then we have to analyze them separately. There are $p(p+1)/2$ parameters for each of Σ_i. The total number of parameters is $sp(p+1)/2$

3.1.2 Maximum Likelihood Estimation

Common principal component analysis (CPCA) is a generalization of principal component analysis (PCA) to s groups. The key assumption is that the $p \times p$ covariance matrices $\Sigma_1, \ldots, \Sigma_s$ of s populations can be diagonalized by the same orthogonal transformation, that is, there exists an orthogonal matrix β such that

$$H_c : \beta^t \Sigma_i \beta = \Lambda_i (\text{diagonal}) \quad (i = 1, 2, \ldots, s) \tag{3.10}$$

holds. We call H_c the hypothesis of common principal components (CPC). In the one sample case $s = 1$, CPC reduces to ordinary Principal Components (PC).

We assume that all CPC are well defined, that is, for each j, $1 \leq j \leq p$, there is at least one population i in which the eigenvalues λ_{ij} are simple. This assumption can identify all β_j. Moreover, β_j will identify all $\lambda_{i'j}$ for $i' \neq i$. Thus, all parameters are identifiable.

Let $S_i,\ i = 1, 2, \ldots, s$, be the sample covariance matrices from a sample of size n_i in group i, so that $(n_i - 1)S_i$ has the Wishart distribution $W_p(n_i - 1,\ \Sigma_i)$ of a symmetric $p \times p$ matrix with $n_i - 1$ degrees of freedom and parameter matrix Σ_i. The joint likelihood function of $\Sigma_1, \Sigma_2, \ldots, \Sigma_s$ given $S_1, S_2, \ldots, S_s$ is

$$L(\Sigma_1, \Sigma_2, \ldots, \Sigma_s) = C \cdot \prod_{i=1}^{s} \exp\{-\frac{1}{2} tr(\Sigma_i^{-1} S_i)\} \cdot |\Sigma_i|^{\frac{n_i}{2}}, \tag{3.11}$$

where the factor C does not depend on the parameters. Introducing Lagrange multipliers ρ_j for the p constraints $\beta_j^t \beta_j = 1$ and ρ_{hj} for the $p(p-1)/2$ constraints $\beta_h^t \beta_j = 0$ $(h \neq j)$, maximizing likelihood is equivalent to minimizing the function

$$\begin{aligned} g(\beta, \Lambda_1, \ldots, \Lambda_s) &= -2 \log L + 2 \log C - \sum_{j=1}^{p} \rho_j (\beta_j^t \beta_j - 1) - 2 \sum_{h=1}^{p} \sum_{j=h+1}^{p} \rho_{hj} \beta_h^t \beta_j \\ &= \sum_{i=1}^{s} n_i [\log |\Sigma_i| + tr(\Sigma_i^{-1} S_i)] - \sum_{j=1}^{p} \rho_j (\beta_j^t \beta_j - 1) \\ &\quad -2 \sum_{j=1}^{p} \sum_{h=1}^{j-1} \rho_{hj} \beta_h^t \beta_j \\ &= \sum_{i=1}^{s} n_i \sum_{j=1}^{p} (\log \lambda_{ij} + \frac{\beta_j^t S_i \beta_j}{\lambda_{ij}}) - \sum_{j=1}^{p} \rho_j (\beta_j^t \beta_j - 1) \\ &\quad -2 \sum_{j=1}^{p} \sum_{h=1}^{j-1} \rho_{hj} \beta_h^t \beta_j. \end{aligned} \tag{3.12}$$

Taking partial derivatives with respect to the λ_{im} and setting them equal to zero

yields

$$\lambda_{im} = \beta_m^t S_i \beta_m, \quad i = 1, 2, \ldots, s; \quad m = 1, 2, \ldots, p. \tag{3.13}$$

Next take partial derivatives with respect to β_j and set them equal to zero. Multiplying on the left by β_m^t $(j \neq m)$, substituting (3.13), and solving for ρ_j, ρ_{hj}, then yields the system of equations

$$\beta_m^t (\sum_{i=1}^{s} n_i \frac{\lambda_{im} - \lambda_{ij}}{\lambda_{im}\lambda_{ij}} S_i) \beta_j = 0, \quad m, j = 1, 2, \ldots, p; \quad m \neq j. \tag{3.14}$$

This is the basic system of equations in CPC analysis. It has to be solved under the orthonormality constraints

$$\beta_m^t \beta_j = \delta_{mj}. \tag{3.15}$$

and using (3.13). The FG Algorithm has been proposed by Flury and Gautschi (1986) to numerically solve the equations (3.14) and (3.13) with constraints (3.15). The FG algorithm is viewed as a generalization of the Jacobi algorithm, the oldest known method (1846) for diagonalizing symmetric matrices. An iterative process to reduce off-diagonal elements to zero leads to the classical Jacobi iteration algorithm. From the numerical analyst's point of view, PCA consists mainly of the diagonalization of a single symmetric matrix, and CPC consists of the simultaneous diagonalization of groups of symmetric matrices.

3.1.3 Asymptotic distribution of MLE

We now develop the asymptotic distribution theory for the maximum likelihood estimates $\hat{\beta}$ and $\hat{\Lambda}$ under the CPC model. By the theory of maximum likelihood

estimation, the joint asymptotic distribution of the parameter estimates for unconstrained real parameters is multivariate normal, and the covariance matrix being given by the inverse of the Fisher information matrix. The log-likelihood function of the s samples, up to an additive constant, is given by

$$-\frac{1}{2}\sum_{i=1}^{s} n_i[\sum_{j=1}^{p}(\log \lambda_{ij} + \beta_j^t S_i \beta_j/\lambda_{ij}]. \tag{3.16}$$

Assume that β_j are well defined, i.e., for each pair $j \neq l$ there is at least one i, $1 \leq i \leq s$, such that $\lambda_{ij} \neq \lambda_{il}$. Let $\lambda_{(i)}^t = (\lambda_{i1}, \ldots, \lambda_{ip})$, $d = p(p-1)/2$, and denote by β^* a vector composed of d functionally independent elements of β. Let $n = n_1 + n_2 + \ldots + n_k$, and $r_i = n_i/n,\ i = 1, 2, \ldots, s$. Then the information matrix is

$$\begin{pmatrix} \Lambda & nG^t \\ nG & nA \end{pmatrix}, \tag{3.17}$$

where A and G are not yet determined, and Λ is

$$\Lambda = \begin{pmatrix} \frac{1}{2}nr_1\Lambda_1^{-2} & 0 & \ldots & 0 \\ 0 & \frac{1}{2}nr_2\Lambda_2^{-2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \frac{1}{2}nr_s\Lambda_s^{-2} \end{pmatrix}. \tag{3.18}$$

Since $\hat{\lambda}_{ij} = \hat{\beta}_j^t S_i \hat{\beta}_j$ and $\hat{\beta}_j$ is a consistent estimate of β_j, we can use the asymptotic ($\min_i n_i \to \infty$) normality of $n_i S_i$ (Muirhead 1982) to get the asymptotic distribution of $\hat{\lambda}_{ij}$ as

$$\sqrt{n_i}(\hat{\lambda}_{ij} - \lambda_{ij}) \sim N(0, 2\lambda_{ij}^2). \tag{3.19}$$

Furthermore, the vectors $\sqrt{n_i}(\hat{\lambda}_{(i)} - \lambda_{(i)})$ and $\sqrt{n_h}(\hat{\lambda}_{(h)} - \lambda_{(h)})$ $(h \neq i)$ are asymptotically independent since the S_i are independent.

From (3.17) the joint asymptotic covariance matrix of the k vectors $\sqrt{n}(\hat{\lambda}_{(i)} - \lambda_{(i)})$ is

$$V_\lambda = (\Lambda - G^t A^{-1} G)^{-1}. \tag{3.20}$$

We already know that V_λ must be diagonal, and the diagonal elements of V_λ are $(\frac{2}{r_1}\lambda_{11}^2,\ \frac{2}{r_2}\lambda_{12}^2,\ \ldots,\ \frac{2}{r_s}\lambda_{kp}^2)$. Since, at the maximum of the likelihood, the matrix A is positive definite, it follows that $G = 0$, and the $\hat{\lambda}_{ij}$ are therefore asymptotically independent of $\hat{\beta}$ (Flury 1988).

3.2 Random Effect Factor Model II

3.2.1 Model and Identifiability

Random Effect Factor Model II ($\mathbf{REFM}_2$). Assume that the observable random vector Y_i from the i'th group, $i = 1,\ 2,\ \ldots,\ s$, can be written as

$$Y_i = \sum_{k=1}^{q} c_{ik} P_k + \epsilon_i, \tag{3.21}$$

where $\{P_k,\ k = 1,\ 2,\ \ldots,\ q\}$ are nonrandom orthonormal coordinate directions, the random effects $c_{ik} \sim N(a_{ik}, b_{ik}^2)$, $1 \le k \le q$, the errors $\epsilon_i \sim N_p(0, \sigma_i^2 I_p)$, and the sequences $\{c_{ik}, k \le q\}$ and $\{\epsilon_i\}$ are jointly independent. The mean and variance of random vector Y_i under $\mathbf{REFM}_2$ are

$$E(Y_i) = \sum_{k=1}^{q} a_{ik} P_k = \pi_1 \underline{a}_i \tag{3.22}$$

and

$$Var(Y_i) = \sum_{k=1}^{q} b_{ik}^2 P_k^{\otimes 2} + \sigma_i^2 I_p = \pi_1 B_i \pi_1^t + \sigma_i^2 I_p \tag{3.23}$$

where $\pi_1 = (P_1,\ P_2,\ \ldots,\ P_q)$ is a $p \times q$ matrix with orthonormal column vectors, $\underline{a}_i^t = (a_{i1},\ a_{i2},\ \ldots,\ a_{iq})$ is a q dimensional mean vector from the i'th group in the common factor space $V_1 = col(\pi_1) = span(P_1,\ P_2,\ \ldots,\ P_q)$, and $B_i = Diag(b_{i1}^2,\ b_{i2}^2\ \ldots\ b_{iq}^2)$ is a $q \times q$ matrix with diagonal elements $b_{i1}^2,\ b_{i2}^2\ \ldots\ b_{iq}^2$ and zeros elsewhere. Based on the model assumption, the random vector Y_i from the i'th group follows a normal distribution with mean $\mu_i = \pi_1 \underline{a}_i$ and covariance matrix $\Sigma_i = \pi_1 B_i \pi_1^t + \sigma_i^2 I_p$. That is,

$$Y_i \sim N_p(\pi_1 \underline{a}_i, \pi_1 B_i \pi_1^t + \sigma_i^2 I_p). \tag{3.24}$$

If there are n_i observations from i'th group Y_i, $i = 1,\ 2,\ \ldots,\ s$, say, $y_{i1},\ y_{i2},\ \ldots,\ y_{in_i}$, our data model under $\mathbf{REFM}_2$ is

$$y_{ij} = \sum_{k=1}^{q} c_{ijk} P_k + \epsilon_{ij} \quad , \qquad i = 1, 2, \ldots, s; \quad j = 1, 2, \ldots, n_i, \tag{3.25}$$

where the random effects $c_{ijk} \sim \mathcal{N}(a_{ik}, b_{ik}^2),\ 1 \leq k \leq q,$ are independent, the errors $\{\epsilon_{ij}\}$ are i.i.d. with $\epsilon_{ij} \sim \mathcal{N}_p(0, \sigma_i^2 I_p)$, the series $\{\epsilon_{ij}\}$ and the series $\{c_{ijk}\}$ are independent, and $P_1, P_2, \ldots, P_k$ are orthonormal. Thus,

$$y_{ij} \sim N_p(\mu_i, \Sigma_i), \tag{3.26}$$

where

$$\begin{cases} \Sigma_i = \pi_1 B_i \pi_1^t + \sigma_i^2 I_p \\ \\ \mu_i = \pi_1 \underline{a}_i. \end{cases} \tag{3.27}$$

Let $\theta = (\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_s, Diag(B_1), Diag(B_2), \ldots, Diag(B_s), \sigma_1^2, \sigma_2^2, \ldots, \sigma_s^2,$ $P_1, P_2, \ldots, P_s)$. We define the parameter space as $\Theta = \{\theta : b_{ik}^2 > 0,\ \sigma_i^2 > 0,\ \pi_1^t \pi_1 = I_q\ for\ 1 \leq i \leq s,\ and\ 1 \leq k \leq q\}$. For any fixed group i, the parameters are $\underline{a}_i, B_i, \sigma_i^2$, and π_1. Since all groups share the parameter $\pi_1 \in \mathcal{M}$, we call π_1 the

common factor matrix, and the space spanned by the columns of π_1 the common factor space. Decompose Θ into two subspaces Θ_1 and Θ_2, with $\Theta = \Theta_1 \times \Theta_2$, and $\Theta_2 = \{\theta_2 \in \mathcal{M} : \theta_2^t\theta_2 = I_q\}$. Then, the parameter space Θ_1 can be further decomposed into s subspaces $\Theta_{11}, \Theta_{12}, \ldots, \Theta_{1s}$, with $\Theta_1 = \Theta_{11} \times \Theta_{12} \times \ldots \times \Theta_{1s}$, where $\Theta_{1i} = \{(a_i, B_i, \sigma_i^2) : b_{ik}^2 > 0, \sigma_i^2 > 0, i = 1, 2, \ldots, s;\ k = 1, 2, \ldots, q\} = \mathcal{R}^q \times \mathcal{R}_+^q \times \mathcal{R}_+$. Now, we can write out the parameter space for each group, say i, as $\Theta_{1i} \times \Theta_2$. There are $(2q + 1)$ parameters for $\theta_{1i} \in \Theta_{1i}$. Then the total number of the parameters in $\theta_1 = (\theta_{11}, \theta_{12}, \ldots, \theta_{1s}) \in \Theta_1$ is $s(2q + 1)$. Also, there are $(pq - q(q + 1)/2)$ parameters for $\theta_2 = \pi_1$. Thus, the total of number of parameters under $\mathbf{REFM}_2$ is $2qs+s+pq-q(q+1)/2$. From section 2.1.3, we know there are $q+p(p+1)/2$ equations from each group, thus the total numbers of equations in (3.27) are $s(q+p(p+1)/2)$. The relationship between the numbers of equations and the numbers of parameters is

$$
\begin{aligned}
&(\#\textit{of equations}) - (\#\textit{of parameters}) \\
&= s(q + \frac{p(p+1)}{2}) - \{(2q+1)s + pq - \frac{q(q+1)}{2}\} \\
&= \frac{s}{2}(p - q + 2)(p - q - 1) + (s - 1)(pq - \frac{q(q+1)}{2}).
\end{aligned} \tag{3.28}
$$

Since $p > q$ and $s > 1$, the above expression indicates that there are many more components in $\mu_i, \Sigma_i, 1 \leq i \leq s$ than parameters in $\mathbf{REFM}_2$.

Now, let us discuss the identifiability of all parameters. As we already mentioned, in $\mathbf{REFM}_2$ all groups share the common factor space, which is spanned by the columns of π_1. So it is important to identify π_1. Conventionally, we could choose

one group, say $i = 1$, and use the $\mathbf{REFM}_1$ result from section 2.1.3 that

$$b_{11}^2 > b_{12}^2 > \ldots > b_{1q}^2, \tag{3.29}$$

is a sufficient condition for us to identify all parameters $\underline{a}_1, B_1, \sigma_1^2$, and π_1 in the first group. Since

$$\Sigma_i P_k = \begin{cases} (b_{ik}^2 + \sigma_i^2)P_k & \text{for } k \leq q, \\ \sigma_i^2 P_i & \text{for } k > q. \end{cases} \quad i = 2, 3, \ldots, s \tag{3.30}$$

we can identify b_{ik}^2 and σ_i^2. Since $\mu_i = \pi_1 \underline{a}_i,\ i = 2, 3, \ldots, s$, we can identify $\underline{a}_i$ as well. Therefore, the condition (3.29) is sufficient to identify all parameters under $\mathbf{REFM}_2$. More generally, a condition for identifiability of all parameters under $\mathbf{REFM}_2$ is that there is a specified group i in which

$$b_{i1}^2 > b_{i2}^2 > \ldots > b_{iq}^2. \tag{3.31}$$

3.2.2 Maximum Likelihood Estimates for REFM_2

The probability density function of $Y_i,\ i = 1, 2, \ldots, s$, under $\mathbf{REFM}_2$ is

$$f(y_i) \ = \ \frac{\exp\{-\frac{1}{2}(y_i - \pi_1 \underline{a}_i)^t (\pi_1 B_i {\pi_1}^t + \sigma_i^2 I_p)^{-1} (y_i - \pi_1 \underline{a}_i)\}}{(2\pi)^{p/2} |\pi_1 B_i {\pi_1}^t + \sigma_i^2 I_p|^{1/2}}. \tag{3.32}$$

We are interested in estimating the parameters $\underline{a}_1,\ \underline{a}_2,\ \ldots,\ \underline{a}_s,\ B_1,\ B_2,\ \ldots,\ B_s,$ $\sigma_1^2,\ \sigma_2^2,\ \ldots,\ \sigma_s^2,\ P_1,\ P_2,\ \ldots,\ P_q$. Let B_i^* be the $p \times p$ diagonal matrix with $b_{ik}^2 \equiv 0$ for $k > q$,

$$B_i^* = Diag(b_{i1}^2,\ b_{i2}^2,\ \ldots,\ b_{iq}^2,\ b_{i,q+1}^2,\ \ldots,\ b_{i,p}^2); \tag{3.33}$$

that is,

$$B_i^* = \begin{pmatrix} B_i & 0 \\ 0 & 0 \end{pmatrix}. \tag{3.34}$$

Under $\mathbf{REFM}_2$, the log likelihood function for the sample $\{y_{ij},\ i = 1,\ 2,\ \ldots,\ s;\ j = 1,\ 2,\ \ldots,\ n_i\}$ is

$$\begin{aligned}
l(\theta) &\equiv \log L(\theta) \\
&= \sum_{i=1}^{s}\sum_{j=1}^{n_i} \log f(y_{ij}) \\
&= \sum_{i=1}^{s}\sum_{j=1}^{n_i}\{-\frac{p}{2}\log(2\pi) - \frac{1}{2}\sum_{k=1}^{q}\log(b_{ik}^2+\sigma_i^2) - \frac{1}{2}\sum_{k=q+1}^{p}\log(\sigma_i^2) \\
&\quad -\frac{1}{2}(y_{ij}-\pi_1\underline{a}_i)^t(\pi^t B_i^*\pi + \sigma_i^2 I_p)^{-1}(y_{ij}-\pi_1\underline{a}_i)\} \\
&= -\frac{p}{2}(\sum_{i=1}^{s} n_i)\log(2\pi) - \frac{1}{2}\sum_{i=1}^{s}\sum_{k=1}^{q} n_i\log(b_{ik}^2+\sigma_i^2) \\
&\quad -\frac{1}{2}\sum_{i=1}^{s}\sum_{k=q+1}^{p} n_i\log(\sigma_i^2) - \frac{1}{2}\sum_{i=1}^{s}\sum_{j=1}^{n_i}\sum_{k=1}^{q}\frac{1}{b_{ik}^2+\sigma_i^2}(P_k^t y_{ij} - a_{ik})^2 \\
&\quad -\frac{1}{2}\sum_{i=1}^{s}\sum_{j=1}^{n_i}\sum_{k=q+1}^{p}\frac{1}{\sigma_i^2}(P_k^t y_{ij})^2. \qquad (3.35)
\end{aligned}$$

For $\pi_1 = (P_1,\ P_2,\ \ldots,\ P_q)$ assumed fixed, we take partial derivatives with regard to $a_{ik},\ b_{ik}^2,\ \sigma_i^2$, respectively. Setting all first order derivatives equal to zero and simplifying all equations as in Chapter 2, we have

$$\begin{cases}
\hat{a}_{ik} = \frac{1}{n_i}\sum_{j=1}^{n_i} P_k^t y_{ij} = P_k^t\bar{y}_i \\
\hat{\sigma}_i^2 = \frac{1}{n_i(p-q)}\sum_{j=1}^{n_i}\sum_{q=k+1}^{p}(P_k^t y_{ij})^2 = \frac{1}{p-q}\sum_{k=q+1}^{p} P_k^t C_{ii} P_k \\
\hat{b}_{ik}^2 = P_k^t S_i P_k - \hat{\sigma}_i^2
\end{cases} \qquad (3.36)$$

where $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij},\ C_{ii} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}y_{ij}^t$, and $S_i = \frac{1}{n_i}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^{\otimes 2}$.

Therefore, the restricted maximum likelihood estimator $\hat{\theta}_1(\theta_2)$ is $(\hat{\theta}_{11}\ (\theta_2),\ \hat{\theta}_{12}(\theta_2),\ \ldots, \hat{\theta}_{1s}(\theta_2))$, given by

$$\hat{\theta}_{1i}(\theta_2) = \begin{cases}
\hat{\underline{a}}_i = \pi_1^t\bar{y}_i \\
\hat{B}_i = Diag\{Diag(\pi_1^t S_i\pi_1)\} - \hat{\sigma}_i^2 I_q \\
\hat{\sigma}_i^2 = \frac{1}{p-q}\{tr(C_{ii}) - tr(\pi_1\pi_1^t C_{ii})\}.
\end{cases} \qquad (3.37)$$

Remark: $\sigma_i^2 = \sigma^2$ is a special case in $\mathbf{REFM}_2$. In this case, we completely ignore the differences in error distributions among the s groups since the variations for these error terms are much smaller than these in the common factor space. In this case, we have similarly the restricted maximum likelihood estimator in each group i, when π_1 is given:

$$\hat{\theta}_{1i}(\theta_2) = \begin{cases} \underline{\hat{a}}_i = \pi_1^t \bar{y}_i \\ \hat{B}_i = Diag\{Diag(\pi_1^t S_i \pi_1)\} - \hat{\sigma}^2 I_q \\ \hat{\sigma}^2 = \frac{1}{s(p-q)} \sum_{i=1}^{s} \{tr(C_{ii}) - tr(\pi_1 \pi_1^t C_{ii}\}. \end{cases} \tag{3.38}$$

□

For simplicity we assume that all n_i are equal, that is, $n_i = n,\ i = 1, 2, \ldots, s$. The profile likelihood function under $\mathbf{REFM}_2$ is

$$\begin{aligned} l_p(\theta_2) &= l(\hat{\theta}_1(\theta_2),\ \theta_2) \\ &= \frac{n}{2}\{C - \sum_{i=1}^{s}\sum_{k=1}^{q} \log(P_k^t S_i P_k) \\ &\quad -(p-q)\sum_{i=1}^{s} \log(tr(C_{ii}) - \sum_{k=1}^{q} P_k^t C_{ii} P_k)\} \end{aligned} \tag{3.39}$$

where $C = -sp\log(2\pi) - ps + (p-q)\log(p-q)$.

So far, we have the restricted maximum likelihood estimators $\underline{\hat{a}}_1,\ \underline{\hat{a}}_2,\ \ldots,\ \underline{\hat{a}}_s$, $\hat{B}_1,\ \hat{B}_2,\ \ldots,\ \hat{B}_s$, and $\hat{\sigma}_1^2,\ \hat{\sigma}_2^2,\ \ldots,\ \hat{\sigma}_s^2$ in terms of fixed (assumed known) common factors $P_1,\ P_2,\ \ldots,\ P_q$ under $\mathbf{REFM}_2$. The restricted MLE's are functions of π_1, the sample mean $\bar{y}_i$ in each group, and the sample covariance matrix S_i in each group.

By Lemma 2 in Chapter 2, the maximum likelihood estimator $\hat{\theta}_2$ based on the

profile likelihood exists, and

$$\sup_{\theta\in\Theta} l(\theta) = \max_{\theta_2\in\Theta_2} l_p(\theta_2) \tag{3.40}$$

since the profile log-likelihood function in equation (3.39) is continuous from Θ_2 into $\mathbf{R}$, and $\Theta_2 = \{\theta_2 \in \mathcal{M} : \theta_2^t\theta_2 \ = \ I_q\}$ is a closed set in the space $\mathcal{M}$ of $p \times q$. As in Chapter 2, the profile likelihood does not have a closed-form analytic maximizer for $P_1, P_2, \ldots, P_q$. The Quasi-Newton method will be applied to solve for π_1.

3.3 Random Effect Factor Model III

3.3.1 Model and Identifiability

Random Effect Factor Model III ($\mathbf{REFM_3}$). Assume that the observable random vector Y_i from i'th group, $i = 1,\ 2,\ \ldots,\ s$, can be written as

$$Y_i = \sum_{k=1}^{q} c_{ik}P_k \ + \sum_{k=q+1}^{r} c_{ik}P_k^{(i)} \ + \epsilon_i, \tag{3.41}$$

where the errors $\epsilon_i \sim N_p(0, \sigma_i^2 I_p)$; the random effects $c_{ik} \sim N(a_{ik}, b_{ik}^2),\ 1 \le i \le s,\ 1 \le k \le q$; and the series $\{c_{ik}, k \le r\}$ and the series $\{\epsilon_i\}$ are independent; $\{P_k,\ k = 1,\ 2,\ \ldots,\ q\}$ are nonrandom orthonormal coordinate directions, as are $\{P_k^{(i)},\ k = q+1,\ \ldots,\ r\}$ for each group i; and $P_k^{(i)} \perp P_{k'},\ k = q+1, \ldots, r,\ k' = 1, \ldots, q$, for all i; $\sum_{k=q+1}^{r} c_{ik}P_k^{(i)} \sim N_p(0,\ \Sigma^i)$, and $\text{rank}(\Sigma^i) = r - q$. The mean and variance of the random vector Y_i under $\mathbf{REFM_3}$ are

$$E(Y_i) = \sum_{k=1}^{q} a_{ik}P_k = \pi_1\underline{a}_i \tag{3.42}$$

and

$$Var(Y_i) = \sum_{k=1}^{q} b_{ik}^2 P_k^{\otimes 2} + \Sigma^i + \sigma_i^2 I_p = \pi_1 B_i \pi_1^t + \Sigma^i + \sigma_i^2 I_p \tag{3.43}$$

where $\pi_1 = (P_1,\ P_2,\ \ldots,\ P_q)$ is a $p \times q$ matrix with orthonormal column vectors, $\underline{a}_i^t = (a_{i1},\ a_{i2},\ \ldots,\ a_{iq})$ is a q dimensional mean vector from the i'th group, and $B_i = Diag(b_{i1}^2, b_{i2}^2 \ldots b_{iq}^2)$ is a $q \times q$ matrix with diagonal elements $b_{i1}^2, b_{i2}^2 \ldots b_{iq}^2$ and zeros elsewhere. Based on the model assumption, the random vector Y_i from i'th group follows a normal distribution with mean $\mu_i = \pi_1 \underline{a}_i$ and covariance matrix $A_i = \pi_1 B_i \pi_1^t + \Sigma^i + \sigma_i^2 I_p$. That is,

$$Y_i \sim N_p(\pi_1 \underline{a}_i, \pi_1 B_i \pi_1^t + \Sigma^i + \sigma_i^2 I_p). \tag{3.44}$$

Define $V_1 = \text{span}\{P_1,\ P_2,\ \ldots,\ P_q\}$, and $V_{2i} = \text{span}\{\ col(\Sigma^i)\ \}$. Assume that $\cap_{i=1}^s V_{2i} = \phi$, and let $W_i = V_1 \oplus V_{2i}$, so that $\cap_{i=1}^s W_i = V_1$. The model parameters under REFM$_3$ are $\underline{a}_1,\ \underline{a}_2,\ \ldots,\ \underline{a}_s,\ B_1,\ B_2,\ \ldots,\ B_s,\ \sigma_1^2,\ \sigma_2^2,\ \ldots,\ \sigma_s^2,$ $\Sigma^1,\ \Sigma^2,\ \ldots,\ \Sigma^s$, and π_1. Now, the question is whether there exist $\underline{a}_i,\ B_i^2,\ \sigma_i^2,\ \Sigma^i,$ and π_1 to satisfy (3.42) and (3.43) when mean μ_i and covariance matrix A_i are given. If the answer is yes, then the next question is whether they are unique. Let us first count the numbers of parameters, and of equations to solve for these parameters. We note that the difference between **REFM$_2$** and **REFM$_3$** is the extra parameters Σ^i. Thus, we can use the results from **REFM$_2$**. For the fixed group i, Σ^i has $\frac{1}{2}(r-q)(2p+1-r-q)$ parameters. (To see this, note that: $\Sigma^i = \beta^{(i)} \Lambda_i \beta^{(i)t}$ with $\beta^{(i)t}\beta^{(i)} = I_{r-q}$, where $\beta^{(i)}$ is a $p \times (r-q)$ matrix and Λ_i is a diagonal matrix. Since any column of β_i is orthogonal to any column of π_1, we count the number of parameters for column 1 to column $(r-q)$ of Σ^i as $p-1-q,\ p-1-(q+1),\ \ldots,\ p-1-(r-1)$, respectively. Thus, the total number of parameters for Σ^i is $(r-q)+(p-1)-q+(p-1)-(q+1)+\ldots+(p-1)-(r-1) = (r-q)+$

$(p-1)(r-q)+(r-q)(q+r-1)/2 = (r-q)(2p-r-q+1)/2$). Thus, the total number of parameters under $\mathbf{REFM}_3$ is $2qs+s+pq-q(q+1)/2+(s/2)(r-q)(2p+1-r-q)$. From Section 3.2.1, the total number of equations in (3.42) and (3.43) is $s(q+p(p+1)/2)$. The relationship between the number of equations and the number of parameters is

$$
\begin{aligned}
&(\#\textit{of equations}) - (\#\textit{of parameters}) \\
&= \frac{s}{2}(p-q+2)(p-q-1) + (s-1)(pq - \frac{q(q+1)}{2}) - \frac{s}{2}(r-q)(2p+1-r-q) \\
&= \frac{s}{2}(p-r+2)(p-r-1) + (s-1)(pq - \frac{q(q+1)}{2}). \qquad (3.45)
\end{aligned}
$$

Since $p > r$ and $s > 1$, the above expression indicates that there are more equations than parameters in $\mathbf{REFM}_3$.

Let $\Sigma^i = \sum_{k=q+1}^{r} d_{ik}^2 P_k^{(i)} P_k^{(i)t}$, where $d_{ik}^2 = Var(c_{ik})$, $k = q+1, \ \ldots, \ r, \ i = 1, \ 2, \ \ldots, \ s$, and let $\{u_{ik}, k = r+1, \ldots, p\}$ be an arbitrary orthonormal basis in $V_{3i} \equiv W_i^{\perp}$. Then

$$
I_p = \sum_{k=1}^{q} P_k P_k^t + \sum_{k=q+1}^{r} P_k^{(i)} P_k^{(i)t} + \sum_{k=r+1}^{p} u_{ik} u_{ik}^t. \qquad (3.46)
$$

Thus,

$$
\begin{aligned}
A_i &= \pi_1 B_i \pi_1^t + \Sigma^i + \sigma_i^2 I_p \\
&= \sum_{k=1}^{q} (b_{ik}^2 + \sigma_i^2) P_k^{\otimes 2} + \sum_{k=q+1}^{r} (d_{ik}^2 + \sigma_i^2) P_k^{(i)\otimes 2} + \sum_{k=r+1}^{p} \sigma_i^2 u_{ik}^{\otimes 2} \qquad (3.47)
\end{aligned}
$$

The $p \times p$ covariance matrix A_i has p eigenvectors and eigenvalues, and among them there will be $p-r$ smallest eigenvalues, and $p-r$ eigenvectors corresponding to the $(p-r)$ smallest eigenvalues. This means that all σ_i^2 are identifiable, and the error space V_{3i} is determined. So is the complement of V_{3i}; that is, W_i is determined.

Hence, V_1 is determined since $V_1 = \cap_{i=1}^{s} W_i$. Moreover, the projection from V to V_1 , that is, $\pi_1\pi_1^t$, is determined. Let $Y_i = \alpha_i + \beta_i$, where $\alpha_i = (\pi_1\pi_1^t)Y_i$, and $\beta_i = (I_p - \pi_1\pi_1^t)Y_i$. Since $\alpha_i^t\beta_i = y_i^t(\pi_1\pi_1^t)(I_p - \pi_1\pi_1^t)y_i = 0$, the vectors α_i and β_i are orthogonal. Now, we project the observed random vector into two orthogonal spaces V_1 and $V_{2i} \oplus V_{3i} = V_1^{\perp}$. The probability laws of both α_i and β_i are determined. Let us write down the mean and covariance matrix for α_i and β_i, respectively.

$$\mathrm{E}(\alpha_i) = \mathrm{E}\{(\pi_1\pi_1^t)Y_i\} = (\pi_1\pi_1^t)\pi_1 a_i = \pi_1 a_i, \tag{3.48}$$

and

$$\begin{aligned} Var(\alpha_i) &= Var\{(\pi_1\pi_1^t)y_i\} = (\pi_1\pi_1^t)(\pi_1 B_i\pi_1^t + \Sigma^i + \sigma_i^2 I_p)(\pi_1\pi_1^t) \\ &= \pi_1 B_i\pi_1^t + \sigma_i^2\pi_1\pi_1^t. \end{aligned} \tag{3.49}$$

Thus,

$$\alpha_i \sim N_p(\pi_1 a_i, \pi_1 B_i\pi_1^t + \sigma_i^2\pi_1\pi_1^t). \tag{3.50}$$

Also,

$$\mathrm{E}(\beta_i) = \mathrm{E}\{I_p - \pi_1\pi_1^t)Y_i\} = (I_p - \pi_1\pi_1^t)\pi_1 a_i = 0, \tag{3.51}$$

and

$$\begin{aligned} \mathrm{Var}(\beta_i) &= \mathrm{Var}\{(I_p - \pi_1\pi_1^t)Y_i\} \\ &= (I_p - \pi_1\pi_1^t)(\pi_1 B_i\pi_1^t + \Sigma^i + \sigma_i^2)(I_p - \pi_1\pi_1^t) \\ &= \Sigma^i + \sigma^2(I_p - \pi_1\pi_1^t) \end{aligned} \tag{3.52}$$

Thus ,

$$\beta_i \sim N_p(0, \Sigma^i + \sigma_i^2(I_p - \pi_1\pi_1^t)). \tag{3.53}$$

Now, we try to identify $\underline{a}_i,\ B_i,\ \pi_1$, and Σ^i from the mean and covariance matrix of $\{\alpha_i\}$, and covariance matrix of $\{\beta_i\}$. The law of the random vector α_i only depends on the parameters $\underline{a}_i,\ B_i,\ \pi_1$, and does not depend on Σ^i. Thus α_i follows $\mathbf{REFM}_2$. We can directly use the results from $\mathbf{REFM}_2$. That is, the sufficient condition for identification is that there exists a known group, say i, such that

$$b_{i1}^2 > b_{i2}^2 > \ldots > b_{iq}^2. \tag{3.54}$$

Therefore, we are able under (3.54) to identify $\pi_1,\ \underline{a}_i$, and B_i for $i = 1,\ 2,\ \ldots,\ s$. Since the covariance matrix of $\{\beta_i\}$ is determined, then also Σ^i is determined since π_1 is already determined. Therefore, the condition (3.54) implies identifiability for all parameters under $\mathbf{REFM}_3$.

3.3.2 Maximum Likelihood Estimates for REFM_3

For group i from 1 to s, we have

$$Y_i \sim N_p(\pi_1\underline{a}_i, \pi_1 B_i \pi_1^t + \Sigma^i + \sigma_i^2 I_p). \tag{3.55}$$

The probability density function of $Y_i,\ i = 1, 2, \ldots, s$, under $\mathbf{REFM}_3$ is

$$f(y_i) \ = \ \frac{\exp\{-\frac{1}{2}(y_i - \pi_1\underline{a}_i)^t(\pi_1 B_i {\pi_1}^t + \Sigma^i + \sigma_i^2 I_p)^{-1}(y_i - \pi_1\underline{a}_i)\}}{(2\pi)^{p/2}|\pi_1 B_i {\pi_1}^t + \Sigma^i + \sigma_i^2 I_p|^{1/2}}. \tag{3.56}$$

Let $\{y_{ij},\ i = 1,\ 2,\ \ldots,\ s;\ j = 1,\ 2,\ \ldots,\ n\}$ be n observations from each group. Under $\mathbf{REFM}_3$, the log likelihood function is

$$\begin{aligned} l(\theta) \ &\equiv \ \log L(\theta) \\ &= \ \log \prod_{i=1}^{s}\prod_{j=1}^{n} f(y_{ij}) \end{aligned}$$

$$= \sum_{i=1}^{s}\sum_{j=1}^{n} \log f(y_{ij}). \tag{3.57}$$

Recall for fixed π_1 that $Y_i = \alpha_i + \beta_i,\ i = 1,\ 2,\ \ldots,\ s$, with $\alpha_i \perp \beta_i$, and $\log f(y_i) = \log f(\alpha_i) + \log f(\beta_i)$. Now, the log-likelihood can be written in terms of the log-likelihood of α_i and β_i, that is,

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{s}\sum_{j=1}^{n}(\log f(\alpha_i) + \log f(\beta_i)) \\ &= l_\alpha(\theta) + l_\beta(\theta). \end{aligned} \tag{3.58}$$

After projection, we seperate the parameters as well. Only some of the parameters $\underline{a}_i,\ B_i,\ \sigma_i^2,\ \pi_1$ are related to α; also, B_i and σ_i^2 only appear together in the form $b_{ik}^2 + \sigma_i^2$. Next, only parameters $\Sigma^i,\ \sigma_i^2$, and π_1 depend on β. Again, we use the profile likelihood method by first assuming that π_1 is given. Then, we take partial derivatives with regard to $a_{ik},\ b_{ik}^2,\ \sigma_i^2$, respectively. We have the following equations:

$$\begin{aligned} \frac{\partial l(\theta)}{\partial a_{ik}} &= \frac{\partial l_\alpha(\theta)}{\partial a_{ik}} \\ \frac{\partial l(\theta)}{\partial b_{ik}^2} &= \frac{\partial l_\alpha(\theta)}{\partial b_{ik}^2} \\ \frac{\partial l(\theta)}{\partial \sigma_i^2} &= \frac{\partial l_\alpha(\theta)}{\partial \sigma_i^2} + \frac{\partial l_\beta(\theta)}{\partial \sigma_i^2}. \end{aligned}$$

Since $\underline{a}_i$ and b_{ik}^2 enter only $l_\alpha(\theta)$ but not $l_\beta(\theta)$, we will find equations for $\underline{a}_i$ and b_{ik}^2 first. Since $\alpha_i \sim N_p(\pi_1 a_i, \pi_1 B_i \pi_1^t + \sigma_i^2 \pi_1 \pi_1^t)$, and rank of $(\pi_1 B_i \pi_1^t + \sigma_i^2 \pi_1 \pi_1^t) = q < p$, so α_i has a degenerate normal distribution. Let $\alpha_i^* = \pi_1^t \alpha_i$, so that $\alpha_i^* \sim N_q(a_i, B_i + \sigma_1^2 I_q)$. The log-likelihood function of α_i^* is given, up to a constant not depending on parameters, by

$$l_{\alpha^*}(\theta) = \sum_{i=1}^{s}\sum_{j=1}^{n}\left\{-\frac{1}{2}\sum_{k=1}^{q}\log(b_{ik}^2 + \sigma_i^2) - \frac{1}{2}(\alpha_i^* - a_i)^t (B_i + \sigma_i^2 I_q)^{-1}(\alpha_i^* - \underline{a}_i)\right\}$$

$$= \sum_{i=1}^{s}\sum_{j=1}^{n}\sum_{k=1}^{q}\left\{-\frac{1}{2}\log(b_{ik}^2+\sigma_i^2)-\frac{1}{2(b_{ik}^2+\sigma_i^2)}(P_k^t y_{ij}-a_{ik})^2\right\}.$$

Take partial derivatives of $l_{\alpha^*}(\theta)$ with respect to a_{ik} for $1 \le i \le s,\ 1 \le k \le q$:

$$\frac{\partial l_{\alpha^*}(\theta)}{\partial a_{ik}} = \sum_{j=1}^{n}\frac{-1}{2(b_{ik}^2+\sigma_i^2)}(-2)(P_k^t y_{ij}-a_{ik}). \tag{3.59}$$

Next take derivatives of $l_{\alpha^*}(\theta)$ with respect to b_{ik}^2 for $1 \le i \le s,\ 1 \le k \le q$:

$$\frac{\partial l_{\alpha^*}(\theta)}{\partial b_{ik}^2} = \sum_{j=1}^{n}(-\frac{1}{2(b_{ik}^2+\sigma_i^2)}+\frac{1}{2(b_{ik}^2+\sigma_i^2)}(P_k^t y_{ij}-a_{ik})^2).$$

Setting these derivatives to zero and simplifying , we have

$$\begin{cases} \hat{a}_{ik} = \frac{1}{n}\sum_{j=1}^{n} P_k^t y_{ij} = P_k^t \bar{y}_i \\ \hat{b}_{ik}^2 + \hat{\sigma}_i^2 = P_k^t S_i P_k \end{cases} \tag{3.60}$$

where $\bar{y}_i = n^{-1}\sum_{j=1}^{n} y_{ij}$, and $S_i = n^{-1}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i)^{\otimes 2}$.

Now, we derive the restricted MLE of σ_i^2 and Σ^i from $l_\beta(\theta)$. We know $\beta \sim N_p(0, \Sigma^i + \sigma_i^2(I_p - \pi_1\pi_1^t))$. Since the rank of $\Sigma^i + \sigma_i^2(I_p - \pi_1\pi_1^t)$ is $(p-q) < p$, thus β also has a degenerate normal distribution. Let $\beta_i^* = \pi_2^t\beta_i$, where the columns of π_2 are any fixed orthonormal basis of $V_1^\perp$. Hence $\beta_i^* \sim N_{p-q}(0, \pi_2\Sigma^i\pi_2^t + \sigma_i^2 I_{p-q})$. Let $\Sigma^i = \Lambda_i\Lambda_i^t$, where $\Lambda_i = (d_{i,q+1}P_{q+1}^{(i)},\ \ldots,\ d_{i,r}P_r^{(i)})$. Thus,

$$\begin{aligned} \pi_2\Sigma^i\pi_2^t + \sigma_i^2 I_{p-q} &= \pi_2^t\Lambda_i\Lambda_i^t\pi_2 + \sigma_i^2 I_{p-q} \\ &= W_i W_i^t + \sigma_i^2 I_{p-q} \end{aligned} \tag{3.61}$$

where $W_i = \pi_2^t\Lambda_i$. The following two Theorems will find the restricted MLE $\hat{\sigma}_i^2$ and $\hat{\Sigma}^i$ in the general case.

Theorem 2 *If* $z_j \overset{iid}{\sim} N_a(0, WW^t + \sigma^2 I_a),\ j = 1,\ 2,\ \ldots,\ n$, *where* Λ *has* b $(b < a)$ *orthogonal columns, and* σ^2 *and* W *are unknown parameters, then the maximum*

likelihood estimators of σ^2 *and* W *are*

$$\hat{\sigma}^2_{ML} = \frac{1}{a-b}\sum_{k=b+1}^{a} \lambda_k \tag{3.62}$$

and

$$\hat{W}_{ML} = \mathcal{U}_b(\Lambda_b - \hat{\sigma}^2_{ML} I_b)^{\frac{1}{2}} \tag{3.63}$$

where the b *column vectors in the* $a \times b$ *matrix* $\mathcal{U}_b$ *are the principal (*b *largest eigenvalue) eigenvectors of* $n^{-1}\sum_{j=1}^{n} z_j^{\otimes 2}$*, with corresponding eigenvalues* $\lambda_1, \ldots, \lambda_b$ *as diagonal entries in the* $b \times b$ *diagonal matrix* Λ_b*.*

Proof: See Appendix A in Tipping and Bishop (1999). □

Remark: The parameter $\hat{\sigma}^2_{ML}$ has a clear interpretation as the variance lost in the projection onto the principal subspace, averaged over these lost dimensions. This result echoes the proposal of PPCA by Roweis (1997) and Tipping and Bishop (1999), that is, to recover these directions "outside" the principal subspace as i.i.d. Gaussian noise. □

Theorem 3 *If* $y_j \overset{iid}{\sim} N_p(0, LL^t + \sigma^2\pi_V)$*,* $j = 1, 2, \ldots, n$*, where* V *is a known subspace of* R^p*,* $dim(V) = a$*,* π_V *is a projection from* R^p *to* V*, and* L *has orthogonal columns with* $col(L) \subset V$*, then the maximum likelihood estimator of* σ^2 *is the average of* $a - b$ *smallest nonzero eigenvalues of* $n^{-1}\sum_{j=1}^{n} y_j^{\otimes 2}$*, and the maximum likelihood estimator of* L *is*

$$\hat{L}_{ML} = \mathcal{U}_b^* \cdot (\Lambda_b^* - \hat{\sigma}^2_{ML} I_b)^{\frac{1}{2}} \tag{3.64}$$

where the b *column vectors in the* $p \times b$ *matrix* $\mathcal{U}_b^*$ *are the principal eigenvectors of* $n^{-1}\sum_{j=1}^{n} y_j^{\otimes 2}$*, with corresponding eigenvalues* $\lambda_1^*, \ldots, \lambda_b^*$ *as diagonal entries in the*

$b \times b$ *diagonal matrix* Λ_b^*.

Proof: Let $v_1, v_2, \ldots, v_a$ be an orthonormal basis of $V \subset R^p$. Observe that $\{v_i,\ 1 \leq i \leq a\}$ are known since V is known. Let $M = (v_1 \mid v_2 \mid \ldots \mid v_a)$ be a $p \times a$ transformation matrix. Then the projection π_V onto V is MM^t. Since $col(L) \subset V$, $\pi_V\ L = L$.

Define $z_j \equiv M^t y_j$. Then $z_j \sim N_a(0, M^t L L^t M + \sigma^2 I_a)$. Next, define $W = M^t L$. Note that $W^t W = L^t M M^t L = L^t L$. Now apply Theorem 2 to find $\hat{\sigma}^2_{ML}$ and $\hat{W}_{ML}$. Since $MW = MM^t L = L$, it follows that $\hat{L}_{ML} = M\hat{W}_{ML}$.

Claim: The vector u_i is an eigenvector of $n^{-1}\sum_{j=1}^n z_j^{\otimes 2} = C_{zz}$ if and only if $u^* = Mu_i$ is an eigenvector of $n^{-1}\sum_{j+1}^n y_j^{\otimes 2} = C_{yy}$.

($\Rightarrow$) Recall $y_i \sim N_p(0, LL^t + \sigma^2 \pi_V)$. For any $x \in R^p$ with $x \in V^\perp$, $x^t y_i \sim N(0,0)$ since $x^t L = 0$ and $\pi_V x = 0$. Thus, $y_j \in V$ and $\pi_V C_{yy} = C_{yy}$. Since $C_{zz} u_i = \lambda_i u_i$, multiply M on the left on both sides of this equation. We have

$$MC_{zz}u_i = \lambda_i M u_i. \tag{3.65}$$

Also, $C_{zz} = n^{-1}\sum_{j=1}^n Z_j^{\otimes 2} = n^{-1}\sum_{j+1}^n M^t y_i y_i^t M = M^t C_{yy} M$. Thus

$$C_{yy}Mu_i = MM^t C_{yy} M u_i = MC_{zz}u_i = \lambda_i M u_i.$$

($\Leftarrow$) Since $C_{yy}Mu_i = \lambda_i Mu_i$, multiply M^t on the left in both sides of the equation. We have $M^t C_{yy} M u_i = \lambda_i u_i$; that is, $C_{zz}u_i = \lambda_i u_i$. □

Applying Theorem 3, to obtain the restricted MLE $\hat{\sigma}^2$ as the average of the smallest $(p-r)$ nonzero eigenvalues of $n^{-1}\sum_{j=1}^n \beta_{ij}^{\otimes 2} = (I_p - \pi_1 \pi_1^t) C_{i,yy} (I_p - \pi_1 \pi_1^t)$. That is, $\hat{\sigma}_i^2 = (p-r)^{-1}\sum_{k=r-q+1}^{p-q} \lambda_{i,k}$, where $C_{i,yy} = n^{-1}\sum_{j=1}^n y_{ij}^{\otimes 2}$, and the restricted

MLE of Σ^i is

$$\hat{\Sigma}^i = \hat{\Lambda}_i \hat{\Lambda}_i^t = \mathcal{U}_{i,(r-q)}(\Lambda_{i,(r-q)} - \hat{\sigma}_i^2 I_{(r-q)})\mathcal{U}_{i,(r-q)}^t \tag{3.66}$$

where the $(r-q)$ column vectors in the $p \times (r-q)$ matrix $\mathcal{U}_{i,(r-q)}$ are the principal eigenvectors of $(I_p - \pi_1\pi_1^t)C_{i,yy}(I_p - \pi_1\pi_1^t)$, with corresponding eigenvalues $\lambda_{i,1}, \ \ldots, \ \lambda_{i,(r-q)}$ in the $(r-q)\times(r-q)$ diagonal matrix $\Lambda_{i,(r-q)}$. Now, we can solve for $\hat{b}_{ik}^2$ from (3.60):

$$\hat{b}_{ik}^2 = P_k^t S_i P_k - \hat{\sigma}_i^2. \tag{3.67}$$

Therefore, the restricted maximum likelihood estimators $\hat{b}_{ik}^2$, when the factor directions are given, are

$$\hat{\theta}_{1i}(\theta_2) = \begin{cases} \underline{\hat{a}}_i = \pi_1^t \bar{y}_i \\ \hat{B}_i = Diag\{Diag(\pi_1^t S_i \pi_1)\} - \hat{\sigma}_i^2 I_q \\ \hat{\sigma}_i^2 = \frac{1}{p-r}\sum_{k=r-q+1}^{p-q} \lambda_{i,k} \\ \Sigma^i = \mathcal{U}_{i,r-q}(\Lambda_{i,r-q} - \hat{\sigma}_i I_{r-q}) \cup_{i,r-q}^t . \end{cases} \tag{3.68}$$

The profile likelihood function under REFM_3 is

$$\begin{aligned} l_p(\theta_2) &= l(\hat{\theta}_1(\theta_2), \ \theta_2) \\ &= \frac{n}{2}\{ \ C - \sum_{i=1}^{s}\sum_{k=1}^{q} \log(P_k^t S_i P_k) \\ &\quad - \sum_{i=1}^{s}\sum_{k=1}^{r-q} \log(\lambda_{ik}) - (p-r)\sum_{i=1}^{s} \log(\sum_{k=r-q+1}^{p-q} \lambda_{ij}) \ \} \end{aligned} \tag{3.69}$$

where $C = -sp\log(2\pi) - ps + (p-r)\log(p-r)$.

Applying Lemma 2 in Chapter 2, the maximum likelihood estimator $\hat{\theta}_2$ of the profile likelihood exists, and

$$\sup_{\theta\in\Theta} l(\theta) = \max_{\theta_2\in\Theta_2} l_p(\theta_2) \tag{3.70}$$

Chapter 4

Computational Method

4.1 EM Algorithm

4.1.1 Introduction

In 1977, Dempster, Laird, and Rubin (1997) first gave a general formulation of the EM algorithm, consisting of the expectation step (E-step) and the maximization step (M-step) in their general forms, for deriving maximum likelihood estimates from incomplete data. They also identified some theoretical properties of the EM algorithm and illustrated a wide range of applications in various statistical models. Since then, many papers have been published over the past 27 years, developing new methodologies using the EM algorithm in almost all fields in which statistical analysis is required, including engineering, medical science, sociology, and business administration. According to a survey conducted by Meng and Pedlow (1992), at least 1700 papers involving the EM algorithm exist on more than 1000 subjects. Moreover, Meng (1997) pointed out that more than 1000 papers were published in approximately 300 journals just in 1991 alone. Statistical journals only accounted for 15%. In July 2003, the Institute for Scientific Information released an updated list of the researchers with the most citations between January 1993 and April 2003. Donald B. Rubin has been ranked as the sixth most Cited researcher in the category of mathematics (AMSTAT News, Issue 317), with a total of 792 citations, largely because of his significant contributions on EM algorithm. These facts clearly

indicate that the EM algorithm has already become a very popular tool for statistical analysis based on the maximum likelihood estimation. From the citation point of view, more researchers use the EM algorithm now as a numerical method over those who use the Newton-Raphson method or other methods.

The EM algorithm has received tremendous attention and further extension. When people talk about the EM algorithm, they always mention two major advantages: simplicity and stability. Most multivariate methods require computation of either the inverse of matrices or the extraction of eigenvectors and eigenvalues, but EM often does not. The Newton-Raphson method and the EM algorithm have been programmed as the main numerical methods for performing the maximum likelihood estimation based on multidimensional data including missing values. The log-likelihood function is complicated, but it can be maximized using standard optimization routines. McHugh (1956, 1958) illustrated how this might be done using the standard Newton-Raphson technique. However, as with many other latent variable models, an easier method which enables larger problems to be tackled is offered by the EM algorithm [39]. Some experiments on the same data were conducted in the past to compare the two methods in terms of the numbers of iterations and central processing unit (CPU) time required to reach convergence. The results showed that the EM algorithm was able to determine a convergent value in all cases, while the application of the Newton-Raphson method failed to achieve convergence in most cases. This is why the EM algorithm is considered stable. As a numerically stable method, it avoids overshooting or undershooting a maximizer of likelihood.

In addition, the results indicated that when the methods both converged to the same maximum likelihood solution, the EM algorithm was faster in terms of the CPU time taken to reach convergence although fewer iterations were required by the Newton-Raphson method. The CPU time required per iteration was overwhelmingly shorter for the EM algorithm. The EM algorithm is generally said to suffer from slow convergence, but in practice this causes no major problems. In fact, the simplicity of the EM algorithm seems to be much more attractive, considering the relatively high operating efficiency from formulating the likelihood to deriving and programming an algorithm. Many improved versions of the EM algorithm, aimed at accelerating convergence, have been proposed since DLR (1977). However, they failed to gain wide acceptance because they eliminated some aspects of the simplicity and the stability of the original EM algorithm.

The applications of the EM algorithm are broad because of its flexibility in interpreting the incompleteness of data, and the high extensibility of the application model. However, if the problem becomes complex, simple calculations in the Expectation Step (E-step) and Maximization Step (M-step) in the EM algorithm will not work well. Simulations from the model may be needed in the E-step, or a Newton-type iterative algorithm may have to be included in the M-step. In practice, the convergence time is a measure of the algorithm's success. A significant number of publications have been written about the acceleration of the EM algorithm. In the 1990s, many papers on the systematization of these extensions to the EM algorithm were explored. Rubin (1991) explains four typical algorithms based on

simulation (Multiple Imputation, Data Augmentation Algorithm, Gibbs Sampler, and Sampling/Importance Resampling Algorithm) in a unified manner based on an extended EM framework with a random number mechanism. On the M-step side, Meng and Rubin (1993) introduced the ECM algorithm. A year later, Liu and Rubin (1994) invented the ECME algorithm. The AECM was published by Meng and Dyk in 1997. Their work on the accelerated EM does not eliminate the simplicity and stability of the original EM algorithm. Therefore, many applications directly use the ECM and ECME algorithms. The first book on EM algorithm was published by McLachlan and Krishnan (1997) which covers recent topics relating to them as well. Today, the EM algorithm is a familiar statistical tool for solving real life problems in diverse fields of application.

4.1.2 Newton-Raphson method

Since the properties of the EM algorithm are contrasted with those of Newton-type methods, which are the main alternatives for the computation of Maximum Likelihood Estimates, we now give a brief review of the Newton-Raphson method. In numerical analysis, there are various techniques for finding zeros of a specified function, including the Newton-Raphson method, quasi-Newton methods, and modified Newton methods. In a statistical framework, the modified Newton methods include the scoring algorithm of Fisher and its modified version using the empirical information matrix in place of the expected information matrix.

First of all, we recall some notation from the section on maximum likelihood

and ML equations in Chapter 2. We have $L(\theta)$ as our likelihood function for θ and the observed data Y. Under regularity conditions, we can take the first and second order partial derivatives of the log likelihood function, $l(\theta) = log(L(\theta))$, with respect to the elements of parameters θ. We have

$$S(Y,\theta) = \frac{\partial \log(L(\theta))}{\partial \theta}, \tag{4.1}$$

and

$$\mathrm{I}(Y,\theta) = -\frac{\partial^2 \log(L(\theta))}{\partial \theta \partial \theta^t}. \tag{4.2}$$

The function $S(Y,\theta)$ is the gradient vector of the log likelihood function, and is called the score statistic, when θ is a null-hypothesis value. Finally, the Fisher information matrix $\mathcal{I}(\theta)$ is given by

$$\begin{aligned} \mathcal{I}(\theta) &= E_\theta\{S(Y,\theta)S^t(Y,\theta)\} \\ &= -E_\theta\{\mathrm{I}(\mathrm{Y},\theta)\}. \end{aligned} \tag{4.3}$$

The Newton-Raphson method is the best known procedure for finding the roots of an equation. Now, we attempt to apply the Newton-Raphson method for solving the likelihood equation

$$S(Y,\theta) = 0. \tag{4.4}$$

Using a linear Taylor series expansion on the current parameter $\theta^{(k)}$ for θ, we have

$$S(Y,\theta) \approx S(Y,\theta^{(k)}) - \mathrm{I}(Y,\theta^{(k)})(\theta - \theta^{(k)}). \tag{4.5}$$

A new parameter $\theta^{(k+1)}$ can be obtained when we set the right-hand side of equation (4.5) equal to zero. Hence

$$\theta^{(k+1)} = \theta^{(k)} + \mathrm{I}^{-1}(Y,\theta^{(k)})S(Y,\theta^{(k)}). \tag{4.6}$$

If the log likelihood function $l(\theta)$ is concave, then the sequence of iterates $\{\theta^{(k)}\}$ will converge to the maximum likelihood estimate $\hat{\theta}_{MLE}$. This only takes one step if the log likelihood function $l(\theta)$ is a quadratic function of θ. When the log likelihood function $l(\theta)$ is not concave, the Newton-Raphson method will not be guaranteed to converge from an arbitrary starting value. Under reasonable assumptions on L(θ) and a sufficiently accurate starting value, the sequence of iterates $\{\theta^{(k)}\}$ generated by the Newton-Raphson method has local quadratic convergence to a solution θ^* of our likelihood equation (4.4). The solution θ^* is the maximum likelihood estimate. That is, given a norm $\|.\|$, there is a constant c such that if θ^0 is sufficiently close to θ^*, then for $k = 0,\ 1,\ 2,\ \ldots$

$$\|\theta^{(k+1)} - \theta^*\| \leq c\|\theta^{(k)} - \theta^*\|^2. \tag{4.7}$$

The biggest advantage of the Newton-Raphson method is its extremely fast quadratic convergence. Since the Newton-Raphson method requires computing the Fisher information matrix I(y, $\theta^{(\mathrm{k})}$) at each iteration k, it immediately provides an estimate of the covariance matrix at its limiting value θ^*, through the inverse of the observed Fisher information matrix I(Y, θ^*). Also, the Hessian matrix is the same as the negative of the observed Fisher information matrix I(Y, $\theta^{(\mathrm{k})}$). On the other hand, the computation of each iteration will create some serious problems in applications when the dimension of data becomes large, because it requires calculating the $d \times d$ information matrix I(Y, $\theta^{(\mathrm{k})}$) at each iteration k, where $d = \dim(\theta)$. Thus, the computation required for an iteration of the Newton-Raphson method is likely to take longer and longer when the parameter dimension d increases. Furthermore, the

Newton-Raphson method in its basic form (4.6) requires an impractically accurate initial value for θ for some problems in order that the sequence of iterates $\{\theta^{(k)}\}$ converge to a solution of (4.6). The Newton-Raphson method has a tendency to head toward the local minimum as often as it heads toward a local maximum.

A method is called a Quasi-Newton method if the solution of (4.6) takes the form

$$\theta^{(k+1)} = \theta^{(k)} - A^{-1}S(Y, \theta), \tag{4.8}$$

where A is used as an approximation to the Hessian matrix of $l(\theta)$. The advantage of the Quasi-Newton methods is that they may avoid the explicit evaluation of the Hessian matrix of the log likelihood function at each iteration.

4.1.3 EM algorithm

The EM algorithm is a method for solving incomplete data problems iteratively based on a complete data framework. The idea is simple. Assume that Y is a p-dimensional random vector corresponding to the observed data, having the probability density function $f(y, \theta)$, where θ is a vector of unknown parameters within the parameter space Θ. Let Z denote the random vector containing the missing data portion, and let $X = (Y, Z)$ denote the vector containing both the observed and missing data, called the complete data. Denote by $f_X(x, \theta)$ the probability density function of X.

Let $l_X(\theta) = \log f_X(X, \theta)$, which is the log likelihood function based on the complete data, and $l(\theta) = \log f(Y, \theta)$, 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 parameter point achieving the maximum of $l(\theta)$.

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

$$\theta^{(k+1)} = \arg \max_{\theta \in \Theta} E(l_X(\theta)|Y, \theta^{(k)}) \tag{4.9}$$

Equation (4.9) can be divided into the E-step and the M-step as follows:

E-step: Calculate the conditional expectation of complete data log likelihood given the observation Y and the k'th temporary value of parameter $\theta^{(k)}$:

$$Q(\theta, \theta^{(k)}) = E(l_X(\theta)|Y, \theta^{(k)}) \tag{4.10}$$

M-step: Find $\theta^{(k+1)}$ to maximize $Q(\theta,\ \theta^{(k)})$ regarded as a function of θ with $\theta^{(k)}$ fixed:

$$Q(\theta^{(k+1)}, \theta^{(k)}) \geq Q(\theta, \theta^{(k)}) \tag{4.11}$$

for all $\theta \in \Theta$.

The E-step and M-step are alternated repeatedly and stop according to the smallness of changes in $(\theta^{(k+1)} - \theta^{(k)})$. Dempster, Laird and Rubin (1977) demonstrated that the incomplete data likelihood function $l(\theta)$ is not decreased in each

EM iteration process, that is,

$$l(\theta^{(k+1)}) \geq l(\theta^{(k)}) \tag{4.12}$$

for $k = 0, 1, 2, \ldots$. Hence, convergence of $l(\theta^{(k)})$ values must be obtained if they are bounded above. This aspect is useful in debugging programs for the EM algorithm.

A GEM algorithm replaces the M-step in the EM algorithm with a step to find $\theta^{(k+1)}$ which satisfies the following formula:

$$Q(\theta^{(k+1)}, \theta^{(k)}) \geq Q(\theta^{(k)}, \theta^{(k)}) \tag{4.13}$$

This indicates that it is not always necessary to find the maximum of the Q-function in the M-step, and that it is sufficient to find $\theta^{(k+1)}$ updating it to a larger value. The likelihood $L(\theta)$ is not decreased after a GEM iteration, and so a GEM sequence of likelihood values must converge if bounded above. Therefore, in cases where maximization within the M-step is sought by using the Newton-Raphson method, etc., in M-step, it is possible to stop after just one iteration. Lange (1995) advocated this method as a gradient algorithm.

4.2 REFM_1 and EM Algorithm

4.2.1 REFM_1

The Random Effect Factor Model 1 (REFM_1) is

$$\begin{aligned} y_i &= \sum_{k=1}^{q} c_{ik} P_k + \epsilon_i \qquad (i = 1, 2, \ldots, n), \\ &= \pi_1 c_i + \epsilon_i \end{aligned}$$

where all assumptions regarding to y_i, c_i, π_i, and ϵ_i are specified in Chapter 2. Let X be the complete data, which includes observation vectors Y_i and unobservable vectors $c_i i = 1, 2, \cdots, n$. That is, $X = (Y,\ c)$. Thus, the complete data X becomes a $(p+q)$-dimensional vector. It is assumed that $X_1,\ X_2,\ \ldots,\ X_n$ are independently and identically distributed, and that c_i are independently and identically normally distributed with mean $\underline{a}$ and covariance matrix B_1; that is,

$$c_i \sim \mathcal{N}_q(\underline{a}, B_1), \tag{4.14}$$

where $\underline{a} = (a_1,\ a_2,\ \ldots,\ a_q)^t$, and $B_1 = Diag(b_1^2,\ b_2^2 \ldots\ b_q^2)$. The vectors c_i are independent of the errors ϵ_i, which are assumed to be independently and identically distributed as $\mathcal{N}_p(0, \sigma^2\, I_p)$. Given the unobservable random effect c_i, the conditional probability distribution over y_i is given by

$$y_i | c_i \sim \mathcal{N}_p(\pi_1 c_i, \sigma^2\, I_p). \tag{4.15}$$

Unconditionally, $\{y_i\}$ is independently and identically distributed with

$$y_i \sim \mathcal{N}_p(\pi_1 \underline{a}, \pi_1 B_1 \pi_1^t + \sigma^2 I_p). \tag{4.16}$$

Since the probability density function of the complete data X can be written as $f(x) = f(y|c)f(c)$ with

$$\begin{aligned} f(y|c) &= (2\pi)^{-p/2} |\sigma^2 I_p|^{-1/2} \exp\{-\frac{1}{2}(y - \pi_1 c)^t (\sigma^2 I_p)^{-1} (y - \pi_1 c)\} \\ &= (2\pi)^{-p/2} (\sigma^2)^{-p/2} \exp\{-\frac{1}{2\sigma^2} \|y - \pi_1 c\|^2\}, \end{aligned} \tag{4.17}$$

and

$$f(c) = (2\pi)^{-q/2} |B_1|^{-1/2} \exp\{-\frac{1}{2}(c - \underline{a})^t B_1^{-1} (c - \underline{a})\}, \tag{4.18}$$

then

$$f(x) = (2\pi)^{-\frac{p+q}{2}}(\sigma^2)^{-\frac{p}{2}}(\prod_{k=1}^{q} b_k^2)^{-\frac{1}{2}} \exp\{-\frac{1}{2\sigma^2}\|y-\pi_1 c\|^2 - \frac{1}{2}(c-\underline{a})^t B_1^{-1}(c-\underline{a})\}. \quad (4.19)$$

The complete data log likelihood function is

$$\begin{aligned} l_X(\theta) &= \log \prod_{i=1}^{n} f(x_i), \\ &= -\frac{n(p+q)}{2}\log(2\pi) - \frac{np}{2}\log\sigma^2 - \frac{n}{2}\sum_{k=1}^{q}\log b_k^2 \\ &\quad -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\|y_i - \pi_1 c_i\|^2 - \frac{1}{2}\sum_{i=1}^{n}(c_i - \underline{a})^t B_1^{-1}(c_i - \underline{a}). \end{aligned} \quad (4.20)$$

Here, our parameters are $\theta = (\underline{a}, \mathrm{Diag}(B_1), \sigma^2, \pi_1)$. We can express

$$\begin{aligned} \sum_{i=1}^{n}\|y_i - \pi_1 c_i\|^2 &= \sum_{i=1}^{n}(y_i - \pi_1 c_i)^t(y_i - \pi_1 c_i), \\ &= \sum_{i=1}^{n}(y_i^t y_i - y_i^t \pi_1 c_i - c_i^t \pi_1^t y_i + c_i^t \pi_1^t \pi_1 c_i), \\ &= \sum_{i=1}^{n}\{tr(y_i y_i^t) - 2tr(y_i c_i^t \pi_1^t) + tr(c_i^t c_i)\}, \\ &= n\ \mathrm{tr}(C_{yy}) - 2n\ \mathrm{tr}(C_{yc}\pi_1^t) + n\ \mathrm{tr}(C_{cc}), \end{aligned} \quad (4.21)$$

where $C_{yy} = n^{-1}\sum_{i=1}^{n} y_i y_i^t, C_{yc} = n^{-1}\sum_{i=1}^{n} y_i c_i^t$, and $C_{cc} = n^{-1}\sum_{i=1}^{n} c_i c_i^t$, and also

$$\begin{aligned} \sum_{i=1}^{n}(c_i - \underline{a})^t B_1^{-1}(c_i - \underline{a}) &= \sum_{i=1}^{n}\{c_i^t B_1^{-1} c_i + \underline{a}^t B_1^{-1}\underline{a} - 2\underline{a}^t B_1^{-1} c_i\}, \\ &= n\ \mathrm{tr}(C_{cc}B_1^{-1}) - 2n\underline{a}^t B_1^{-1}\bar{c} + n\underline{a}^t B_1^{-1}\underline{a}, \end{aligned} \quad (4.22)$$

where $\bar{c} = n^{-1}\sum_{i=1}^{n} c_i$. After substituting (4.21) and (4.22) into the complete data log likelihood function (4.20), we immediately find

$$\begin{aligned} l_X(\theta) &= -\frac{n(p+q)}{2}\log(2\pi) - \frac{np}{2}\log\sigma^2 - \frac{n}{2}\sum_{k=1}^{q}\log b_k^2 - \frac{n}{2\sigma^2}\ \mathrm{tr}(C_{yy}) \\ &\quad + \frac{n}{\sigma^2}\ \mathrm{tr}(C_{yc}\pi_1^t) \\ &\quad -\frac{n}{2\sigma^2}\ \mathrm{tr}(C_{cc}) - \frac{n}{2}\ \mathrm{tr}(C_{cc}B_1^{-1}) + n\underline{a}^t B_1^{-1}\bar{c} - \frac{n}{2}\underline{a}^t B_1^{-1}\underline{a}. \end{aligned} \quad (4.23)$$

Thus, $l_X(\theta)$ for $\theta = (\underline{a},\ B_1,\ \sigma^2,\ \pi_1)$, belongs to an exponential family, and the sufficient statistics are $\bar{c},\ C_{yy},\ C_{yc}$, and C_{cc}.

Before we can calculate the E-step, we must find out what is the distribution of unobservable c given the observed data Y and current parameter $\theta = \theta^{(k)}$. We know $c \sim \mathcal{N}_q(\underline{a},\ B_1)$, and $Y \sim \mathcal{N}_p(\pi_1\underline{a},\ \pi_1 B_1 \pi_1^t + \sigma^2 I_p)$, but Y and c are not independent. Their relationship is through the $\mathbf{REFM}_1$. By their joint normality, there must exist a $q \times p$ transformation matrix D such that $c - DY$ is uncorrelated with Y. Thus, the matrix D must satisfy the equation:

$$\begin{aligned} 0 &= E(c - DY - E(c - DY))(Y - EY)^t \\ &= E(c - \underline{a} - D(Y - \pi_1\underline{a}))(Y - \pi_1\underline{a})^t \\ &= E(c - \underline{a})(\pi_1 c + \epsilon - \pi_1\underline{a})^t - DE(Y - \pi_1\underline{a})(Y - \pi_1\underline{a})^t \\ &= E(c - \underline{a})(c - \underline{a})^t\pi_1^t + E(c - \underline{a})\epsilon^t - DVar\{Y\} \\ &= B_1\pi_1^t - D(\pi_1 B_1 \pi_1^t + \sigma^2 I_p). \end{aligned} \tag{4.24}$$

It is easy to solve this matrix equation for D as a function of θ, that is,

$$D = B_1\pi_1^t(\pi_1 B_1 \pi_1^t + \sigma^2 I_p)^{-1}. \tag{4.25}$$

Remark: In practice, we prefer to calculate a lower dimensional $q \times q$ matrix instead of a higher dimensional $p \times p$ matrix, by applying Woodbury's Identity (Rubin & Thayer, 1982, page 72):

$$(\tau^2 + \beta^t R \beta)^{-1} = \tau^{-2} - (\tau^{-2}\beta^t)(R^{-1} + \beta\tau^{-2}\beta^t)^{-1}(\beta\tau^{-2}) \tag{4.26}$$

When $R = B_1, \tau^2 = \sigma^2 I_p$, and $\beta = \pi^t$, we have the following equation:

$$(\pi_1 B_1 \pi_1^t + \sigma^2 I_p)^{-1} = \frac{1}{\sigma^2} I_p - \left(\frac{1}{\sigma^2}\right)^2 \pi_1 \left(B_1^{-1} + \frac{1}{\sigma^2} I_q\right)^{-1} \pi_1^t. \tag{4.27}$$

We will apply Woodbury's Identity (4.27) in a later section. □

The expectation of unobservable c given the observed data Y is

$$\begin{aligned} E(c|Y) &= E(c - DY + DY|Y) \\ &= E(c - DY|Y) + E(DY|Y) \\ &= E(c - DY) + DY \\ &= \underline{a} - D\pi_1\underline{a} + DY \\ &= \underline{a} + D(Y - \pi_1\underline{a}), \end{aligned} \tag{4.28}$$

and since the relationship between c and Y is from our $\mathbf{REFM}_1$, we can calculate the covariance between c and Dy as follows:

$$\begin{aligned} \text{Cov}\ (c,\ DY) &= \text{Cov}(c - DY + DY,\ DY) \\ &= \text{Cov}(c - DY,\ DY) + \text{Cov}(DY,\ DY) \\ &= Var(DY) \\ &= DVar(Y)D^t. \end{aligned} \tag{4.29}$$

The conditional covariance of unobservable c given observed data Y is

$$\begin{aligned} Var(c|Y) &= Var(c - DY + DY|Y) \\ &= Var(c - DY|Y) + Var(DY|Y) \\ &= Var(c - DY) \\ &= Var(c) + DVar(Y)D^t - 2Cov(c, DY) \\ &= B_1 - DVar(Y)D^t \\ &= B_1 - D\pi_1 B_1 \end{aligned}$$

$$= \quad (I_q - D\pi_1)B_1. \tag{4.30}$$

Therefore, given observed the Y_i, the conditional probability distribution of c_i is specified by

$$c_i|Y_i \sim \mathcal{N}_q(\underline{a} + D(y - \pi_1\underline{a}), (I_q - D\pi_1)B_1) \tag{4.31}$$

where $D = B_1\pi_1^t(\pi_1 B_1 \pi_1^t + \sigma^2 I_p)^{-1}$.

4.2.2 E-Step

Now, we calculate the conditional expectation of the complete data log likelihood function, $l_X(\theta)$, given the observation Y and the k'th temporary values of the parameter $\theta = \theta^{(k)}$:

$$\begin{aligned} Q(\theta, \theta^{(k)}) &= E(l_X(\theta)|Y,\ \theta^{(k)}), \\ &= -\frac{n(p+q)}{2}\log(2\pi) - \frac{np}{2}\log\sigma^2 - \frac{n}{2}\sum_{k=1}^{q}\log b_k^2 \\ &\quad -\frac{n}{2\sigma^2}\ \mathrm{tr}(C_{yy}^*) + \frac{n}{\sigma^2}\ \mathrm{tr}(C_{yc}^*\pi_1^t) - \frac{n}{2\sigma^2}\ \mathrm{tr}(C_{cc}^*) \\ &\quad -\frac{n}{2}\ \mathrm{tr}(C_{cc}^* B_1^{-1}) + n\underline{a}^t B_1^{-1}\bar{c}^* - \frac{n}{2}\underline{a}^t B_1^{-1}\underline{a}, \end{aligned} \tag{4.32}$$

where $\bar{c}^* = E(\ \bar{c}|\ Y,\ \theta^{(k)})$, $C_{yy}^* = E(\ C_{yy}|\ Y,\ \theta^{(k)})$, $C_{yc}^* = E(\ C_{yc}|\ Y,\ \theta^{(k)})$, and $C_{cc}^* = E(\ C_{cc}|\ Y,\ \theta^{(k)})$. Given the current parameter $\theta^{(k)}$ and observed data Y, we have the transformation matrix

$$D^{(k)} = B_1^{(k)}(\pi_1^{(k)})^t(\pi_1^{(k)}B_1^{(k)}(\pi_1^{(k)})^t + (\sigma^2)^{(k)}I_p)^{-1}. \tag{4.33}$$

We calculate the conditional expectation of the sufficient statistics as follows:

$$C_{yy}^* = C_{yy}, \tag{4.34}$$

$$\begin{aligned}
\bar{c}^* &= \frac{1}{n}\sum_{i=1}^{n} E(\ c_i|\ Y,\ \theta^{(k)}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\{\underline{a}^{(k)} + D^{(k)}(y_i - \pi_1^{(k)}\underline{a}^{(k)})\} \\
&= \underline{a}^{(k)} + D^{(k)}(\bar{y} - \pi_1^{(k)}\underline{a}^{(k)}), \qquad (4.35)
\end{aligned}$$

where $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$,

$$\begin{aligned}
C_{yc}^* &= \frac{1}{n}\sum_{i=1}^{n} y_i E(c_i^t|Y,\theta^{(k)}) \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i(\underline{a}^{(k)} + D^{(k)}(y_i - \pi_1^{(k)}\underline{a}^{(k)}))^t \\
&= \bar{y}(\underline{a}^{(k)})^t + \frac{1}{n}\sum_{i=1}^{n} y_i y_i^t (D^{(k)})^t - \frac{1}{n}\sum_{i=1}^{n} y_i(\underline{a}^{(k)})^t(\pi_1^{(k)})^t(D^{(k)})^t \\
&= \bar{y}(\underline{a}^{(k)})^t(I_q - (\pi_1^{(k)})^t(D^{(k)})^t) + C_{yy}(D^{(k)})^t, \qquad (4.36)
\end{aligned}$$

and

$$\begin{aligned}
C_{cc}^* &= \frac{1}{n}\sum_{i=1}^{n} E(c_i c_i^t|Y,\theta^{(k)}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\{Var(c_i|Y,\theta^{(k)}) + E(c_i|y,\theta^{(k)})E^t(c_i|y,\theta^{(k)})\} \\
&= (I_q - D^{(k)}\pi_1^{(k)})B_1^{(k)} + \frac{1}{n}\sum_{i=1}^{n}(\underline{a}^{(k)} + D^{(k)}(y_i - \pi_1^{(k)}\underline{a}^{(k)}))(\underline{a}^{(k)} + D^{(k)}(y_i - \pi_1^{(k)}\underline{a}^{(k)}))^t \\
&= (I_q - D^{(k)}\pi_1^{(k)})B_1^{(k)} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}((I_q - D^{(k)}\pi_1^{(k)})\underline{a}^{(k)} + D^{(k)}y_i)((\underline{a}^{(k)})^t(I_q - D^{(k)}\pi_1^{(k)})^t + y_i^t(D^{(k)})^t) \\
&= (I_q - D^{(k)}\pi_1^{(k)})\underline{a}^{(k)}(\underline{a}^{(k)})^t(I_q - D^{(k)}\pi_1^{(k)})^t + D^{(k)}\bar{y}(\underline{a}^{(k)})^t(I_q - D^{(k)}\pi_1^{(k)})^t \\
&\quad + (I_q - D^{(k)}\pi_1^{(k)})(B_1^{(k)} + \underline{a}^{(k)}\bar{y}^t(D^{(k)})^t) + D^{(k)}C_{yy}(D^{(k)})^t. \qquad (4.37)
\end{aligned}$$

We consider the E-step of the $(k+1)$'th iteration of the EM algorithm, where $\theta^{(k)}$ denotes the value of θ after the k'th EM iteration.

4.2.3 M-Step

After we finish calculating the conditional expectation of the complete data log likelihood function, we implement an M-step, which maximizes $Q(\theta,\ \theta^{(k)})$, that is, maximizes the equation (4.32) with respect to θ. There are four parameters in θ within our Q-function: $\underline{a},\ B_1,\ \sigma^2$, and π_1. We will work on $\underline{a}$ first, and take the partial derivative of $Q(\theta, \theta^{(k)})$ with regard to $\underline{a}$,

$$\frac{\partial Q(\theta, \theta^{(k)})}{\partial \underline{a}} = nB_1^{-1}\bar{c}^* - \frac{n}{2} \cdot 2B_1^{-1}\underline{a}. \tag{4.38}$$

Setting the expressions equal to 0, we can solve for our first parameter $\underline{a}$,

$$\hat{\underline{a}} = \bar{c}^*. \tag{4.39}$$

Secondly, for fixed k, $k = 1, 2, \ldots, q$, take the partial derivative of $Q(\theta, \theta^{(k)})$ with respect to b_k^2,

$$\frac{\partial Q(\theta, \theta^{(k)})}{\partial b_k^2} = -\frac{n}{2}\frac{1}{b_k^2} - \frac{n}{2}C^*_{cc,kk} \cdot \left(-\frac{1}{b_k^4}\right) - n\frac{a_k\bar{c}_k^*}{b_k^4} - \frac{n}{2}\left(-\frac{a_k^2}{b_k^4}\right). \tag{4.40}$$

since

$$\begin{cases} tr(C^*_{cc}B_1^{-1}) = \sum_{k=1}^q C^*_{cc,kk}/b_k^2 \\ \underline{a}^t B_1^{-1}\bar{c}^* = \sum_{k=1}^q a_k\bar{c}_k^*/b_k^2 \\ \underline{a}^t B_1^{-1}\underline{a} = \sum_{k=1}^q a_k^2/b_k^2 \end{cases} \tag{4.41}$$

where a_k and $\bar{c}_k^*$ are the k'th elements of the q dimensional vectors $\underline{a}$ and $\bar{c}$, and b_k^2 and $C^*_{cc,kk}$ are the k'th elements on the diagonal of $q \times q$ matrix B_1 and C^*_{cc}. Setting the expressions (4.40) equal to zero, we have

$$\begin{aligned} \hat{b}_k^2 &= C^*_{cc,kk} + \hat{a}_k^2 - 2\hat{a}_k\bar{c}_k^*, \\ &= C^*_{cc,kk} - \hat{a}_k^2. \end{aligned} \tag{4.42}$$

Therefore, the second parameter B_1 is estimated by

$$\hat{B}_1 = Diag(C^*_{cc}) - Diag(\underline{\hat{a}}\hat{a}^t). \tag{4.43}$$

The third parameter σ^2 is a scalar variable. Take the partial derivative of $Q(\theta,\ \theta^{(k)})$ with regard to σ^2 to find

$$\begin{aligned}\frac{\partial Q(\theta,\theta^{(k)})}{\partial \sigma^2} &= -\frac{np}{2}\frac{1}{\sigma^2} + \frac{n}{2}(\frac{1}{\sigma^2})^2 \cdot tr(C^*_{yy}) - n(\frac{1}{\sigma^2})^2 \cdot tr(C^*_{yc}\pi_1^t) \\ &\quad + \frac{n}{2}(\frac{1}{\sigma^2})^2 \cdot tr(C^*_{cc}).\end{aligned} \tag{4.44}$$

Setting (4.44) equal to zero, we have

$$\hat{\sigma}^2 = \frac{1}{p}(tr(C^*_{yy}) + tr(C^*_{cc}) - 2 \cdot tr(C^*_{yc} \cdot \hat{\pi}_1^t)) \tag{4.45}$$

From equation (4.45), we see that $\hat{\sigma}^2$ is a function of $\hat{\pi}_1$. Thus, we must find an estimate of π_1. Then we can obtain the estimate of σ^2. We use the following Lemmas to obtain the estimate of π_1.

Lemma 11 *Assume that $q > 1$ and $\lambda_k \geq 0,\ k = \ 1,\ 2,\ \ldots,\ q$, and that $\{v_k : 1 \leq k \leq q\}$ is an orthonormal set of vectors. Then, the unique maximizer of $tr(P\Lambda^{\frac{1}{2}}M)$ over the class $\mathcal{M}$ of $q \times q$ matrices with the property $MM^t = I_q$ is $M = P^t$. That is,*

$$\max_{M:MM^t=I_q} tr(P\Lambda^{\frac{1}{2}}M) = \sum_{k=1}^{q}\sqrt{\lambda_k}, \tag{4.46}$$

where $P = (v_1, v_2, ..., v_q)$ is a $p \times q$ matrix with columns v_i, and $\Lambda = Diag(\lambda_1,\ \lambda_2,\ \ldots,\ \lambda_q)$.

Proof. Define vectors $M_k,\ 1 \leq k \leq q$, as the row vectors of M. Then

$$\mathrm{tr}(P\Lambda^{\frac{1}{2}}M) \ = \ \mathrm{tr}(MP\Lambda^{\frac{1}{2}})$$

$$= \sum_{k=1}^{q} \sqrt{\lambda}_k M_k v_k. \tag{4.47}$$

Since both v_k and M_k are unit vectors, apply the Cauchy-Schwartz inequality to find

$$M_k v_k \leq \|v_k\|^2 \cdot \|M_k\|^2 = 1 \tag{4.48}$$

with equality if and only if $M_k = v_k^t,\ k = 1,\ 2,\ \ldots,\ q$. Therefore, the maximum of $\mathrm{tr}(P\Lambda^{\frac{1}{2}}M)$ over $\mathrm{M} \in \mathcal{M}$ is $M = P^t$. The unique maximum of $tr(P\Lambda^{\frac{1}{2}}M)$ is

$$\begin{aligned} \max_{M:MM^t=I_q} \mathrm{tr}(P\Lambda^{\frac{1}{2}}M) &= \mathrm{tr}(P\Lambda^{\frac{1}{2}}P^t) \\ &= \mathrm{tr}(\Lambda^{\frac{1}{2}}) \\ &= \sum_{k=1}^{q} \sqrt{\lambda_k}. \end{aligned}$$ □

Lemma 12 *Given $q > 1$, let $\{\lambda_k : 1 \leq k \leq q\}$ be the eigenvalues of a $q \times q$ full rank symmetric matrix Z and let $\{v_k : 1 \leq k \leq q\}$ be the orthonormal eigenvectors corresponding to eigenvalues $\{\lambda_k : 1 \leq k \leq q\}$. Then the maximum over $q \times q$ matrices W of $tr(WZ)$ subject to the constraint $W^tZW = I_q$ is*

$$W = P\Lambda^{-1/2}P^t, \tag{4.49}$$

where $P = (v_1,\ v_2,\ \ldots,\ v_q)$, and $\Lambda = Diag(\lambda_1,\ \lambda_2,\ \ldots,\ \lambda_q)$.

Proof. By the spectral decomposition theorem,

$$Z = \sum_{k=1}^{q} \lambda_k v_k v_k^t = P\Lambda P^t. \tag{4.50}$$

Define $W_0 = (v_1/\lambda_1^{1/2},\ v_2/\lambda_2^{1/2},\ \ldots,\ v_q/\lambda_q^{1/2}) = P\Lambda^{-1/2}$. Then, using $P^tP = I_q$ we have

$$W_0^t Z W_0 = (P\Lambda^{-1/2})^t P\Lambda P^t P\Lambda^{-1/2} = I_q. \tag{4.51}$$

If $W = W_0 M$, where M is any $q \times q$ orthogonal matrix, then the constraint function of W is

$$W^t Z W = (W_0 M)^t Z W_0 M = M^t (W_0^t Z W_0) M = M^t M = I_q, \tag{4.52}$$

and

$$\mathrm{tr}(WZ) = \mathrm{tr}(W_0 M Z) = \mathrm{tr}(Z W_0 M). \tag{4.53}$$

Note that (4.52) also shows that the only matrices $W = W_0 M$ satisfying the constraint are those with orthogonal matrices M. Now, our problem transfers from finding the maximum of $\mathrm{tr}(WZ)$ with $W^t Z W = I_q$ to finding the maximum of $\mathrm{tr}(Z W_0 M)$ with $M M^t = I_q$. Since

$$Z W_0 M = P \Lambda P^t P \Lambda^{-1/2} M = P \Lambda^{1/2} M, \tag{4.54}$$

we can apply Lemma 11: $M = P^t$ uniquely attains the maximum of $\mathrm{tr}(P \Lambda^{1/2} M)$ subject to $M^t M = I_q$. Thus,

$$W = W_0 M = P \Lambda^{-1/2} P^t. \qquad \square$$

Lemma 13 *Let A be a $p \times q$ matrix with orthonormal columns, i.e., $A^t A = I_q$, where $q \leq p$. Let B be a $p \times q$ matrix with rank q and, let $V_1 = col(B)$ be the column space of B, where $V_1 \subset R^p$. Then*

(1) *There exists a maximum for $tr(A^t B)$.*

(2) *Let $\hat{A}$ be the maximizer of $tr(A^t B)$, that is,*

$$\max_A tr(A^t B) = tr(\hat{A}^t B). \tag{4.55}$$

Then the space $\hat{V}_1$ spanned by the columns of $\hat{A}$ is V_1, that is,

$$\hat{V}_1 = V_1. \tag{4.56}$$

(3) *When q=1, $A = B/\|B\|$ is the maximizer of $tr(A^tB)$.*

(4) *When $q > 1$, the maximizer over A of $tr(A^tB)$ subject to the constraint $A^tA = I_q$ is*

$$\hat{A} = BP\Lambda^{-1/2}P^t, \tag{4.57}$$

where $P = (v_1,\ v_2,\ \ldots,\ v_q)$, and $\Lambda = Diag(\lambda_1,\ \lambda_2,\ \ldots,\ \lambda_q)$, λ_k and v_k are eigenvalues and eigenvectors of B^tB.

Proof.

(1) Let $A_1,\ A_2, \ldots,\ A_q$ be the q column vectors of A, that is, $A = (A_1,\ A_2,\ \ldots,\ A_q)$, and let $B_1,\ B_2,\ \ldots,\ B_q$ be the q column vectors of B. For $1 \le k \le q$, we apply the Cauchy-Schwarz inequality

$$A_k^tB_k \le \|A_k\|\|B_k\| \le \frac{\|A_k\|^2 + \|B_k\|^2}{2} = \frac{1 + \|B_k\|^2}{2}. \tag{4.58}$$

Also,

$$\begin{aligned} \mathrm{tr}(A^tB) &= \sum_{k=1}^{q} A_k^tB_k \\ &\le \sum_{k=1}^{q} \frac{1 + \|B_k\|^2}{2}. \end{aligned} \tag{4.59}$$

Since $\mathrm{tr}(A^tB)$ is a continuous function of $A_1,\ A_2,\ \ldots,\ A_q$ and has an upper bound, and the set of matrices $\{A_i : i = 1,\ 2,\ \ldots,\ q\}$ with orthonormal columns is compact, there must exist a maximum.

(2) Since the rank of B is q, the columns of B are linearly independent,

$$V_1 = \mathrm{span}\{B_1, B_2, \ldots, B_q\}, \tag{4.60}$$

and V_1 is a q dimensional space. To maximize $tr(A^tB)$ subject to the constraint $A^tA = I_q$, we can apply the Lagrange multiplier method,

$$L = \sum_{k=1}^{q} A_k^t B_k + \sum_{k=1}^{q} \lambda_k (A_k^t A_k - 1) + \sum_{j=1}^{q} \sum_{k=1, k\neq j}^{q} \lambda_{jk} (A_j^t A_k) \tag{4.61}$$

where $(\lambda_{jk})_{q\times q}$ is a symmetric matrix of unknowns with $\lambda_k \equiv \lambda_{kk}$. For $1 \leq k \leq q$, we take the gradient of the Lagrange function L with respect to A_k, then set $\nabla_{A_k} L$ to zero. We have the following equation:

$$B_k + 2\hat{\lambda}_k \hat{A}_k + 2 \sum_{j=1, j\neq k}^{q} \hat{\lambda}_{jk} \hat{A}_j = 0. \tag{4.62}$$

Multiply by $\hat{A}_i^t$ on both sides of (4.62) to obtain

$$\hat{A}_i^t B_k + 2\hat{\lambda}_k \delta_{ik} + 2 \sum_{j=1, j\neq k}^{q} \hat{\lambda}_{jk} \delta_{ij} = 0. \tag{4.63}$$

If $i = k$, the equation (4.63) implies

$$\hat{\lambda}_k = -\frac{1}{2} \hat{A}_k^t B_k. \tag{4.64}$$

For $i \neq k$, the equation (4.63) implies

$$\hat{\lambda}_{ik} = -\frac{1}{2} \hat{A}_i^t B_k. \tag{4.65}$$

Substitute equations (4.64) and (4.65) into equation (4.62), to find for all k

$$\begin{aligned}
& B_k - \hat{A}_k \hat{A}_k^t B_k - \sum_{j=1, j\neq k}^{q} \hat{A}_j \hat{A}_j^t B_k = 0 \\
\Rightarrow \quad & (I_p - \hat{A}_k \hat{A}_k^t - \sum_{j=1, j\neq k}^{q} \hat{A}_j \hat{A}_j^t) B_k = 0 \\
\Rightarrow \quad & (I_p - \hat{A}\hat{A}^t) B_k = 0 \\
\Rightarrow \quad & (I_p - \hat{A}\hat{A}^t) B = 0
\end{aligned} \tag{4.66}$$

Thus,

$$\text{span}\{\hat{A}_1,\ \hat{A}_2,\ \ldots,\ \hat{A}_q\} = \text{span}\{B_1,\ B_2,\ \ldots,\ B_q\}, \tag{4.67}$$

that is,

$$\hat{V}_1 = V_1. \tag{4.68}$$

(3) When $q = 1$, A and B are p-dimensional vectors, and $\text{tr}(A^tB) = A^tB$. The maximum of A^tB is achieved if only if A and B are in the same direction. Also, A is a unit vector. Therefore,

$$A = B/\|B\|. \tag{4.69}$$

(4) When $q > 1$, since $\hat{V}_1 = V_1$, then there exists a $q \times q$ matrix W such that $\hat{A} = BW$. Hence,

$$\hat{A}^t\hat{A} = W^tB^tBW = W^tZW = I_q,$$

where $Z = B^tB$, and

$$tr(\hat{A}^tB) = tr(W^tB^tB) = tr(W^tZ).$$

The problem of maximizing $\text{tr}(\hat{A}^tB)$ subject to the constraint $\hat{A}^t\hat{A} = I_q$ is equivalent to the problem of maximizing $\text{tr}(W^tZ)$ with constraint $W^tZW = I_q$. Applying Lemma 12, we have

$$W = P\Lambda^{-1/2}P^t.$$

Thus,

$$\hat{A} = BP\Lambda^{-1/2}P^t. \qquad \square$$

Now, we return to the M-step, and maximize the Q-function (4.32) with respect to the parameter π_1. There is only one term in the Q-function which involves π_1, which is the term $\mathrm{tr}(C_{yc}^* \pi_1^t)$. In order to maximize the Q-function, we only need maximize $\mathrm{tr}(C_{yc}^* \pi_1^t)$ subject to the constraint $\pi_1^t \pi_1 = I_q$. We apply Lemma 13. The constrained maximum of π_1 is

$$\hat{\pi}_1 = \begin{cases} C_{yc}^* / \|C_{yc}^*\| & \text{for } q = 1, \\ C_{yc}^* P \Lambda^{-1/2} P^t & \text{for } q > 1. \end{cases} \tag{4.70}$$

where Λ is a diagonal matrix of eigenvalues of the $q \times q$ matrix $Z = (C_{yc}^*)^t C_{yc}^*$, and the column vectors of P are eigenvectors corresponding to these eigenvalues.

Before we finish the M-step, we need to verify that the estimates of all parameters $\underline{a},\ B_1,\ \sigma^2$, and π_1 are jointly maximizing the Q-function. First, for fixed $\underline{a},\ B_1,\ \sigma^2$, the Q-function achieves its maximum with respect to π_1 when $\pi_1 = \hat{\pi}_1$ because of the Lagrange multiplier method and Cauchy-Schwartz inequality. Secondly, if $B_1,\ \sigma^2$, and π_1 are given, the Q-function is a quadratic function of $\underline{a}$ with a negative second order coefficient. Thus, the Q-function attains the maximum value when $\underline{a} = \hat{\underline{a}}$. Third, we rewrite the Q-function when $\underline{a},\ B_1$, and π_1 are given as

$$Q(\sigma^2) = c(\underline{a},\ B_1,\ \pi_1) - \frac{np}{2} \log \sigma^2 - \frac{1}{2\sigma^2} c_1, \tag{4.71}$$

where $c_1 = E\{\sum_{i=1}^n \|y_i - \pi c_i\|^2 | Y, \theta^{(k)}\} > 0$. Since $\sigma^2 > 0$, we evaluate the boundary values when $\sigma^2 \to 0+$ or $\sigma^2 \to +\infty$. When $\sigma^2 \to +\infty$, $1/\sigma^2 \to 0$ and $\log(\sigma^2) \to +\infty$. From the expression (4.71), $Q(\sigma^2) \to -\infty$. When $\sigma^2 \to 0+$, $1/\sigma^2 \to +\infty$ and

$\log(\sigma^2) \to -\infty$. Since

$$\lim_{\sigma^2 \to 0+} \frac{\log \sigma^2}{1/\sigma^2} = \lim_{\sigma^2 \to 0+} (-\sigma^2) = 0, \tag{4.72}$$

that is, $1/\sigma^2$ goes to $+\infty$ much faster than $\log \sigma^2$ goes to $-\infty$ when $\sigma^2 \to 0+$. Therefore we have $Q(\sigma^2) \to -\infty$ since $c_1 > 0$. Noting that the Q-function is $-\infty$ at both boundary points, we conclude that the continuous Q-function reaches its maximum when $\sigma^2 = \hat{\sigma}^2$ from (4.45). Similarly, the Q-function of b_k^2 for $1 \leq k \leq q$ takes its maximum value when $\hat{b}_k^2 = C^*_{cc,kk} - a_k^2$ because $Q(b_k^2) \to -\infty$ when $b_k^2 \to 0+$ or $b_k^2 \to +\infty$. Now, we have demonstrated that all parameters have jointly achieved the maximum of the Q-function. Thus, we have completed the M-step.

4.3 Results of Estimation on Simulated Data

4.3.1 Simplifying the EM algorithm

In order to speed up the computation process, we simplify the EM algorithm by avoiding loops, reducing the high dimensional matrix computation, and simplifying the inverse matrix. After examining all formulas, there are three places we can simplify. Since

$$C^*_{yy} = C_{yy} = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^t, \tag{4.73}$$

and since matrix multiplication is much faster than a loop with n index values, especially when n is large, we rewrite $C^*_{yy} = n^{-1} Y Y^t$, where Y is a $p \times n$ data matrix.

The $q \times p$ matrix D appears everywhere in the EM algorithm. In order to compute D, we have to invert a $p \times p$ matrix. Specifically, in our intended data applica-

tion (the tongue dataset discussed in Chapter 5), p is extremely large. Woodbury's Identity helps us to avoid calculating a high dimensional inverse matrix. Instead, we need only to calculate a much lower dimensional $q \times q$ inverse matrix. By applying Woodbury's Identity and substituting equation (4.27) into the expression (4.25) for D, we find

$$\begin{aligned} D &= B_1\pi_1^t(\pi_1 B_1 \pi_1^t + \sigma^2 I_p)^{-1} \\ &= B_1\pi_1^t\{\frac{1}{\sigma^2}I_p - (\frac{1}{\sigma^2})^2\pi_1(B_1^{-1} + \frac{1}{\sigma^2}I_q)^{-1}\pi_1^t\} \\ &= \frac{1}{\sigma^2}B_1\pi_1^t - (\frac{1}{\sigma^2})^2 B_1(B_1^{-1} + \frac{1}{\sigma^2}I_q)^{-1}\pi_1^t \\ &= \frac{1}{\sigma^2}B_1\{I_q - \frac{1}{\sigma^2}(B_1^{-1} + \frac{1}{\sigma^2}I_q)^{-1}\}\pi_1^t \\ &= \frac{1}{\sigma^2}B_1\{I_q - (I_q + \sigma^2 B_1^{-1})^{-1}\}\pi_1^t. \end{aligned} \tag{4.74}$$

Define G as a $q \times q$ matrix, according to the equation:

$$G = \frac{1}{\sigma^2}B_1\{I_q - (I_q + \sigma^2 B_1^{-1})^{-1}\}. \tag{4.75}$$

Clearly G is a diagonal and symmetric matrix, which is very simple and fast to compute. The two inverse operations in G are both occurring on $q \times q$ diagonal matrices. Then $D = G\pi_1^t$ from equation (4.74), which is the form we use in our S-plus calculation.

We calculate the likelihood function $l(\theta)$ as a check in our EM Algorithm Splus function because Dempster, Laird, and Rubin (1977) showed that the incomplete-data likelihood function $l(\theta)$ is not decreased in each EM iteration process, that is, $l(\theta^{(k+1)}) \geq l(\theta^{(k)})$ for $k = 0, 1, 2, \ldots$. Under $\mathbf{REFM}_1$, the likelihood function $l(\theta)$ is

$$l(\theta) = -\frac{np}{2}\log(2\pi) - 2\log(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{n}(y_i - \pi_1\underline{a})^t\Sigma^{-1}(y_i - \pi_1\underline{a}), \tag{4.76}$$

where $\Sigma = \pi_1 B_1 \pi_1^t + \sigma^2 I_p$. We can simplify $l(\theta)$ by rewriting

$$\begin{aligned} & \sum_{i=1}^{n} (y_i - \pi_1 \underline{a})^t \Sigma^{-1} (y_i - \pi_1 \underline{a}) \\ = \; & tr(Y_1^t \Sigma^{-1} Y_1) \\ = \; & tr(Y_1 Y_1^t \Sigma^{-1}), \end{aligned} \tag{4.77}$$

where Y_1 is the $p \times n$ matrix whose column vectors are $y_i - \pi_1 \underline{a}$.

4.3.2 Splus function for MLE in REFM_1

We have written two Splus functions, EM_1 and EM_2, for the cases of $q = 1$ and $q > 1$. The Splus functions implement the E-step, M-step, and likelihood calculation in each EM iteration. Each function uses as inputs a data matrix Y, starting points for all parameters and a number of iterations, and outputs the parameter estimates and likelihood values. Given the data set Y, we calculate the sample mean $\bar{y}$, the sample variance S_y, and sample second moments C_{yy} before we start the iteration.

E-step: To calculate $Q(\theta, \theta^{(k)})$, the conditional expectation of the complete data log likelihood function, given the observation Y and the k'th iteration parameter $\theta^{(k)}$, we only need to calculate these conditional sufficient statistics: $\bar{c}^*, C^*_{yy}, C^*_{yc}, C^*_{cc}$. All conditional sufficient statistics depend on the matrix $D = G\pi_1^t$. We find

$$\begin{aligned} \bar{c}^* &= \underline{a} + D(\bar{y} - \pi_1 \underline{a}) \\ &= \underline{a} + G\pi_1^t (\bar{y} - \pi_1 \underline{a}) \\ &= (I_q - G)\underline{a} + G\pi_1^t \bar{y}; \end{aligned} \tag{4.78}$$

$$C^*_{yc} = \bar{y}\underline{a}^t (I_q - \pi_1^t D^t) + C_{yy} D^t$$

$$
\begin{aligned}
&= \bar{y}\underline{a}^t(I_q - \pi_1^t\pi_1 G) + C_{yy}\pi_1 G \\
&= \bar{y}\underline{a}^t(I_q - G) + C_{yy}\pi_1 G; \qquad (4.79)
\end{aligned}
$$

and

$$
\begin{aligned}
C_{cc}^* &= (I_q - D\pi_1)\underline{a}\underline{a}^t(I_q - D\pi_1)^t + D\bar{y}\underline{a}^t(I_q - D\pi_1)^t \\
&\quad +(I_q - D\pi_1)(B_1 + \underline{a}\bar{y}^t D^t) + DC_{yy}D^t \\
&= (I_q - G)\underline{a}\underline{a}^t(I_q - G) + G\pi_1^t\bar{y}\underline{a}^t(I_q - G) \\
&\quad +(I_q - G)(B_1 + \underline{a}\bar{y}^t\pi_1 G) + G\pi_1^t C_{yy}\pi_1 G. \qquad (4.80)
\end{aligned}
$$

The equations (4.75), (4.78), (4.79), and (4.80) complete our E-step computation. Specifically, when $q = 1$, the parameter θ is a, b^2, σ^2, and P, the equations for the E-step are as follows:

$$
\begin{cases}
G = \frac{b^2}{b^2+\sigma^2} \\
c^* = (1-G)a + GP^t\bar{y} \\
C_{yc}^* = a(1-G)\bar{y} + GC_{yy}P \\
C_{cc}^* = a^2(1-G)^2 + (1-G)(b^2 + 2aGP^t\bar{y}) + G^2P^tC_{yy}P.
\end{cases} \qquad (4.81)
$$

M-step: To maximize the function $Q(\theta, \theta^{(k)})$, we have taken first order partial derivatives with respect to $\underline{a}, B_1$, and σ^2. Since there is an orthonormality constraint on π_1, we apply the Lagrange multiplier method to maximize over π_1. We separate the problem into two cases, with $q = 1$ and $q > 1$, in the M-step. When $q = 1$, the

M-step is as follows:

$$\begin{cases} a = \bar{c}^* \\ b^2 = C^*_{cc} - \bar{c}^{*2} \\ P_1 = C^*_{yc}/\|C_{yc}\| \\ \sigma^2 = \frac{1}{p}(\mathrm{tr}(C^*_{yy}) + C^*_{cc} - 2\|C^*_{yc}\|). \end{cases} \tag{4.82}$$

In the case $q > 1$, the M-step includes the equations (4.39), (4.43), (4.70), and (4.45).

Likelihood function: Computing the likelihood does not affect the numerical results of the maximization in applying the EM algorithm. But, since the incomplete data likelihood is a nondecreasing function of iterations, we can use the value of likelihood as an aid to debugging the EM programs. Thus, we add the computation of likelihood inside the EM iterative procedure. Because the purpose of computing likelihood is checking monotonicity, we can drop the large constant term in the likelihood function, which relates to n. In the Splus function, we calculate likelihood in four steps:

$$\begin{cases} d = \pi_1 a \\ \Sigma = \pi_1 Diag(b^2)\pi_1 + \sigma^2 I_p \\ D = Y_1 Y_1^t \\ l = -\frac{n}{2} log|\Sigma| - \frac{1}{2}\mathrm{tr}(D\Sigma^{-1}). \end{cases} \tag{4.83}$$

4.3.3 Computational results on simulated data

In this section, we implement these Splus functions on simulated data. First, we choose the size indices ($p,\ q$) of data , the size n of the sample, and the parameters

$\theta_0 = (\underline{a},\ \underline{b}^2,\ \sigma^2,\ \pi_1)$. Since $Y \sim N_p(\pi_1\underline{a},\ \pi_1 B_1\pi_1^t + \sigma^2 I_p)$, we can use the Splus command "rmvnorm" to randomly generate the sample data.

We choose initially $q = 1$, along with $p = 3$, which determine the parameter dimension. In this case, the number of parameters is 5. Next, we choose the sample size $n = 100$ for the case of small sample size and $n = 1000$ for the case of large sample size. We are interested in $a = 0$ because REFM_1 then becomes a standard Factor Analysis Model. Also, we choose $b^2 = 1,\ \sigma^2 = 1$, and $P = (1,\ 0,\ 0)^t$. Now, we simulate two data sets with the same parameters, but different sample sizes.

Using the true parameters as the starting points, we apply EM_1, which is a Splus function to perform the EM algorithm when $q = 1$. Referring to Table 4.1, we can see, in the case of small sample size, the estimates $\hat{a} = -.042,\ \hat{b}^2 = 1.016,\ \hat{\sigma}^2 = .885$ and $\hat{P} = (.991,\ -.136,\ .003)^t$. These numbers are close to the true parameters. Consider the iteration stopping criterion according to

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

In table 4.1, we record the iteration number k at which each component first shows a change smaller than 10^{-3}. Most of these iteration numbers are less than 10. Obviously, most of the estimates and the numbers of iterations are better in the large sample case ($n = 1000$).

We tested different starting points for the sample size of $n = 100$, choosing as a first set of starting parameters: $a = 1,\ b^2 = 2,\ \sigma^2 = 2,\ P = (1,\ 0,\ 0)^t$, and as a second set: $a = 3,\ b^2 = 4,\ \sigma^2 = 5,\ P = (1,\ 1,\ 1)^t$. In both cases, the EM estimates for all parameters converge to the same values displayed in Table 4.1, taking only a

few more iterations to achieve the maximum.

Next, we select another two sets of starting parameters for the sample size $n = 1000$, the first being $(a = 1,\ b^2 = 2,\ \sigma^2 = 2,\ P = (1,\ 0,\ 0)^t)$, and the second $(a = 2,\ b^2 = 3,\ \sigma^2 = 4,\ P = (1,\ 1,\ 1)^t)$. We have all parameters again converging to the same value as Table 4.1, also taking a few more iterations to achieve the maximum. This behavior recalls the conclusion of Lemma 4 from the section 2.3.2, that $\hat{P} = P_0$ is not only the local but also the global maximum of $\tilde{g}_p(\cdot; \theta_{10})$.

For the next data set, we select $p = 10$ and $q = 3$. This is a very large parameter size with a total of 31 (3+3+1+9+8+7=31) free parameters to be estimated. Again, we simulate two data sets with $n = 100$ and $n = 1000$. Here the true parameters θ_0 were chosen to be $\underline{a} = (1,\ 1,\ 1)^t,\ b^2 = (3,\ 2,\ 1)^t,\ \sigma^2 = 1,\ P_1 = (1,\ 0,\ \ldots,\ 0)^t,\ P_2 = (0,\ 1,\ 0,\ \ldots,\ 0)^t$, and $P_3 = (0,\ 0,\ 1,\ 0,\ \ldots,\ 0)^t$. The data are random samples from a multivariate normal distribution. Now, both data sets should be considered as small samples in view of the 31 parameters.

	a=0	b^2=1	σ^2=1	P_1=1	P_2=0	P_3=0
n=100	-.04245	1.01643	.88472	.99074	-0.13573	.00347
#iteration	7	12	6	5	8	7
n=1000	-.00144	1.05643	1.00972	.99823	-.00795	.05887
#iteration	4	11	3	1	3	7

Table 4.1: Results of the EM Algorithm when q=1, where the estimator value is taken at iteration step = 100, and # iteration represents the number of the iteration step which first shows a change smaller than 10^{-3}.

Using the true parameters as starting points, we apply the Splus function for the case $q > 1$. The results are in Tables 4.2 and 4.3. Some of the estimates converge to values rather far from the true parameters. The largest iteration number is 22 in the sample of size $n = 100$, and the largest iteration number is 63 in the dataset of size 1000. This may indicate that the sample size increasing from $n = 100$ to $n = 1000$ is not sufficient to ensure large sample behavior, because of the large parameter dimension. The overall parameter estimation with $n = 1000$ still looks better than with $n = 100$. However, the improvement in results for the two sample sizes between Tables 4.2 and 4.3 is still smaller than the difference shown in Table 4.1. Because of the high parameter dimension, we still consider the Splus function of the EM2 algorithm to be working well.

In order to test the convergence from the different starting points, we set two different starting points for the sample size $n = 100$ and $n = 1000$, respectively. We had slight differences among the resulting parameter estimates, but the worst difference is around 10^{-4}. Considering the large dimension of parameters, small sample size, moderate number (100) of iterations, we do not think these small differences are a serious problem. When we increase the number of iteration to 2000, the differences vanish. This means that all 4 tests are ending at the same convergence points. In the next section, we verify these results from the EM algorithm by using a Quasi-Newton method, and calculate component-wise standard errors.

It takes different numbers of iterations to reach the convergence criterion for each component of parameter θ, as we have seen from Tables 4.2 and 4.3. Sometimes, we need to measure the overall performance of the parameter, not based on the

n=100	a_1=1	a_2=1	a_2=1	b_1^2=3	b_2^2=2	b_3^2=1	σ^2=1
	.84388	.77687	1.06085	4.61825	1.79705	.72465	.97555
#iter	21	14	19	11	6	22	5
	P_{11}=1	P_{12}=0	P_{13}=0	P_{14}=0	P_{15}=0	P_{16}=0	P_{17}=0
	.97696	.01643	-.12962	.01749	-.06218	-.08372	.01509
#iter	8	10	18	8	8	8	8
	P_{18}=0	P_{19}=0	$P_{1,10}$=0	$P_{21}=0$	$P_{22}=1$	$P_{23}=0$	$P_{24}=0$
	-.03481	-.11487	-.05152	.01077	.97129	.05658	.067059
#iter	8	8	8	7	7	13	7
	$P_{25}=0$	$P_{26}=0$	$P_{27}=0$	$P_{28}=0$	$P_{29}=0$	$P_{2,10}=0$	$P_{31}=0$
	.00760	.18088	-.09511	-.41432	-.01577	.07063	.15349
#iter	7	7	7	7	7	7	18
	$P_{32}=0$	$P_{33}=1$	$P_{34}=0$	$P_{35}=0$	$P_{36}=0$	$P_{37}=0$	$P_{38}=0$
	-.07432	.97524	-.01472	.03903	.08838	.00721	-.04173
#iter	13	7	8	6	9	6	6
	$P_{39}=0$	$P_{3,10}=0$					
	.09184	.00666					
#iter	12	6					

Table 4.2: Result of EM Algorithm 2 for sample size 100, where the estimator value is taken at iteration step = 100, and # iteration represents the number of the iteration step which first shows a change smaller than 10^{-3}.

n=1000	a_1=1	a_2=1	a_2=1	b_1^2=3	b_2^2=2	b_3^2=1	σ^2=1
	1.08347	1.04866	.85460	3.08448	1.78886	.98409	.98074
#iter	22	29	63	7	9	4	3
	P_{11}=1	P_{12}=0	P_{13}=0	P_{14}=0	P_{15}=0	P_{16}=0	P_{17}=0
	.99784	.01828	.041186	-.01232	-.01456	.00277	-.03565
#iter	1	2	13	3	3	3	4
	P_{18}=0	P_{19}=0	$P_{1,10}$=0	$P_{21}=0$	$P_{22}=1$	$P_{23}=0$	$P_{24}=0$
	-.00306	.01791	.0175	-.02357	.992023	.1153	-.00380
#iter	2	3	3	2	1	21	4
	$P_{25}=0$	$P_{26}=0$	$P_{27}=0$	$P_{28}=0$	$P_{29}=0$	$P_{2,10}=0$	$P_{31}=0$
	-.3509	-01574	-.01346	.01205	-.00431	-.01392	-.03560
#iter	5	3	4	4	2	3	13
	$P_{32}=0$	$P_{33}=1$	$P_{34}=0$	$P_{35}=0$	$P_{36}=0$	$P_{37}=0$	$P_{38}=0$
	-.11220	.98807	.03432	.06083	-.00472	.05211	-.04708
#iter	52	3	5	6	2	5	5
	$P_{39}=0$	$P_{3,10}=0$					
	-.00601	.00143					
#iter	3	2					

Table 4.3: Result of EM Algorithm 2 for sample size 1000, where the estimator value is taken at iteration step = 100, and # iteration represents the number of the iteration step which first shows a change smaller than 10^{-3}.

individual component. We introduce a value $\mathcal{R}$ to measure the difference between iteration steps, as follows:

$$\mathcal{R}(k) = \sqrt{(\theta_i^{(k+1)} - \theta_i^{(k)})/d}.$$

Later on, a similar measure of distance is also used to measure the difference between two vectors describing tongue curves.

	Case 1	Case 2	Case 3	Case 4
$\mathcal{R}(100)$	8.250191e-15	8.317894e-14	2.471331e-07	1.691174e-05
$\mathcal{R}(2000)$	x	x	2.395897e-16	7.527503e-16

Table 4.4: $\mathcal{R}$-Value in 4 different cases: Case 1: p=3, q=1, n=100. Case 2: p=3, q=1, n=1000. Case 3: p=10, q=3, n=100. Case 4: p=10, q=3, n=1000.

When the number of EM iterations is 100, in all $q = 1$ cases the results are good enough to claim convergence. In all $q = 3$ cases, convergence has not yet been achieved. But, when we we increase number of iterations to 2000, we do achieve convergence.

4.3.4 Quasi-Newton methods on the profile likelihood

If the likelihood function has a unique local maximum, then the maximum likelihood estimators should be the same regardless of the different numerical approaches. Thus, we use Quasi-Newton methods on the profile likelihood to verify the results we got from the EM algorithm on the simulated data.

There is a Splus function, "nlmin", which finds a local minimum of a non-linear function using a general Quasi-Newton method optimizer for an input Splus

function, Based on "nlmin", we wrote another Splus function for case $q = 1$, whose input is a data set, a starting point, 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 $\hat{\theta}_1(\hat{\theta}_2)$. When we apply this function to the two simulated data sets in the case $p = 3$ and $q = 1$, we have exactly the same value $\hat{\theta}_2$ as we found in the EM algorithm , and the profile likelihood from the Quasi-Newton method is equal to the likelihood from the EM algorithm. This result echoes Lemma 2 in Chapter 2. Also, the value $\hat{\theta}_1(\hat{\theta}_2)$ from Quasi-Newton methods is the same as that from the EM algorithm. The total numbers of iteration steps needed to converge is 5 for sample size $n = 100$, and 4 for sample size $n = 1000$ (with respect to overall convergence criterion). This number is less than the 12 and 11 iterations, respectively, needed when applying the EM algorithm. But it takes EM a short time to finish compared to the Quasi-Newton method. This agrees with the same claim made by Watanabe and Yamaguchi (2003).

To see how good the estimation is, we have to check the standard error of $\hat{\theta}_i$, which is given by

$$SE(\hat{\theta}_i) \approx (I^{-1}(\hat{\theta}))_{ii}^{\frac{1}{2}} \quad (i = 1, 2, \ldots, d),$$

where $(A)_{ij}$ means the element in the i'th row and j'th column of a matrix A. Let us discuss the case $q = 1$ first. Since $\|P\| = 1$, that is, $P_1^2 + P_2^2 + P_3^2 = 1$, we can write $P_3 = \sqrt{1 - P_1^2 - P_2^2}$ (just choosing a single sign). Then, the likelihood function $l(\theta)$ is a function $l(a,\ b^2,\ \sigma^2,\ P_1,\ P_2)$ of the parameter $\theta^* = (a,\ b^2,\ \sigma^2,\ P_1,\ P_2)$. We can calculate the 5×5 Hessian matrix $\nabla_{\theta^*}^{\otimes 2} l(\theta^*)$. According to Cox and Hinkley

(1974), the asymptotic covariance matrix of the MLE $\hat{\theta}^*$ is equal to the inverse of the expected information matrix $I(\theta^*)$, which can be approximated by $I(\hat{\theta}^*)$.

Let $I(\hat{\theta}^*) = (-\nabla^{\otimes 2}_{\theta^*} l(\theta^*)|_{\theta^*=\hat{\theta}^*})^{-1}$, where $\hat{\theta}^* = (\hat{a},\ \hat{b}^2,\ \hat{\sigma}^2,\ \hat{P}_1,\ \hat{P}_2)^t$. Thus, the standard error is

$$SE(\hat{\theta}^*_i) \approx \sqrt{I(\hat{\theta}^*)}.$$

	a=0	b^2=1	σ^2=1	P_1=1	P_2=0
n=10000	.00141	.02998	.01007	.00319	.00009
n=5000	.02033	.04183	.01390	.01253	.00020
n=3000	.02605	.0554	.01801	.010667	.00044

Table 4.5: Standard Error

We calculate the Hessian matrix by taking the first order and second order partial derivatives with respect to $a,\ b^2,\ \sigma^2,\ P_1,\ P_2$. We implement a Splus function to compute the standard error including all derivatives we just calculated. Table 4.5 displays the calculated standard errors for different large sample size of the individual parameter components.

For the case $q > 1$, there will be more parameters and more constraints. For example, in the case $q = 3$ and $p = 10$, there are 31 free parameters, and 6 constraints on $\pi_1 = \theta_2$ as an element of the space $\mathcal{M}$ of 10×3 matrices. The idea to obtain the standard error of θ is the same as above. But since $\pi_1 = (P_1,\ P_2,\ P_3)$ has 6 constraints, we remove them by solving 6 constraint equations. Entries $(\pi_1)_{1,10}, (\pi_1)_{2,9}, (\pi_1)_{2,10}, (\pi_1)_{3,8}, (\pi_1)_{3,9}, (\pi_1)_{3,10}$ can then be written as functions of

the other entries, but the expressions are complicated. Moreover, the analytical calculation of the Hessian matrix will be a nightmare. The inverse of a 31×31 Hessian matrix may also cause problems. Thus, the numerical calculation of standard errors in cases as large as $p = 10, q = 3$, will be deferred to future work.

4.4 REFM_2 and EM algorithm

In this section, the Random Effect Factor Model II (REFM_2) is

$$\begin{aligned} y_{ij} &= \sum_{k=1}^{q} c_{ijk} P_k + \epsilon_{ij} \qquad (i = 1, 2, \ldots, s;\ j = 1, 2, \ldots, n), \\ &= \pi_1 c_{ij} + \epsilon_{ij}. \end{aligned} \tag{4.84}$$

where y_{ij} is a p-dimensional random observation vector, the random effect c_{ij} is a q-dimensional unobservable random vector, and $\pi_1 = (P_1,\ P_2,\ ...,\ P_q)$ with $\pi_1^t \pi_1 = I_q$. Let X be the complete data, which includes the random observation vector y_{ij} and the random unobservable vector c_{ij}, that is, $x_{ij} = (y_{ij},\ c_{ij})$. Thus, the complete data X becomes a $(p+q)$-dimensional vector. We have

$$c_{ij} \sim \mathcal{N}_q(\underline{a}_i,\ B_i), \tag{4.85}$$

where $\underline{a}_i = (a_{i1},\ a_{i2},\ \ldots,\ a_{iq})^t$, and $B_i = Diag(b_{i1}^2, b_{i2}^2 \ldots b_{iq}^2)$. The conditional probability distribution over y_{ij}, when the unobservable random effects c_{ij} are given, is

$$y_{ij} | c_{ij} \sim \mathcal{N}_p(\pi_1 c_{ij},\ \sigma_i^2\, I_p). \tag{4.86}$$

Unconditionally, for any fixed group i, $\{y_{ij}\}$ is independently and identically distributed with

$$y_{ij} \sim \mathcal{N}_p(\pi_1 \underline{a}_i,\ \pi_1 B_i \pi_1^t + \sigma_i^2 I_p). \tag{4.87}$$

Then, the complete-data log likelihood function is

$$\begin{aligned} l_X(\theta) &= log(\prod_{i=1}^{s}\prod_{j=1}^{n} f(x_{ij})), \\ &= -\frac{sn(p+q)}{2}\log(2\pi) - \frac{np}{2}\sum_{i=1}^{s}\log\sigma_i^2 - \frac{n}{2}\sum_{i=1}^{s}\sum_{k=1}^{q}\log b_{ik}^2 \\ &\quad -\frac{n}{2}\sum_{i=1}^{s}\frac{1}{\sigma_i^2}\mathrm{tr}(C_{i,yy}) + n\sum_{i=1}^{s}\frac{1}{\sigma_i^2}\,\mathrm{tr}(C_{i,yc}\pi_1^t) - \frac{n}{2}\sum_{i=1}^{s}\frac{1}{\sigma_i^2}\,\mathrm{tr}(C_{i,cc}) \\ &\quad -\frac{n}{2}\sum_{i=1}^{s}\mathrm{tr}(C_{i,cc}B_i^{-1}) + n\sum_{i=1}^{s}\underline{a}_i^t B_i^{-1}\bar{c}_i - \frac{n}{2}\sum_{i=1}^{s}\underline{a}_i^t B_i^{-1}\underline{a}_i \end{aligned} \tag{4.88}$$

where $\bar{c}_i = n^{-1}\sum_{j=1}^{n} c_{ij}$, $C_{i,yy} = n^{-1}\sum_{j=1}^{n} y_{ij}^{\otimes 2}$, $C_{i,yc} = n^{-1}\sum_{j=1}^{n} y_{ij}c_{ij}^t$, and $C_{i,cc} = n^{-1}\sum_{j=1}^{n} c_{ij}^{\otimes 2}$. Thus, $l_X(\theta)$ for $\theta = (\underline{a}_i,\ B_i,\ \sigma_i^2,\ \pi_1)$, belongs to an exponential family, and the sufficient statistics are $\bar{c}_i,\ C_{i,yy},\ C_{i,yc}$, and $C_{i,cc}$.

We can obtain the $q \times p$ transformation matrix D_i by the a procedure similar to that in the previous section 4.2.1 under the condition that $c_{ij} - D_i y_{ij}$ is uncorrelated with y_{ij}. This leads to

$$D_i = B_i\pi_1^t(\pi_1 B_i \pi_1^t + \sigma_i^2 I_p)^{-1}. \tag{4.89}$$

The expectation of the unobservable c_{ij} given the observed data y_{ij} is

$$E(c_{ij}|Y) = \underline{a}_i + D_i(y_{ij} - \pi_1\underline{a}_i), \tag{4.90}$$

and the conditional covariance of unobservable c_{ij} given observed data y_{ij} is

$$Var(c_{ij}|Y) = (I_q - D_i\pi_1)B_i. \tag{4.91}$$

Therefore, the conditional probability distribution over c_{ij}, when the random observable vectors y_{ij} are given, is given by

$$c_{ij}|Y \sim \mathcal{N}_q(\underline{a}_i + D_i(y_{ij} - \pi_1\underline{a}_i), (I_q - D_i\pi_1)B_i) \tag{4.92}$$

where $D_i = B_i\pi_1^t(\pi_1 B_i \pi_1^t + \sigma_i^2 I_p)^{-1}$.

E-Step: We calculate the conditional expectation of the complete data log likelihood function, $l_X(\theta)$, given the observation Y and the k'th temporary values of parameter $\theta = \theta^{(k)}$:

$$\begin{aligned} Q(\theta, \theta^{(k)}) &= E(l_X(\theta)|Y,\ \theta^{(k)}), \\ &= -\frac{sn(p+q)}{2}\log(2\pi) - \frac{np}{2}\sum_{i=1}^{s}\log\sigma_i^2 - \frac{n}{2}\sum_{i=1}^{s}\sum_{k=1}^{q}\log b_{ik}^2 \\ &\quad -\frac{n}{2}\sum_{i=1}^{s}\frac{1}{\sigma_i^2}\mathrm{tr}(C_{i,yy}^*) + n\sum_{i=1}^{s}\frac{1}{\sigma_i^2}\ \mathrm{tr}(C_{i,yc}^*\pi_1^t) - \frac{n}{2}\sum_{i=1}^{s}\frac{1}{\sigma_i^2}\ \mathrm{tr}(C_{i,cc}^*) \\ &\quad -\frac{n}{2}\sum_{i=1}^{s}\mathrm{tr}(C_{i,cc}^* B_i^{-1}) + n\sum_{i=1}^{s}\underline{a}_i^t B_i^{-1}\bar{c}_i^* - \frac{n}{2}\sum_{i=1}^{s}\underline{a}_i^t B_i^{-1}\underline{a}_i \end{aligned} \tag{4.93}$$

where $\bar{c}_i^* = E(\ \bar{c}_i|\ Y,\ \theta^{(k)})$, $C_{i,yy}^* = E(\ C_{i,yy}|\ Y,\ \theta^{(k)})$, $C_{i,yc}^* = E(\ C_{i,yc}|\ Y,\ \theta^{(k)})$, and $C_{i,cc}^* = E(\ C_{i,cc}|\ Y,\ \theta^{(k)})$. Given the current parameter $\theta^{(k)}$ and observed data Y, we have the transformation matrix

$$D_i^{(k)} = B_i^{(k)}(\pi_1^{(k)})^t(\pi_1^{(k)}B_i^{(k)}(\pi_1^{(k)})^t + (\sigma_i^2)^{(k)}I_p)^{-1}, \tag{4.94}$$

and also calculate the conditional expectation of these sufficient statistics as follows:

$$C_{i,yy}^* = C_{i,yy}, \tag{4.95}$$

$$\bar{c}_i^* = \underline{a}_i^{(k)} + D_i^{(k)}(\bar{y}_i - \pi_1^{(k)}\underline{a}_i^{(k)}), \tag{4.96}$$

where $\bar{y}_i = n^{-1}\sum_{j=1}^{n} y_{ij}$,

$$C_{i,yc}^* = \bar{y}_i(\underline{a}_i^{(k)})^t(I_q - (\pi_1^{(k)})^t(D_i^{(k)})^t) + C_{i,yy}(D_i^{(k)})^t, \tag{4.97}$$

and

$$\begin{aligned} C_{i,cc}^* &= (I_q - D_i^{(k)}\pi_1^{(k)})\underline{a}_i^{(k)}(\underline{a}_i^{(k)})^t(I_q - D_i^{(k)}\pi_1^{(k)})^t + D_i^{(k)}\bar{y}_i(\underline{a}_i^{(k)})^t(I_q - D_i^{(k)}\pi_1^{(k)})^t \\ &\quad +(I_q - D_i^{(k)}\pi_1^{(k)})(B_i^{(k)} + \underline{a}_i^{(k)}\bar{y}_i^t(D_i^{(k)})^t) + D_i^{(k)}C_{i,yy}(D_i^{(k)})^t. \end{aligned} \tag{4.98}$$

These equations give the E-step of the $(k+1)$'th iteration of the EM algorithm, where $\theta^{(k)}$ denotes the value of θ after the k'th EM iteration.

M-Step: After we calculate the conditional expectation of the complete data log likelihood function, we perform the M-step, which maximizes $Q(\theta,\ \theta^{(k)})$, that is, maximizes equation (4.93). There are parameters $\underline{a}_i,\ B_i,\ \sigma_i^2$, and π_1 in our Q-function. We find $\underline{a}_i,\ B_i$, and σ_i^2 by taking the partial derivative of $Q(\theta, \theta^{(k)})$ with regard to $\underline{a}_i,\ B_i$, and σ_i^2, respectively. Setting them equal to 0, and simplifying all equations, we have

$$\hat{\underline{a}}_i = \bar{c}_i^*. \tag{4.99}$$

and

$$\hat{B}_i = Diag(C^*_{i,cc}) - Diag(\hat{\underline{a}}_i \hat{\underline{a}}_i{}^t) \tag{4.100}$$

Since M-step estimation of σ_i^2 involves π_1, we first find $\hat{\sigma}_i^2$ as a function of $\hat{\pi}_1$. Thus, we must find an estimate of π_1 in terms of $(\sigma_1^2,\ \sigma_2^2,\ \ldots,\ \sigma_s^2)$ by maximizing

$$\max_{\pi_1} \sum_{i=1}^{s} \frac{1}{\sigma_i^2} tr(C^*_{i,yc}\pi_1^t) = \max_{\pi_1} \mathrm{tr}(\sum_{i=1}^{s} \frac{1}{\sigma_i^2}\pi_1^t). \tag{4.101}$$

From Lemma 13, we obtain $\hat{\pi}_1$ as a function of $(\sigma_1^2, \ldots, \sigma_s^2)$. After substituting $\hat{a}_i$, $\hat{B}_i$, and $\hat{\pi}_1$ into the Q-function, this 'profile' Q-function only depends on $(\sigma_1^2, \ldots, \sigma_s^2)$. We will use a Newton-Raphson method to obtain the numerical value of all σ_i^2 in each M-step iteration.

Chapter 5

2-D Coronal Tongue Data

5.1 Data Set

The tongue is the major contributor to vocal tract shape and the resulting speech signal, and is also an unusual structure in the human body because the internal musculature of the tongue provides support as both a skeletal bone and muscle, while the organ itself maintains a constant volume. A change in one dimension will result in change in at least one other dimension. Theoretically speaking, the tongue has an infinity of degrees of freedom. The statistical model, although not directly representative of the underlying muscles, will be used initially to reduce the complexity of vocal tract and tongue surface behavior through some simple features, such as openness and shape. The shapes captured by the statistical model should be explicable by what the underlying muscles can produce.

The methodology of data collection includes a Head And Transducer Support (HATS) [35] system designed to hold the subject's head and the transducer steady in a fixed position. The coronal tongue images were extracted from digitized ultrasound images recorded on a VCR, using the μ-Tongue software package [37] during natural speech. The cross-sectional tongue surface was recorded and measured for six subjects (3 Caucasian females, 2 African-American males, 1 Hispanic male), three different sessions and twenty-two different sounds (two consonants and eleven vowels) by ultrasound, VCR and the μ-Tongue software package in the Vocal Tract Visual-

ization Laboratory of M. Stone in Baltimore. Each subject speaks each sound five times successively at each session. Thus we obtained a total of $22 \times 5 \times 3 \times 6 = 1980$ cross-sectional tongue images. Each curve image is represented by 120 pairs (x, y), and different curves do not necessarily have the same x values. Let (x_{abcdi}, y_{abcdi}), for $a = 1, 2, \ldots, 6,\ b = 1, 2, 3,\ c = 1, 2, \ldots, 22,\ d = 1, \ldots, 5,\ i = 1, \ldots 120,$ be our raw data set, where a indexes subject, b indexes session, c indexes sound, d indexes replications within session, and i indexes lateral location on the image curves.

The ultrasound measuring system is set differently for different subjects and sessions, which results in arbitrary shifts in the x and y coordinates. Furthermore, the coronal tongue width varies across subjects and sounds, even for the same speech sound, session, and speaker. Pre-processing strategies were introduced by Slud et al. (2002), involving translation in the x and y directions, extension (padding) or truncation within session, and subtracting constants by sound, session, and speaker. Hence, the final data set on a common (x, y) coordinate system based upon five replicated measurements in three sessions for each of six subjects, is

$$(x_i,\ y_{abcd,i}), \tag{5.1}$$

where subject is indexed by $a = 1,\ 2,\ \ldots,\ 6$, session by $b = 1,\ 2,\ 3$, sound by $c = 1,\ 2,\ \ldots,\ 22$, replication by $d = 1,\ 2,\ 3,\ 4,\ 5$, and observations (points) along the image curve by $i = 1,\ 2,\ \ldots,\ 101$. We have a total of $6 \times 3 \times 22 \times 5 = 1980$ cross-sectional tongue images available to analyze.

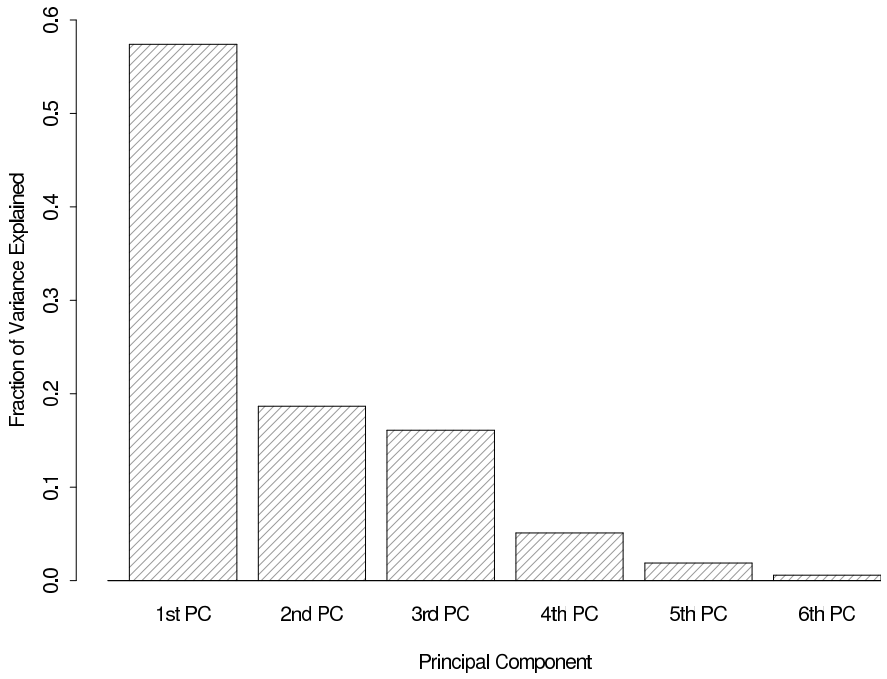

Figure 5.1: Percent of variance in the i'th PC from the Tongue data.

5.2 Data Analysis Using EM Algorithm

First, we ran a Principal Component Analysis on the final Tongue data Y, a 1980×101 matrix, obtained by pre-processing Plan 5 of Slud et al. (2002) combining all pre-processing steps. The percentage of variance (in order) accounted for by the successive PC's are: 57.4%, 18.6%, 16.1%, 5.1%, 1.9%, .6%, and .2% (see Figure 5.1). The first six components capture over 99% of variation. Therefore we decide to reduce the data by projecting onto only the first 6 PCs. The principal space is $V = \text{span}\{PC1,\ PC2,\ \ldots,\ PC6\}$. We use PC1 and PC2 as a basis with which to compare the estimated directions which we will later find in the factor space by the EM algorithm.

We directly apply the EM algorithm for REFM_1 with $q = 1,\ a = 1,\ b^2 = 1,\ \sigma^2 = 1, P = (1, 0, \ldots, 0)$ on the final Tongue data. We find the estimated

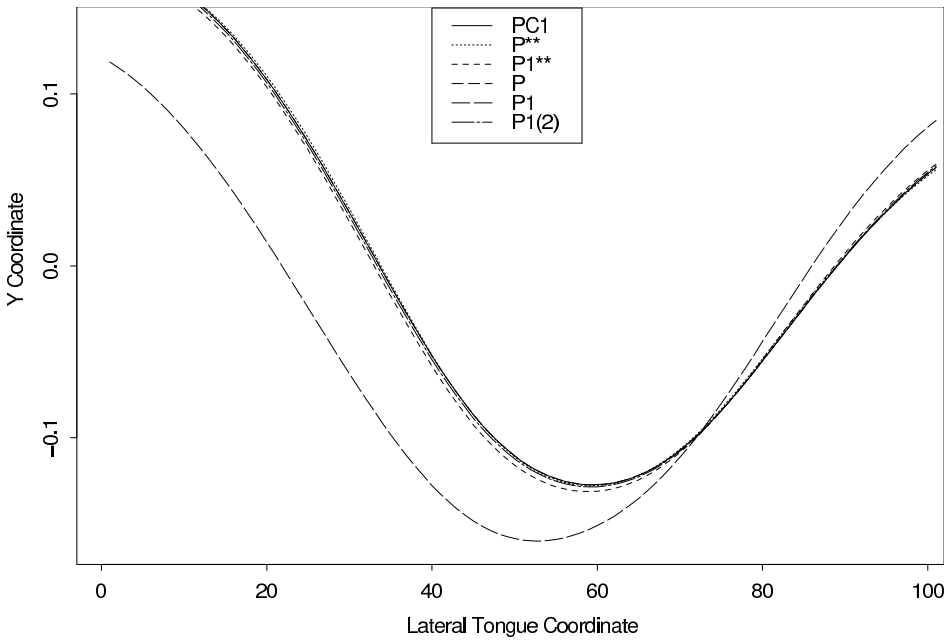

Figure 5.2: Tongue Data: PC1 vs. the first basis factor directions from the EM algorithm with different approaches (six curves total).

direction P at the 100th iteration step in factor space is almost the same as the PC1 of Slud et al. (2002). The accuracy $\mathcal{R}(100)$ is 4.46e-14, and the difference between PC1 and P is 1.82e-4 ($\mathcal{R}$ value). Next, we directly apply the EM algorithm for REFM_1 using $q = 2$ on the same data with starting points: $a = (1, 1)$, $b^2 = (3, 1)$, $\sigma^2 = 1$, $P_1 = (1, 0, \ldots, 0)$, and $P_2 = (0, 1, 0, \ldots, 0)$. First, we iterate 100 times. The accuracy for the EM algorithm is $\mathcal{R}(100) = 2.36e - 4$. We estimate two basis directions P_1 and P_2 which are completely different from the first two Principal Components. When we increase the number of iteration from 100 to 1000, P_1 and P_2 move toward PC1 and PC2, respectively. But, the change in parameter estimates from one iteration to the next does not increase since $\mathcal{R}(1000) = 2.40e - 4$. We will further investigate P_1 and P_2 convergence of estimate in the future.

Since the dimension of the Tongue data is high, $p = 101$, and also the first six

PCs dominate the total variance, we project the Tongue data on the principal space V by $Y^* = Y\pi_1$ where $\pi_1 = (PC1,\ PC2,\ \ldots,\ PC6)$ is a 101×6 matrix. That is, we project the 1980×101 matrix orthogonally to a 1980×6 matrix. Now, the projected data consist of 1980 observations on 6 dimensions. It is much easier to estimate all parameters in the case $p = 6$ by the EM algorithm. We implement two Splus functions on the projected data set by choosing $q = 1$ and $q = 2$, respectively. When we iterate 100 times, the single iteration index $\mathcal{R}$ of change is 1.16e-14 and 2.73e-5, respectively. We obtain P^* as the single ($q = 1$) 6-dimensional estimated basis vector, P_1^* as the first basis direction in case $q = 2$, and P_2^* as the second basis direction in case $q = 2$. Now, we transform back P^*, P_1^*, and P_2^* to $\mathcal{R}^{101}$ by $P^{**} = \pi_1 P^*$, $P_1^{**} = \pi_1 P_1^*$, and $P_2^{**} = \pi_1 P_2^*$. The operator projecting to the residuals from the mapping and transform is $I_{101} - \pi_1\pi_1^t$. Then P^{**}, P_1^{**}, and P_2^{**} are 101 dimensional vectors. The difference (root mean-square component-wise difference) between P^{**} and PC1 is 4.08e-3; the difference between P_1^{**} and PC1 is 2.76e-2; and the difference between P_2^{**} and PC2 is 1.57e-4.

Using P_1^{**} and P_2^{**} as starting points to rerun REFM_1 with q=2, we find $\mathcal{R}(100) = 3.54e-5$, and the estimate vectors $P_1^{(2)}$ and $P_2^{(2)}$ are close to PC1 and PC2 with root mean-square component-wise difference 1.01e-2 and 5.50e-4.

We see from the simulated data and the Tongue data that REFM_1 with $q = 1$ converges quickly to a global maximizer of log-likelihood. Then we can apply the same model again on data projected to the orthocomplement of the basis direction already found. Let $Y^{***} = Y(I_{101} - PP^t)$. Then REFM_1 with $q = 1$ applied to Y^{***}, yields the estimated ($q = 1$) basis direction P^{***}. We compare P^{***} to PC2. The

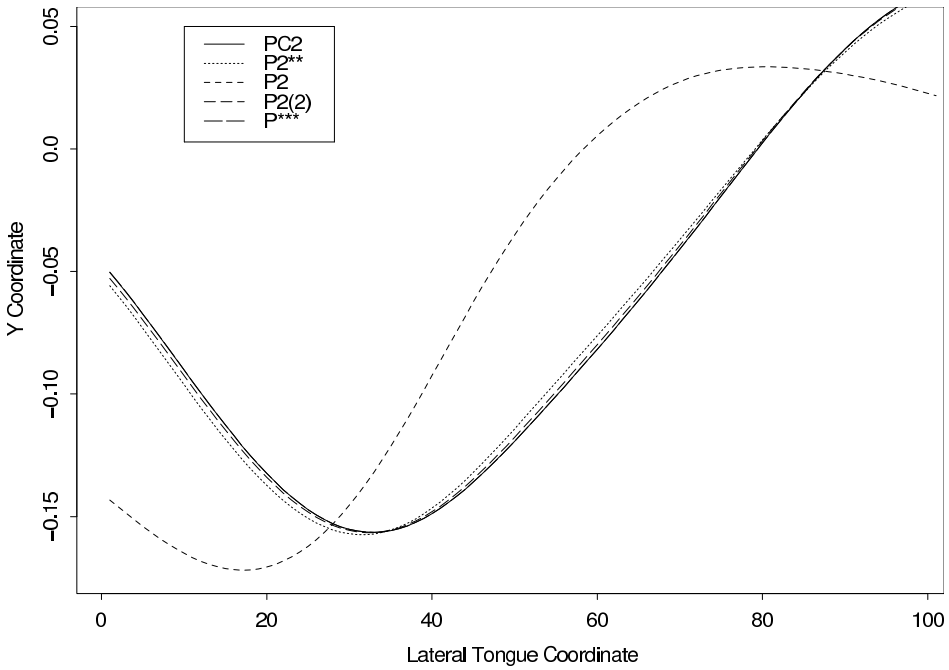

Figure 5.3: Tongue Data: PC2 vs. the second basis factor directions from the EM algorithm with different approaches (five curves total).

difference between them is 5.16e-5. To summarize, we make two plots, respectively of PC1, P^{**}, P_1^{**}, P, P_1, and $P_1^{(2)}$ (Figure 5.2), and of PC2, P_2^{**}, P_2, $P_2^{(2)}$, and P^{***} (Figure 5.3). All basis directions from the EM algorithm for REFM_1 coincide with PC1 and PC2 to high accuracy except P_1 and P_2.

Our real data example shows that $\mathbf{REFM}_1$ can estimate the principal space about as well as PCA does. Besides that, $\mathbf{REFM}_1$ has other nice properties on other parameters. Moreover, $\mathbf{REFM}_1$ could be used to simulate artificial data, but PCA could not.

Remark: As an alternative computational method, we could first project P_1 and P_2 onto the Principal space V by $P_1\ \pi_1$ and $P_2\ \pi_1$. Using them as starting points to run the EM algorithm in REFM_1 with $q = 2$, the two estimate basis directions at the 100th iteration are almost the same as P_1^* and P_2^*. □

Chapter 6

Summary

We introduced a new model, Random Effect Factor Model I (REFM_1), and found a sufficient condition to identify all parameters. We characterize the maximum likelihood estimators (MLE's) under REFM_1 by a profile likelihood method. That is, we maximize the likelihood first with respect to $\theta_1 = (\underline{a},\ B,\ \sigma^2)$, with the other parameter component θ_2 fixed, find closed-form restricted MLE's $\hat{\underline{a}},\ \hat{B},\ \hat{\sigma}^2$ in terms of the factor directions. We then substitute $\hat{\underline{a}},\ \hat{B},\ \hat{\sigma}^2$ into the likelihood, and finally maximize the profile likelihood with respect to the factor directions. We prove that there exists a unique local maximum of the profile likelihood . In the special case when $q = 1$, the maximum is the global maximum. Also, we show that the restricted MLE and MLE from the profile likelihood are consistent. We show that the Hessian matrix for θ_1 is negative definite and also prove without calculating the derivatives that the Hessian matrix for θ_2 is negative definite. From that, we conclude the positive definiteness of the Fisher Information matrix in terms of free parameters. This ensures that the asymptotic properties of the MLE such as asymptotic normal distribution hold. Finally, we show that the maximizer of the profile likelihood function $l_p(\theta_2)$ over the factor directions, combined with the restricted MLE for other parameters, is the joint MLE of the likelihood function.

We extend our new model to multivariate data from s groups ($s > 1$). We further introduce two more new models. REFM_2 is a model which assumes all s

groups have a common factor space but differing mean and variance parameters for factor loadings and error terms, and REFM_3 is a model which has not only a common factor space but also an additional individual space belonging to each group only. We find sufficient conditions to identify all parameters, and give the closed-form expressions for the restricted MLE's $\hat{\theta}_1$.

We find the EM algorithm formula for REFM_1 to compute MLE and also a slightly less explicit EM algorithm formula in REFM_2. The performance of the algorithm on simulated data for REFM_1 is described. Quasi-Newton methods are also used to calculate the MLE of the profile likelihood $l_p(\theta_2)$ and yield the same results as the EM algorithm. Finally, we apply the EM algorithm for REFM_1 estimation to a real data set on ultrasound cross-sectional images of the tongue during speech.

In the next phase of work, we will focus on the following three areas: computational, theoretical and applications. In the computational area, we will explore methods to calculate standard errors for all models, implement Quasi-Newton methods using Splus to find MLE for the case $q > 1$, and extend the EM algorithm to our models REFM_2 and REFM_3. In further theoretical work, we will establish consistency and MLE asymptotic normality in REFM_2 and REFM_3. We have applied our model to a real tongue dataset. The question we often ask is "Where else can we apply our models?".

BIBLIOGRAPHY

[1] Andersen, P.K. and Gill, R.D. (1982). Cox's Regression Model for Counting Processes: A Large Sample Study. The Annals of Statistics, 10, 1100-1120.

[2] Anderson, T.W. (1963). Asymptotic Theory for Principal Component Analysis. Ann. of Math. Stat., Vol. 34, 122-148.

[3] Anderson, T.W. (1984). An Introduction to Multivariate Statistical Analysis. New York, Wiley.

[4] Bartholomew, D.J. (1987). Latent Variable Models and Factor Analysis. Oxford, Oxford University Press.

[5] Basilevsky, A. (1994). Statistical Factor Analysis and Related Methods: Theory and Applications. New York, Wiley.

[6] Cox, D.R. and Hinkley, D.V. (1974). Theoretical Statistics. London, Chapman & Hall.

[7] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm with Discussion. Journal of the Royal Statistical Society B, 39, 1-38.

[8] Flury, B.D. (1984). Common Principal Components in k Groups. J. Amer. Statist. Assoc., 79, 892-898.

[9] Flury, B.D. (1986). Asymptotic Theory for Common Principal Component Analysis. The Annals of Statistics, 14, 418-430.

[10] Flury, B. (1988). Common Principal Components and Related Multivariate Models. New York, Wiley.

[11] Flury, B. and Gautschi, W. (1986). An Algorithm for Simultaneous Orthogonal Transformation of Several Positive Definite Symmetric Matrices to Nearly Diagonal Form. SIAM Journal on Scientific and Statistical Computing, 7, 169-184.

[12] Harman, H.H. (1976). Modern Factor Analysis. Chicago, The University of Chicago Press.

[13] Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. Journal of Educational Psychology, 24, 417-441.

[14] Jolliffe, I.T. (1986). Principal Component Analysis. New York, Springer-Verlag.

[15] Karlin, S. and Taylor, H.M. (1975). A First Course in Stochastic Processes. San Diego, Academic Press Inc.

[16] Lange, K. (1995). A Gradient Algorithm Locally Equivalent to the EM Algorithm. Journal of the Royal Statistical Society B, 57, 425-437.

[17] Lehmann, E.L. (1991). Theory of Point Estimation. Belmont, California, Wadsworth, Inc.

[18] Liu, C. and Rubin, D.B. (1994). The ECME Algorithm: a Simple Extension of EM and ECM with Faster Monotone Convergence. Biometrika, 81, 633-648.

[19] McHugh, R.B. (1956). Efficient Estimation and Local Identification in Latent Class Analysis. Psychometrika, 21, 331-347.

[20] McHugh, R.B. (1958). Note on Efficient Estimation and Local Identification in Latent Class Analysis. Psychometrika, 23, 2743-274.

[21] McLachlan, G.J. and Krishnan, T. (1997). The EM Algorithm and Extensions. New York, Wiley.

[22] Meng, X.L. (1997). The EM Algorithm and Medical Studies: a Historical Link. Stat. Meth. Med. Res., 6, 3-23.

[23] Meng, X.L. and Dyk, D. (1997). The EM Algorithm - an Old Folk Song Sung to a New Tune. Journal of the Royal Statistical Society B, 59, 511-567.

[24] Meng, X.L., and Pedlow, S. (1992). EM: a Bibliographic Review with Missing Article. Proceedings of the Statistical Computing Section, 24-27, American Statistical Association, Washington, DC.

[25] Meng, X.L. and Rubin, D.B. (1993). Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. Biometrika, 80, 267-278.

[26] Muirhead, R.J. (1982). Aspects of Multivariate Statistical Theory. New York, Wiley.

[27] Neyman, J. and Scott E.L. (1948). Consistent Estimates Based on Partially Consistent Observations. Econometrica, 16, 1-32.

[28] Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine, Ser. 6, 2, 559-572.

[29] Roweis, S. (1997). EM Algorithms for PCA and SPCA. Neural Information Processing Systems, 10, 626-632.

[30] Rubin, D.B. (1991). EM and Beyond. Psychometrika, 56, 241-254.

[31] Rubin, D.B. and Thayer, D.T. (1982). EM Algorithms for ML Factor Analysis. Psychometrika, 47, 69-76.

[32] Schölkopf, B., Smola, A. and Müller, K.R. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Computation, 10(5), 1299-1319.

[33] Silverman, B.W. (1996). Smoothed Functional Principal Components Analysis by Choice of Norm. The Annals of Statistics, 24, 1-24.

[34] Slud, E.V., Stone, M., Smith, P.J., and Goldstein, M. (2002). Principal Components Representation of the Two-Dimensional Coronal Tongue Surface. Phonetica, 59, 108-133.

[35] Stone, M. and Davis, E.P. (1995). A Head and Transducer Support System for Making Ultrasound Images of Tongue/Jaw Movement. J. Acoust. Soc. Am., 98, 3107-3112.

[36] Tipping, M.E. and Bishop, C.M. (1999). Probabilistic Principal Component Analysis. Journal of the Royal Statistical Society B, 61, 611-622.

[37] Unser, M. and Stone, M. (1991). Automated Detection of the Tongue Surface in Sequences of Ultrasound Images. J. Acoust. Soc. Am., 94, 3001-3007.

[38] Wahba, G. (1990). Spline Models for Observational Data. Society for Industrial and Applied Mathematics.

[39] Watanabe, M. and Yamaguchi, K. (2003). The EM Algorithm and Related Statistical Models. New York, Marcel Dekker, Inc.

[40] Wu, C.F.J. (1983). On the Convergence Properties of the EM Algorithm. The Annals of Statistics, 11, 95-103.

[41] Young, G. (1940). Maximum Likelihood Estimation and Factor Analysis. Psychometrika, 6, 49-53.