A thesis submitted for the degree of *Master of Mathematical Science*

27th June 2024

Model Selection in Factor Analysis

Zhining Wang

Mathematical Sciences Institute

supervised by

Dr. Emi Tanaka, Dr. Qinian Jin

Australian National University

Master of Mathematical Science Thesis

Author:	Zhining Wang
Supervisors:	Dr. Emi Tanaka, Dr. Qinian Jin
Project period:	2024-03-01 – 2024-10-01

Mathematical Sciences Institute
Australian National University

Table of contents

List of Figures

List of Tables

Preface

Chapter 1

Introduction

Notation is presented in Table ??.

1.1 What is factor analysis and why is it important?

Factor analysis is a mathematical model which tries to use fewer underlying factors to explain the correlation between a large set of observed variables (?). It provides a useful tool for exploring the covariance structure among observable variables (?). One of the major assumptions that factor analytic model stands on is that it is impossible for us to observe those underlying factors directly. This assumption is especially suited to subjects like psychology where we cannot observe exactly some concept like how intelligent our subjects are (?).

Suppose we have a observable random vector $y \in \mathbb{R}^p$ with mean $\mathbb{E}[y] = \mu$ and variance $\mathbb{V}[y] = \Sigma$. Then a k-order factor analysis model for y can be given by

$$y = \Lambda f + \mu + \epsilon, \tag{1.1}$$

where $\Lambda \in \mathbb{R}^{p\times k}$ is called *loading matrix*, we call $f \in \mathbb{R}^k$ as *common factors* and $\epsilon \in \mathbb{R}^p$ is *unique factors*. To make the model well-defined, we may assume

$$\mathbb{E}[f] = 0_k, \mathbb{V}[f] = I_{k\times k}, \mathbb{E}[\epsilon] = 0_p, \mathbb{V}[\epsilon] =: \Psi = \text{diag}(\Psi_{11}, \dots, \Psi_{pp})$$

and also the independence between any elements from f and ϵ separately, i.e.

$$Cov[f_i, \epsilon_j] = 0, \text{ for all } i \in \{1, 2, \dots, k\} \text{ and } j \in \{1, 2, \dots, p\}$$

Straightforwardly, the covariance of observable vector y can be modelled by

$$\mathbb{V}[y] = \Lambda\Lambda^\top + \Psi \tag{1.2}$$

1.2 Indeterminacy of the loading matrix

One can easily see that if our factor analytic model is given by (1), then it can also be modelled as

$$y = (\Lambda M)(M^\top f) + \mu + \epsilon$$

where the matrix M is orthogonal and simultaneously the variance of y given by (2) still holds, since

$$\mathbb{V}[y] = (\Lambda M M^\top)\mathbb{V}[f](\Lambda M M^\top)^\top + \Psi = \Lambda\Lambda^\top + \Psi.$$

Therefore a rotated loading matrix ΛM is still a valid loading matrix for a factor analytic model. Sometimes we resolve this problem by making the loading matrix to satisfy some constraints like (?)

$$\Lambda^\top \Psi^{-1} \Lambda \text{ is diagonal.}$$

1.3 Traditional Estimation of Parameters in Factor Analytic Models

We denote the set of parameters by $\beta := \{\text{vec}(\Lambda), \text{vec}(\Psi)\}$ where $\text{vec}(\cdot)$ is the vectorisation of the input.

Traditionally, a two-step procedure is used to construct a factor analytic model: estimate parameters by maximum likelihood estimation (aka, MLE) and then use rotation techniques to find an interpretable model.

1.3.1 Maximum Likelihood Estimation

Suppose we have n independent and identically distributed observations $y_1, y_2, \dots, y_N$ from a p-dimensional multi-variate normal distribution $N_p(\mu, \Sigma)$ and by our hypothesis, we have $\Sigma = \Lambda\Lambda^\top + \Psi$. Then the likelihood function is given by

$$L(\Lambda, \Psi) = \prod_{i=1}^{n} \left[(2\pi)^{-\frac{p}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp(-\frac{1}{2}(y_i - \mu)^\top \Sigma^{-1}(y_i - \mu)) \right].$$