Tutorial on Structural Equation Modeling

John Sum
Institute of Technology Management, National Chung Hsing University
Taichung 40227, Taiwan
Email: pfsum@nchu.edu.tw

December 25, 2019

Contents

List of Figures

List of Tables

1 Introduction

Structural equation model (SEM) is a statistical model to analyze the data collected from a batch of questionnaires. The data collected is assumed to be linear related to some latent variables which are unobservable (or unmeasurable). Schematically, a SEM consists of a pair of input/output $(\boldsymbol{x}, \boldsymbol{y})$ which are observable, Figure 1. The output vector $\boldsymbol{y}$ is linear related to a vector of latent factors $\boldsymbol{\eta}$. The input vector $\boldsymbol{x}$ is linear related to a vector of latent factors $\boldsymbol{\xi}$. The latent vector $\boldsymbol{\eta}$ is linear related to $\boldsymbol{\xi}$. From another perspective, the observable facts $(\boldsymbol{x}, \boldsymbol{y})$ are generated (equivalently governed) by those latent factors $\boldsymbol{\xi}$, $\boldsymbol{\eta}$ and the noise factors $\boldsymbol{\zeta}$, $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$.

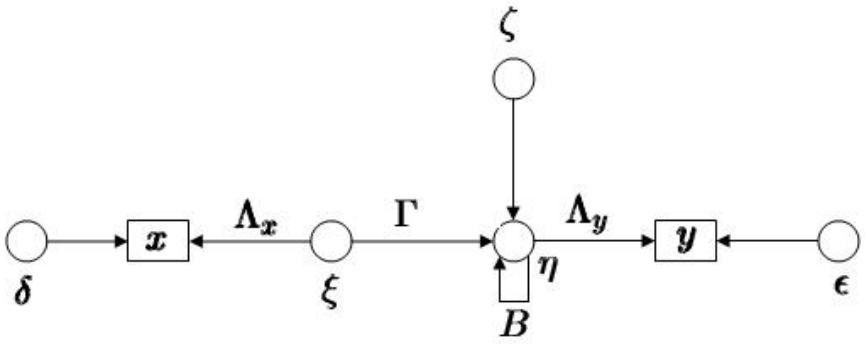

Figure 1: Schematic diagram of a structural equation model.

A good exposition on SEM can be referred to [5]. Although this book focuses on Bayesian SEM, it has two introductory chapters (Chapter 2 and 3) on SEM, its relations to other models such as CFA Model and Bentler-Weeks Model [2] and the theorems for understanding the properties of different estimators.

Generally speaking, the methodology for conducting a management research could follow the steps as shown in Figure 2.

Conceptual model design: Based upon literature survey, the relationships (equivalently the implications) amongst factors should be put together as the hypotheses of a conceptual model. Some of these factors could be observable (equivalently measurable) and some might be invisible (equivalently non-measurable). Almost in all management researches, the factors appeared in a conceptual model are assumed to be invisible. These factors indeed are the $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ in Figure 1.

Questionnaire design: Questions are designed to reveal the quantities of the factors, including both measurable and non-measurable. Special care has to be taken for those non-measurable. Multiple questions should be designed for a single factor.

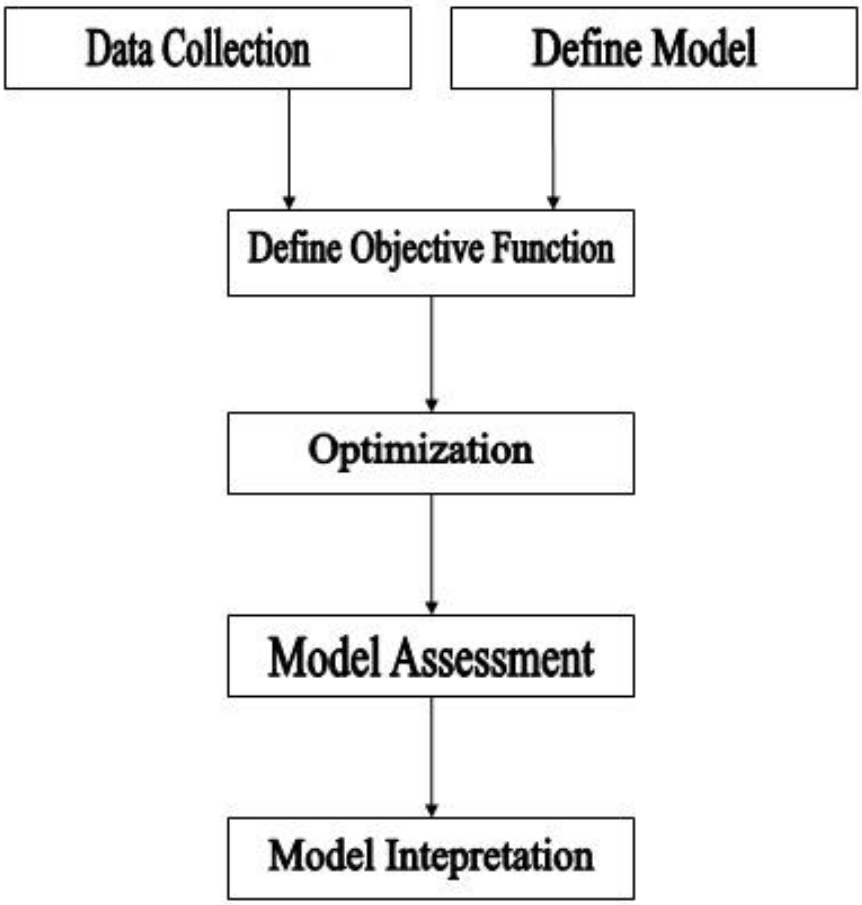

Figure 2: Steps for conducting survey research.

These questions are the variables $\boldsymbol{x}$ and $\boldsymbol{y}$ in Figure 1. Therefore, the factors in the conceptual model and the questions designed in the questionnaire are the variables in the structural equation model.

Data collection: Once a questionnaire has been carefully designed, the questionnaires would be distributed for *data collection*.

Statistical analysis: After sufficient samples has been collected, an *objective function* for estimation have to be defined. Precisely, estimation refers to finding the parameters in the matrices $\boldsymbol{B}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Lambda_y}$ and $\boldsymbol{\Lambda_x}$; the covariance matrices for $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, $\boldsymbol{\zeta}$, $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$, as stated in (1), (2) and (3). The objective function could be defined based on comparison between the elements of the sample covariance matrix and the estimated covariance matrix. It could also be defined based on likelihood function. Then, the analyst has to select amongst different *optimization techniques* the one that can search an optimal solution efficiently.

Model assessment: An analyst can also conduct multiple-search for more than one solution and assess the models' viability with reference to *model assessment* indices.

Implications: While the best model has been selected, further analysis and interpretation on the model can be made.

In the rest of the paper, except on the design of questionnaire and data collection method, various steps in using SEM as a tool for survey research will be summarized.

1.1 SE Model Definition

In accordance with the terminologies in SAS, a structural equation model (SEM) can be defined as follows :

$$\boldsymbol{\eta} = \boldsymbol{B\eta} + \boldsymbol{\Gamma\xi} + \boldsymbol{\zeta} \quad (1)$$

$$\boldsymbol{y} = \boldsymbol{\Lambda_y\eta} + \boldsymbol{\epsilon} \quad (2)$$

$$\boldsymbol{x} = \boldsymbol{\Lambda_x\xi} + \boldsymbol{\delta}, \quad (3)$$

where $\boldsymbol{\eta} \in R^m$, $\boldsymbol{\xi} \in R^n$, $\boldsymbol{y} \in R^p$ and $\boldsymbol{x} \in R^q$. In which, the elements in $\boldsymbol{\eta}$ correspond to endogenous latent variables. Elements in $\boldsymbol{\xi}$ correspond to exogenous latent variables. Elements in vectors $\boldsymbol{y}$ and $\boldsymbol{x}$ correspond to the manifest variables that are observable (or measurable). Vectors $\boldsymbol{\zeta} \in R^m$, $\boldsymbol{\epsilon} \in R^p$ and $\boldsymbol{\delta} \in R^q$ are the error vectors of mean zero.

$\boldsymbol{B}$ is an $m \times m$ matrix with diagonal elements all zeros. Its off-diagonal elements specify the interaction amongst the endogenous latent variables. $\boldsymbol{\Gamma}$ is an $m \times n$ matrix specifying the dependence of the endogenous variables on exogenous variables. Without

loss of generality, the expectations of random vectors $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, $\boldsymbol{y}$ and $\boldsymbol{x}$ are all zeros. That is,

$$E[\boldsymbol{\eta}] = 0, \quad E[\boldsymbol{\xi}] = 0, \quad E[\boldsymbol{y}] = 0, \quad E[\boldsymbol{x}] = 0.$$

While the covariance matrices for the random vectors $\boldsymbol{\xi}$, $\boldsymbol{\epsilon}$, $\boldsymbol{\delta}$ and $\boldsymbol{\zeta}$ are depicted as follows :

$$E[\boldsymbol{\xi}\boldsymbol{\xi}^T] = \Phi, \quad E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \boldsymbol{\Theta}_{\boldsymbol{\epsilon}}, \quad E[\boldsymbol{\delta}\boldsymbol{\delta}^T] = \boldsymbol{\Theta}_{\boldsymbol{\delta}}, \quad E[\boldsymbol{\zeta}\boldsymbol{\zeta}^T] = \boldsymbol{\Psi}.$$

By convention, it is assumed that $\boldsymbol{\Theta}_{\boldsymbol{\epsilon}}$, $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$ and $\boldsymbol{\Psi}$ are diagonal matrices, meaning that the random variables in $\boldsymbol{\epsilon}$, $\boldsymbol{\delta}$ and $\boldsymbol{\zeta}$ are all independent. Furthermore, the random vectors $\boldsymbol{\zeta}$, $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ are independent of $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, $\boldsymbol{y}$ and $\boldsymbol{x}$.

$$E[\boldsymbol{\eta}\boldsymbol{\zeta}^T] = 0, \quad E[\boldsymbol{\xi}\boldsymbol{\zeta}^T] = 0, \quad E[\boldsymbol{y}\boldsymbol{\epsilon}^T] = 0, \quad E[\boldsymbol{x}\boldsymbol{\delta}^T] = 0.$$

Given $\boldsymbol{B}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Lambda}_{\boldsymbol{y}}$, $\boldsymbol{\Lambda}_{\boldsymbol{x}}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Theta}_{\boldsymbol{\epsilon}}$, $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$ and $\boldsymbol{\Psi}$, the covariance matrices for $\boldsymbol{\eta}$, y and x can readily be deduced.

$$\begin{aligned}
E[\boldsymbol{\eta}\boldsymbol{\eta}^T] &= (\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^T(\boldsymbol{I}-\boldsymbol{B})^{-T} + (\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{\Psi}(\boldsymbol{I}-\boldsymbol{B})^{-T}, && (4)\\
E[\boldsymbol{y}\boldsymbol{y}^T] &= \boldsymbol{\Lambda}_{\boldsymbol{y}}E[\boldsymbol{\eta}\boldsymbol{\eta}^T]\boldsymbol{\Lambda}_{\boldsymbol{y}}^T + \boldsymbol{\Theta}_{\boldsymbol{\epsilon}}, && (5)\\
E[\boldsymbol{x}\boldsymbol{x}^T] &= \boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{\Phi}\boldsymbol{\Lambda}_{\boldsymbol{x}}^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}}. && (6)
\end{aligned}$$

For simplicity, we let $\boldsymbol{\theta}$ be the parametric vectors augmenting all the parameters in $\boldsymbol{B}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Lambda}_{\boldsymbol{y}}$, $\boldsymbol{\Lambda}_{\boldsymbol{x}}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Theta}_{\boldsymbol{\epsilon}}$, $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$ and $\boldsymbol{\Psi}$. Besides, the covariance matrix for random variables $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, $\boldsymbol{y}$ and $\boldsymbol{x}$ are denoted by $\boldsymbol{\Sigma}_{\boldsymbol{\eta\eta}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\eta\xi}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\xi\eta}}$, $\boldsymbol{\Sigma}_{\boldsymbol{yy}}$, $\boldsymbol{\Sigma}_{\boldsymbol{xx}}$, $\boldsymbol{\Sigma}_{\boldsymbol{yx}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{xy}}$.

1.2 Standardized form

Note that the model parameters specified by Equations (1) (2) and (3) has any restriction. Except that there are a few mild conditions on the latent vectors and the error vectors, such as mean zeros on $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$. The covariance between latent vectors and error vectors are independent.

In such case, a model can have infinite many equivalent representation which leads to a confusion in comparison amongst different models. Let us have a simple example. Suppose we define

$$\boldsymbol{\eta}' = \boldsymbol{\eta}/2, \quad \boldsymbol{\xi}' = \boldsymbol{\xi}/2, \quad \boldsymbol{\zeta}' = \boldsymbol{\zeta}/2, \quad \boldsymbol{\Lambda}'_{\boldsymbol{y}} = 2\boldsymbol{\Lambda}_{\boldsymbol{y}}, \quad \boldsymbol{\Lambda}'_{\boldsymbol{x}} = 2\boldsymbol{\Lambda}_{\boldsymbol{x}},$$

the aforementioned model will be equivalent to the following model.

$$\begin{aligned}
\boldsymbol{\eta}' &= \boldsymbol{B}\boldsymbol{\eta}' + \boldsymbol{\Gamma}\boldsymbol{\xi}' + \boldsymbol{\zeta}'\\
\boldsymbol{y} &= \boldsymbol{\Lambda}'_{\boldsymbol{y}}\boldsymbol{\eta}' + \boldsymbol{\epsilon}\\
\boldsymbol{x} &= \boldsymbol{\Lambda}'_{\boldsymbol{x}}\boldsymbol{\xi}' + \boldsymbol{\delta}.
\end{aligned}$$

In such case, the impact of the endogenous variables on $\boldsymbol{y}$ will be confused.

One approach to circumvent such confusion is by standardization. In which, the variances of the elements in the random vectors $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, $\boldsymbol{\zeta}$, $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ are set to unity. This standardization can be accomplished by the following steps.

First, evaluate the covariance matrix for $\boldsymbol{\eta}$, $\boldsymbol{y}$ and $\boldsymbol{x}$ by using $\boldsymbol{B}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Lambda_y}$, $\boldsymbol{\Lambda_x}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Theta_\epsilon}$, $\boldsymbol{\Theta_\delta}$ and $\boldsymbol{\Psi}$.

$$\begin{aligned}
\boldsymbol{\Sigma_{\eta\eta}} &= (\boldsymbol{I}-\boldsymbol{B})^{-1}\left(\boldsymbol{\Gamma\Phi\Gamma}^T+\boldsymbol{\Psi}\right)(\boldsymbol{I}-\boldsymbol{B})^{-T}, && (7)\\
\boldsymbol{\Sigma_{yy}} &= \boldsymbol{\Lambda_y\Sigma_{\eta\eta}\Lambda_y}^T+\boldsymbol{\Theta_\epsilon}, && (8)\\
\boldsymbol{\Sigma_{xx}} &= \boldsymbol{\Lambda_x\Phi\Lambda_x}^T+\boldsymbol{\Theta_\delta}. && (9)
\end{aligned}$$

Second, define random vectors $\boldsymbol{\eta'}$, $\boldsymbol{\xi'}$, $\boldsymbol{\zeta'}$, $\boldsymbol{y'}$, $\boldsymbol{x'}$, $\boldsymbol{\epsilon'}$ and $\boldsymbol{\delta'}$, and the diagonal matrix $\boldsymbol{V_\eta}$, $\boldsymbol{V_\xi}$, $\boldsymbol{V_\zeta}$, $\boldsymbol{V_y}$, $\boldsymbol{V_x}$, $\boldsymbol{V_\epsilon}$ and $\boldsymbol{V_\delta}$ such that

$$\boldsymbol{\eta}=\boldsymbol{V_\eta\eta'},\ \ \boldsymbol{\xi}=\boldsymbol{V_\xi\xi'},\ \ \boldsymbol{\zeta}=\boldsymbol{V_\zeta\zeta'},\ \ \boldsymbol{y}=\boldsymbol{V_y y'},\ \ \boldsymbol{x}=\boldsymbol{V_x x'},\ \ \boldsymbol{\epsilon}=\boldsymbol{V_\epsilon\epsilon'},\ \ \boldsymbol{\delta}=\boldsymbol{V_\delta\delta'}.$$

Denote $(M)_{ii}$ be the i^{th} diagonal element of a matrix, the diagonal matrix can then be obtained by the following equations.

$$\begin{aligned}
\boldsymbol{V_\eta} &= \mathbf{diag}\left\{\sqrt{(\boldsymbol{\Sigma_{\eta\eta}})_{11}},\sqrt{(\boldsymbol{\Sigma_{\eta\eta}})_{22}},\cdots,\sqrt{(\boldsymbol{\Sigma_{\eta\eta}})_{mm}}\right\}. && (10)\\
\boldsymbol{V_\xi} &= \mathbf{diag}\left\{\sqrt{(\boldsymbol{\Phi})_{11}},\sqrt{(\boldsymbol{\Phi})_{22}},\cdots,\sqrt{(\boldsymbol{\Phi})_{nn}}\right\}. && (11)\\
\boldsymbol{V_\zeta} &= \mathbf{diag}\left\{\sqrt{(\boldsymbol{\Psi})_{11}},\sqrt{(\boldsymbol{\Psi})_{22}},\cdots,\sqrt{(\boldsymbol{\Psi})_{mm}}\right\}. && (12)\\
\boldsymbol{V_y} &= \mathbf{diag}\left\{\sqrt{(\boldsymbol{\Sigma_{yy}})_{11}},\sqrt{(\boldsymbol{\Sigma_{yy}})_{22}},\cdots,\sqrt{(\boldsymbol{\Sigma_{yy}})_{pp}}\right\}. && (13)\\
\boldsymbol{V_x} &= \mathbf{diag}\left\{\sqrt{(\boldsymbol{\Sigma_{xx}})_{11}},\sqrt{(\boldsymbol{\Sigma_{xx}})_{22}},\cdots,\sqrt{(\boldsymbol{\Sigma_{xx}})_{qq}}\right\}. && (14)\\
\boldsymbol{V_\epsilon} &= \mathbf{diag}\left\{\sqrt{(\boldsymbol{\Theta_\epsilon})_{11}},\sqrt{(\boldsymbol{\Theta_\epsilon})_{22}},\cdots,\sqrt{(\boldsymbol{\Theta_\epsilon})_{pp}}\right\}. && (15)\\
\boldsymbol{V_\delta} &= \mathbf{diag}\left\{\sqrt{(\boldsymbol{\Theta_\delta})_{11}},\sqrt{(\boldsymbol{\Theta_\delta})_{22}},\cdots,\sqrt{(\boldsymbol{\Theta_\delta})_{qq}}\right\}. && (16)
\end{aligned}$$

Then, substitute the above equations to the original model, we can have that

$$\begin{aligned}
\boldsymbol{V_\eta\eta'} &= \boldsymbol{BV_\eta\eta'}+\boldsymbol{\Gamma V_\xi\xi'}+\boldsymbol{V_\zeta\zeta'},\\
\boldsymbol{V_y y'} &= \boldsymbol{\Lambda_y V_\eta\eta'}+\boldsymbol{V_\epsilon\epsilon'}\\
\boldsymbol{V_x x'} &= \boldsymbol{\Lambda_x V_\xi\xi'}+\boldsymbol{V_\delta\delta'}.
\end{aligned}$$

Equivalently,

$$\begin{aligned}
\boldsymbol{\eta'} &= \boldsymbol{V_\eta^{-1} B V_\eta \eta' + V_\eta^{-1} \Gamma V_\xi \xi' + V_\eta^{-1} V_\zeta \zeta'}, && (17)\\
\boldsymbol{y'} &= \boldsymbol{V_y^{-1} \Lambda_y V_\eta \eta' + V_y^{-1} V_\epsilon \epsilon'}, && (18)\\
\boldsymbol{x'} &= \boldsymbol{V_x^{-1} \Lambda_x V_\xi \xi' + V_x^{-1} V_\delta \delta'}. && (19)
\end{aligned}$$

In the last step, the standardized form of the SEM can readily be obtained by setting

$$\boldsymbol{B' = V_\eta^{-1} B V_\eta}, \quad \boldsymbol{\Gamma' = V_\eta^{-1} \Gamma V_\xi},$$

$$\boldsymbol{\Lambda'_y = V_y^{-1} \Lambda_y V_\eta}, \quad \boldsymbol{\Lambda'_x = V_x^{-1} \Lambda_x V_\xi}.$$

The model will then be given by

$$\begin{aligned}
\boldsymbol{\eta'} &= \boldsymbol{B'\eta' + \Gamma'\xi' + \zeta''}, && (20)\\
\boldsymbol{y'} &= \boldsymbol{\Lambda'_y \eta' + \epsilon''}, && (21)\\
\boldsymbol{x'} &= \boldsymbol{\Lambda'_x \xi' + \delta''}. && (22)
\end{aligned}$$

In which, $\boldsymbol{\zeta'' = V_\eta^{-1} V_\zeta \zeta'}$, $\boldsymbol{\epsilon'' = V_y^{-1} V_\epsilon \epsilon'}$ and $\boldsymbol{\delta'' = V_x^{-1} V_\delta \delta'}$. By standardization, the covariance matrix for the latent vectors and the observable vectors will have the form of diagonal elements all ones, i.e.

$$\begin{bmatrix} 1 & * & \dots & * \\ * & 1 & \dots & * \\ \vdots & \vdots & & \vdots \\ * & * & \dots & 1 \end{bmatrix}.$$

A "$*$" symbol in the matrix corresponds a real number element.

1.3 Variants of SEM

Apart from defining the latent vectors and the measurement vectors have a linear relation with the latent vectors, the model can be defined in many other ways. Basically, an SEM (linear or nonlinear) can be defined following form.

$$\begin{aligned}
\boldsymbol{\eta} &= \boldsymbol{f(\eta, \xi, \zeta | \theta_f)}, && (23)\\
\boldsymbol{y} &= \boldsymbol{g(\eta, \epsilon | \theta_g)}, && (24)\\
\boldsymbol{x} &= \boldsymbol{h(\xi, \delta | \theta_h)}. && (25)
\end{aligned}$$

In which, $\boldsymbol{f}(\cdot)$, $\boldsymbol{g}(\cdot)$ and $\boldsymbol{h}(\cdot)$ are nonlinear vector functions with corresponding parametric vectors $\boldsymbol{\theta}_f$, $\boldsymbol{\theta}_g$ and $\boldsymbol{\theta}_h$.

Particular attention has to be paid when this model is applied. One reason is because no standard software tool has developed for this type of model. Even for a simple nonlinear model with single output ($p = 1$) like this.

$$\begin{aligned} \boldsymbol{\eta} &= \boldsymbol{B\eta} + \boldsymbol{\zeta}, \\ y &= \boldsymbol{\eta}^T \boldsymbol{G\eta} + \epsilon. \end{aligned}$$

Researcher usually defines an extended vector

$$\boldsymbol{\eta^*} = (\eta_1, \eta_2, \cdots, \eta_m, \eta_1\eta_1, \eta_1\eta_2, \cdots, \eta_i\eta_j, \cdots, \eta_m\eta_m)^T.$$

The model is re-expressed in the following form :

$$\begin{aligned} \boldsymbol{\eta^*} &= \boldsymbol{B^*\eta} + \boldsymbol{\zeta^*}, \\ y &= \boldsymbol{\Lambda^*\eta^*} + \epsilon, \end{aligned}$$

and then apply standard software to solve the problem. Special care has to be aware as the elements in $\boldsymbol{\eta^*}$ are no longer Gaussian distributed. Analysis results obtained from this setting can only be a reference. Further analysis is needed.

2 Parametric Estimation

Normally, the true model is unknown. One can have a set of observations, $\mathcal{D} = \{\boldsymbol{x}_k, \boldsymbol{y}_k\}_{k=1}^N$. Remind that $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ are q-vector and p-vector. For clarification, we denote the elements in $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ by

$$\begin{aligned} \boldsymbol{x}_k &= (x_{k1}, x_{k2}, \cdots, x_{ki}, \cdots, x_{kq})^T, \\ \boldsymbol{y}_k &= (y_{k1}, y_{k2}, \cdots, y_{ki}, \cdots, y_{kp})^T. \end{aligned}$$

Let $\boldsymbol{S} \in R^{(p+q)\times(p+q)}$ be the sample covariance matrix for $\mathcal{D}$. Without loss of generality, we assume that the mean of $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ are zero vectors. Then,

$$\boldsymbol{S} = \begin{bmatrix} \frac{1}{N}\sum_{k=1}^N \boldsymbol{y}_k\boldsymbol{y}_k^T & \frac{1}{N}\sum_{k=1}^N \boldsymbol{y}_k\boldsymbol{x}_k^T \\ \frac{1}{N}\sum_{k=1}^N \boldsymbol{x}_k\boldsymbol{y}_k^T & \frac{1}{N}\sum_{k=1}^N \boldsymbol{x}_k\boldsymbol{x}_k^T \end{bmatrix}. \tag{26}$$

In accordance with Equations (7) to (9), the respective covariance matrix of a model with parametrix vector $\boldsymbol{\theta}$ (denoted by $\boldsymbol{\Sigma(\theta)}$) will be given by,

$$\boldsymbol{\Sigma(\theta)} = \begin{bmatrix} \boldsymbol{\Lambda_y}(\boldsymbol{I}-\boldsymbol{B})^{-1}\left(\boldsymbol{\Gamma\Phi\Gamma}^T + \boldsymbol{\Psi}\right)(\boldsymbol{I}-\boldsymbol{B})^{-T}\boldsymbol{\Lambda_y}^T + \boldsymbol{\Theta_\epsilon} & \boldsymbol{\Lambda_y}(\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{\Phi\Lambda_x}^T \\ \boldsymbol{\Lambda_x\Phi}(\boldsymbol{I}-\boldsymbol{B})^{-T}\boldsymbol{\Lambda_y}^T & \boldsymbol{\Lambda_x\Phi\Lambda_x}^T + \boldsymbol{\Theta_\delta} \end{bmatrix}. \tag{27}$$

To estimate the true model parameters, a few common objective (fitting) functions (to be minimized) are usually applied.

$$
\begin{aligned}
F_{LS}(\boldsymbol{\theta}) &= \mathbf{Tr}\left\{(\boldsymbol{S}-\boldsymbol{\Sigma}(\boldsymbol{\theta}))^2\right\}. && (28)\\
F_{LL}(\boldsymbol{\theta}) &= \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \mathbf{Tr}\left\{\boldsymbol{S}\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\right\}. && (29)\\
F_{LR}(\boldsymbol{\theta}) &= \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \log|\boldsymbol{S}| + \mathbf{Tr}\left\{\boldsymbol{S}\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\right\} - (p+q). && (30)\\
F_{PL}(\boldsymbol{\theta}) &= \mathbf{Tr}\left\{(\boldsymbol{S}-\boldsymbol{\Sigma}(\boldsymbol{\theta}))^2\right\} + \mathcal{R}_{PL}(\boldsymbol{\theta}). && (31)\\
F_{AP}(\boldsymbol{\theta}) &= \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \mathbf{Tr}\left\{\boldsymbol{S}\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\right\} + \mathcal{R}_{AP}(\boldsymbol{\theta}). && (32)
\end{aligned}
$$

The subscripts LS, LL, LR, PL and LB stand for *least square*, *log likelihood*, *log of likelihood ratio*, *penalized least-square* and *log a posterior*.

2.1 Unweighted least squares

The first objective evaluates the deviation between the estimated covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and the sample covariance matrix $\boldsymbol{S}$ in the sense of sum square errors. That is,

$$
\mathbf{Tr}\left\{(\boldsymbol{S}-\boldsymbol{\Sigma}(\boldsymbol{\theta}))^2\right\} = \sum_{i=1}^{p+q}\sum_{j=1}^{p+q}\left(\boldsymbol{S}_{ij}-\boldsymbol{\Sigma}(\boldsymbol{\theta})_{ij}\right)^2. \tag{33}
$$

2.2 Log likelihood

The second objective applies under the normality assumption. Consider the sample vectors in the dataset $\mathcal{D}$ are random drawn from a Gaussian distribution which is given by

$$
\mathbf{Pr}(\boldsymbol{w}|\boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi)^{(p+q)}|\boldsymbol{\Sigma}(\boldsymbol{\theta})|}}\exp\left\{-\frac{1}{2}\,\mathbf{Tr}\{\boldsymbol{w}\boldsymbol{w}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\}\right\}, \tag{34}
$$

where $\boldsymbol{w} = (\boldsymbol{y}^T\ \boldsymbol{x}^T)^T$. The log-likelihood of $\mathcal{D}$ conditioned that the covariance matrix is $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is given as follows :

$$
\mathcal{L}(\mathcal{D}|\boldsymbol{\theta}) = -\frac{N(p+q)}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \frac{N}{2}\mathbf{Tr}\left\{\boldsymbol{S}\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\right\}. \tag{35}
$$

Thus, the second objective function $F_{LL}(\boldsymbol{\theta})$ has a direct relation to log-likelihood given by the following equality.

$$
F_{LL}(\boldsymbol{\theta}) = -\frac{2}{N}\left\{\mathcal{L}(\mathcal{D}|\boldsymbol{\theta}) - (p+q)\log(2\pi)\right\}.
$$

Besides, it can be shown that the maximum of $\mathcal{L}(\mathcal{D}|\boldsymbol{\theta})$ in Equation (35) is attained at $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{S}$. (See p.62 of [1] for the proof!)

$$
\mathcal{L}(\mathcal{D}|\boldsymbol{S}) = -\frac{N(p+q)}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{S}| - (p+q). \tag{36}
$$

2.3 Log likelihood ratio

The third objective function $F_{LR}(\boldsymbol{\theta})$ has a direct relation to $\mathcal{L}(\mathcal{D}|\boldsymbol{S})$ and $\mathcal{L}(\mathcal{D}|\boldsymbol{\theta})$.

$$F_{LR}(\boldsymbol{\theta}) = -\frac{2}{N}\left\{\mathcal{L}(\mathcal{D}|\boldsymbol{S}) - \mathcal{L}(\mathcal{D}|\boldsymbol{\theta})\right\}.$$

The beauty of $F_{LR}(\boldsymbol{\theta})$ is that its value can be used for likelihood ratio test to reject or accept a model $\boldsymbol{\theta}$ [3, 4].

$$\text{Likelihood Ratio} = \exp\left\{-\frac{N}{2}F_{LR}(\boldsymbol{\theta})\right\}.$$

Since p, q and $\boldsymbol{S}$ are independent of $\boldsymbol{\theta}$, minimizing F_{LL} will give the same solution as minimizing F_{LR}. That is,

$$\arg\min_{\boldsymbol{\theta}}\left\{F_{LL}(\boldsymbol{\theta})\right\} = \arg\min_{\boldsymbol{\theta}}\left\{F_{LR}(\boldsymbol{\theta})\right\}.$$

For some advanced techniques, like feature selection and model reduction, the solutions obtained by minimizing F_{LL} will be different from the solution obtained minimizing F_{LR}.

2.4 Regularized ULS

The last two objective functions are basically a natural extension of the F_{LS} and F_{LL} by adding constraints on the parametric vector $\boldsymbol{\theta}$). The additional terms are called regularizer which has the following properties.

(i) $\mathcal{R}(\boldsymbol{\theta}) \geq 0$ for all $\|\boldsymbol{\theta}\| \geq 0$. Equality holds when $\|\boldsymbol{\theta}\| = 0$.

(ii) $\mathcal{R}(\boldsymbol{\theta'}) > \mathcal{R}(\boldsymbol{\theta''})$ if $\|\boldsymbol{\theta'}\| > \|\boldsymbol{\theta''}\|$.

A common choice for $\mathcal{R}(\boldsymbol{\theta})$ is that

$$\mathcal{R}(\boldsymbol{\theta}) = \text{cont.} \times\ \boldsymbol{\theta}^T\boldsymbol{R}\boldsymbol{\theta},$$

where $\boldsymbol{R}$ is a positive definite matrix.

2.5 Log *a posterior* probability

For F_{MAP}, the regularization term plays a role as adding a logarithm of *a prior* probability ($\log \mathbf{Pr}(\boldsymbol{\theta})$) to the log-likelihood ($\log \mathbf{Pr}(\mathcal{D}|\boldsymbol{\theta})$) for the evaluation of the fitness of a model. The idea is essentially the same as evaluating the model fitness by its *a posterior* probability. Consider the *a posterior* probability of getting $\boldsymbol{\theta}$ given $\mathcal{D}$.

$$\mathbf{Pr}(\boldsymbol{\theta}|\mathcal{D}) = \frac{\mathbf{Pr}(\mathcal{D}|\boldsymbol{\theta}) \times\ \mathbf{Pr}(\boldsymbol{\theta})}{\mathbf{Pr}(\mathcal{D})}.$$

By virtue that the solution of $\max\{\mathbf{Pr}(\boldsymbol{\theta}|\mathcal{D})\}$ is identical to the solution of $\max\{\log \mathbf{Pr}(\boldsymbol{\theta}|\mathcal{D})\}$, we consider the logarithm of the above equation for simplicity.

$$\log \mathbf{Pr}(\boldsymbol{\theta}|\mathcal{D}) = \underbrace{\log \mathbf{Pr}(\mathcal{D}|\boldsymbol{\theta})}_{\text{Likelihood}} + \underbrace{\log \mathbf{Pr}(\boldsymbol{\theta})}_{\text{Prior}} - \log \mathbf{Pr}(\mathcal{D}). \tag{37}$$

Define the *a prior* distribution as follows :

$$\mathbf{Pr}(\boldsymbol{\theta}) = \kappa^{-1} \exp\left\{-\mathcal{R}(\boldsymbol{\theta})\right\}. \tag{38}$$

Here κ is the normalization constant, i.e. $\kappa = \int \exp\left\{-\mathcal{R}(\boldsymbol{\theta})\right\} d\boldsymbol{\theta}$. Together with the Equation (35),

$$\log \mathbf{Pr}(\boldsymbol{\theta}|\mathcal{D}) = -\frac{N}{2} \log |\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \frac{N}{2} \mathbf{Tr}\left\{\boldsymbol{S}\, \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\right\} - \mathcal{R}(\boldsymbol{\theta}) - \text{constant}. \tag{39}$$

The constant term is equal to $\frac{N(p+q)}{2} \log(2\pi) + \log \mathbf{Pr}(\mathcal{D})$. Hence,

$$F_{MAP}(\boldsymbol{\theta}) = F_{LL}(\boldsymbol{\theta}) + \mathcal{R}(\boldsymbol{\theta}) + \text{constant}. \tag{40}$$

3 Solution Space for F_{LL}

Although the objective function provides a mean for searching an optimal estimator, there might exist infinite number of estimators that have the same optimality.

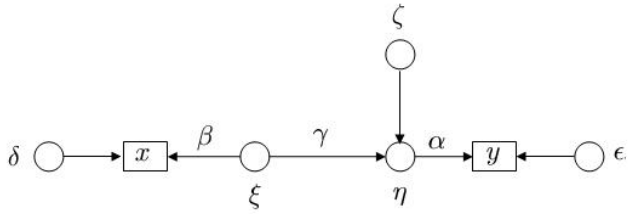

Figure 3: Simple SEM.

3.1 Case I : $d' < d$

To explain such situation, let us have the following simple SEM, Figure 3. All variables and parameters are scalars.

$$\eta = \gamma\xi + \zeta, \quad y = \alpha\eta + \epsilon, \quad x = \beta\xi + \delta.$$

By convention, we assume that the variance of random variables η and ξ are unity, it is not difficult to show that the covariance matrix for $(y\ x)^T$ is a Gaussian distribution with

$$\Sigma(\theta) = \begin{bmatrix} \alpha^2 + \Theta_\epsilon & \alpha\gamma\beta \\ \alpha\gamma\beta & \beta^2 + \Theta_\delta \end{bmatrix}.$$

In this example, it is clear that $d' = 3$ and $d = 8$. Without further constraints on model parameters, the total number of equations is small than the total number of parameters.

Suppose, the sample covariance for y and x are obtained as follows :

$$\Sigma_{yy} = 1, \quad \Sigma_{xx} = 2.25, \quad \Sigma_{xy} = \Sigma_{yx} = 1.5.$$

Using either F_{LL} or F_{LR} to be the objective function, the optimal solution for θ will be the one satisfying the following condition.

$$\Sigma(\theta) = \begin{bmatrix} 1 & 1.5 \\ 1.5 & 2.25 \end{bmatrix}.$$

With an addition condition that

$$E[\eta^2] = 1 \quad \text{or equivalently} \quad \gamma^2\phi + \Psi = 1, \tag{41}$$

$$E[\xi^2] = 1 \quad \text{or equivalently} \quad \phi = 1, \tag{42}$$

the variables α, γ and β will fulfill the following system of equations.

$$\alpha^2 = 1 - \Theta_\epsilon, \quad \alpha\gamma\beta = 1.5, \quad \beta^2 = 2.25 - \Theta_\delta. \tag{43}$$

As a result, there are five equations for the determination of eight parameters. There will have infinite many solutions for the model.

After simple algebraic manipulation on Equations (41), (42) and (43),

$$\gamma^2 = \frac{2.25}{(1 - \Theta_\epsilon)(2.25 - \Theta_\delta)}.$$

Taking Θ_ϵ and Θ_δ be the free parameters with the constraints that

$$0 \le \Theta_\epsilon < \Sigma_{yy} = 1 \ \text{ and } \ 0 \le \Theta_\delta < \Sigma_{xx} = 2.25,$$

the projection of the solution surface (manifold) on the $(\alpha^2, \gamma^2, \beta^2)$-Space is shown in Figure 4. As all the points on the surface have the same optimal F_{LL} (and F_{LR}) values. Different initial estimate on the parameters or different stopping criteria for an estimation method could come up with different solutions.

It should also be noted that the surface shown in Figure 4 is in the $(\alpha^2, \gamma^2, \beta^2)$-Space. For (α, γ, β)-Space, there will have more than one surface because there is no restriction on the signs of α, γ and β. They can take both positive as well as negative values. With reference to the above example and given $(\Theta_\epsilon, \Theta_\delta)$, there are four solutions, including (α, γ, β), $(-\alpha, \gamma, -\beta)$, $(-\alpha, -\gamma, \beta)$ and $(\alpha, -\gamma, -\beta)$, satisfying the equalities in Equation (43). The solution set appears as four surfaces in the (α, γ, β)-Space. Discontinuities exist at either $\alpha = 0$ or $\beta = 0$.

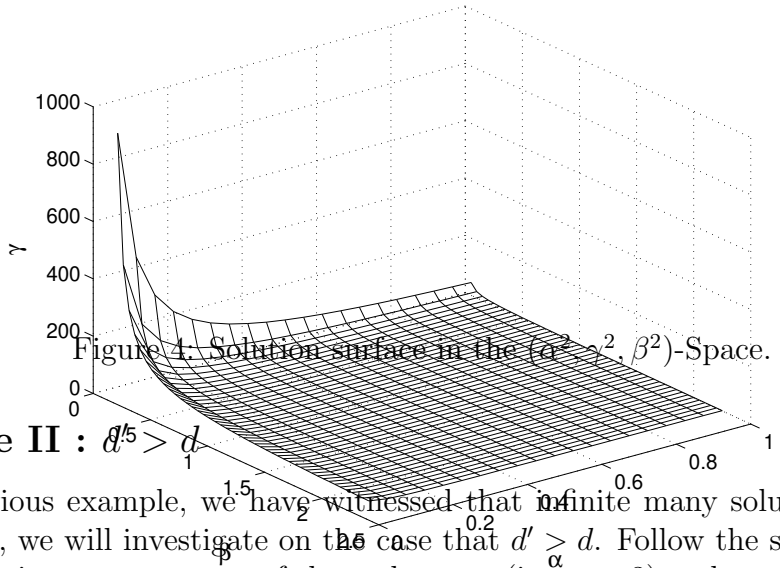

Figure 4: Solution surface in the $(\alpha^2 = \gamma^2, \beta^2)$-Space.

3.2 Case II : $d' > d$

For the previous example, we have witnessed that infinite many solutions can exist if $d' < d$. Next, we will investigate on the case that $d' > d$. Follow the same example but we define the input as a vector of three elements (i.e. $q = 3$) and output as a vector of two elements (i.e. $p = 2$). The model is defined in a similar fashion as before.

$$\eta = \gamma\xi + \zeta, \quad \boldsymbol{y} = \boldsymbol{\alpha}\eta + \boldsymbol{\epsilon}, \quad \boldsymbol{x} = \boldsymbol{\beta}\xi + \boldsymbol{\delta}.$$

Given the sample covariance matrix, the MLE will be the solution satisfies the following matrix equations,

$$\begin{aligned} \boldsymbol{\alpha}\boldsymbol{\alpha}^T + \boldsymbol{\Theta}_\epsilon &= \boldsymbol{\Sigma}_{xx}, & (44)\\ \gamma\,\boldsymbol{\alpha}\boldsymbol{\beta}^T &= \boldsymbol{\Sigma}_{xy}, & (45)\\ \boldsymbol{\beta}\boldsymbol{\beta}^T + \boldsymbol{\Theta}_\delta &= \boldsymbol{\Sigma}_{xx}, & (46) \end{aligned}$$

under the constraints that

$$\gamma^2\phi + \Psi = 1 \quad \text{and} \quad \phi = 1.$$

Since the covariance matrices $\boldsymbol{\Theta}_\epsilon$ and $\boldsymbol{\Theta}_\delta$ are diagonal matrices, the total number of parameters excluding ϕ and Ψ (i.e. $d-2$) will be 12. The total number of equations d' is 15. It should have enough information for the determination of the parameters.

However, it has no guarantee. It can have three possible situations : (1) no solution, (2) one solution and (3) infinite many solution. While a model has been estimated, one can investigate on the RMR index to identify whether $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{S}$. If it is non-zero and the magnitude is large, one can increase the complexity of the model by adding more parameters. So that the possibility of getting a close to optimal solution can be made.

4 Optimization Techniques

Once an objective function has been defined for optimality, two questions are remained for answer.

(1) How to find this *optimal* estimator $\hat{\boldsymbol{\theta}}$ in the parametric space R^{p+q} ?

(2) Whether there are more than one optimal solution.

The answer for the first question relies on the use of optimization technique. For the second question, we provide an answer in the Appendix. Suppose the objective function is defined as *log likelihood*, there might exist infinite many optimal solutions. As the analysis on the uniqueness of optimal solution is always a complicated problem, we leave it open here.

To find an optimal solution for an objective function, it is simply a problem in optimization. In accordance with optimization theory, many techniques can be applied. For a smooth function, gradient descent and Newton's method are two common iterative procedures that can search step by step and eventually reach to an optimal solution. In the context of parametric estimation, the idea of applying optimization technique can be visualized by Figure 5.

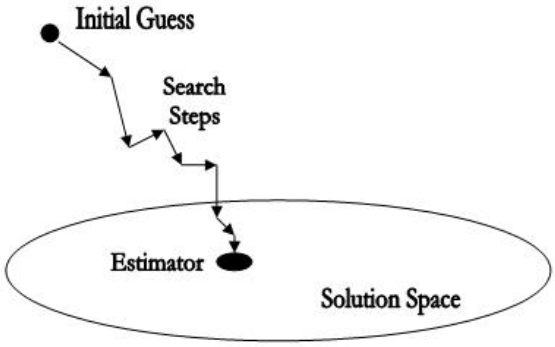

Figure 5: The idea of optimization search.

To explain the idea mathematically, let us introduce t as the index for the number of iteration steps and μ as the step size. The search can be viewed as a realization of a sequence $\{\hat{\boldsymbol{\theta}}(t)|t = 1, 2, \cdots\}$ such that $\lim_{t\to\infty} \hat{\boldsymbol{\theta}}(t)$ is an optimal solution based on an initial guess $\hat{\boldsymbol{\theta}}(0)$.

$$\hat{\boldsymbol{\theta}}(0) \to \hat{\boldsymbol{\theta}}(1) \to \hat{\boldsymbol{\theta}}(2) \to \ldots \to \hat{\boldsymbol{\theta}}(t) \to \hat{\boldsymbol{\theta}}(t+1) \to \ldots$$

For gradient descent, the arrow corresponds to the following iteration.

$$\hat{\boldsymbol{\theta}}(t+1) = \hat{\boldsymbol{\theta}}(t) - \mu \frac{\partial}{\partial \boldsymbol{\theta}} F(\hat{\boldsymbol{\theta}}(t)). \tag{47}$$

For Newton's method,

$$\hat{\boldsymbol{\theta}}(t+1) = \hat{\boldsymbol{\theta}}(t) - \mu \left[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} F(\hat{\boldsymbol{\theta}}(t))\right]^{-1} \frac{\partial}{\partial \boldsymbol{\theta}} F(\hat{\boldsymbol{\theta}}(t)). \tag{48}$$

The success of applying gradient descent and Newton's method rely on the conditions that

(1) $F(\boldsymbol{\theta})$ is differentiable, and

(2) $\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} F(\hat{\boldsymbol{\theta}}(t))$ should not be near-singular

for all $\boldsymbol{\theta}$. One should be aware that these conditions might not always be ensured, especially Condition (2). In this regard, various techniques extended from Newton's method have been proposed. Interested reader can refer to any optimization theory textbook for detail.

As there might have infinite optimal solutions for an objective function, (i) different search techniques, (ii) different initial guesses, (iii) different step size μ and (iv) different stopping criteria might lead to different estimators for $\boldsymbol{\theta}$.

Certainly, optimization technique is not the only approach for the search. For some situations, (see Appendix for an example), analytical equations can be obtained for the solution. Iterative procedures in *numerical method* could be applied.

5 EM Algorithm

EM (expectation-maximization) algorithm is an iterative procedure that can maximize the marginal log-likelihood of a model with missing information, i.e.

$$\ell(Y|\boldsymbol{\theta}) = \int \mathbf{Pr}(Z|\boldsymbol{\theta}) \log \mathbf{Pr}(Y|Z, \boldsymbol{\theta}) dZ, \tag{49}$$

by repeating the following **E-Step** and **M-Step** until converge. Here $\boldsymbol{\theta}$ denotes the parametric vector of the model.

E-Step : Evaluate the expectation of $\log \mathbf{Pr}(Y, Z|\boldsymbol{\theta})$ on Z using $\mathbf{Pr}(Z|Y, \hat{\boldsymbol{\theta}}^t)$.

$$Q(\theta|Y, \hat{\theta}^t) = \int \mathbf{Pr}(Z|Y, \hat{\boldsymbol{\theta}}_t) \log \mathbf{Pr}(Y, Z|\boldsymbol{\theta}) dZ.$$

M-Step : Maximize $Q(\boldsymbol{\theta}|Y)$ and set

$$\hat{\boldsymbol{\theta}}^{t+1} = \arg\max_{\boldsymbol{\theta}} \{Q(\boldsymbol{\theta}|Y)\}.$$

Here $Y = \{y_k\}_{k=1}^N$ is the set of observable data. While $Z = \{z_k\}_{k=1}^N$ is the set of missing information. In the following text, the function $Q(\theta|Y, \hat{\theta}^t)$ is also denoted by $Q(\theta|\hat{\theta}^t)$ or $Q(\theta|Y, \hat{Z}^t)$ for simplicity.

5.1 Factor analysis model

A FA model is a model which output $\boldsymbol{y} \in R^p$ is depended on an un-observed latent factors $\boldsymbol{\eta} \in R^m$, i.e.

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\epsilon}. \tag{50}$$

$\boldsymbol{\mu}$ is a constant vector representing the mean of $\boldsymbol{y}$. $\boldsymbol{\Lambda}$ is the matrix of factor loadings. By convention, $\boldsymbol{\mu}$ is assumed null. Besides,

$$\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}_m, \boldsymbol{I}_{m\times m}) \quad \text{and} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Theta}). \tag{51}$$

For presentation simplicity, those subscripts in ± 0 and $\boldsymbol{I}$ will not be shown.

Since the latent factor $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$ are Gaussian, the marginal probability distribution of $\boldsymbol{y}$ given $\boldsymbol{\theta}$ and the conditional probability distribution of $\boldsymbol{x}|\boldsymbol{y}$ are also Gaussian. In accordance with the model defined in Equation (50) and the conditions depicted in Equation (51),

$$\begin{bmatrix} \boldsymbol{\eta} \\ \boldsymbol{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I} & \boldsymbol{\Lambda}^T \\ \boldsymbol{\Lambda} & \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Theta} \end{bmatrix} \right). \tag{52}$$

Using the results in the conditional probability, the mean and variance of the conditional probability distribution of $\boldsymbol{\eta}$ given $\boldsymbol{y}$ is given by

$$\begin{aligned} E[\boldsymbol{\eta}|\boldsymbol{y}, \boldsymbol{\theta}] &= \boldsymbol{\Lambda}^T(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Theta})^{-1}\boldsymbol{y}, & (53)\\ E[\boldsymbol{\eta}\boldsymbol{\eta}^T|\boldsymbol{y}, \boldsymbol{\theta}] &= \boldsymbol{I} - \boldsymbol{\Lambda}^T(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Theta})^{-1}\boldsymbol{\Lambda}. & (54)\\ &= (I + \boldsymbol{\Lambda}^T\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda})^{-1}. & (55) \end{aligned}$$

The last equality is based on Equation (107). It should note that the conditional covariance matrix is independent of $\boldsymbol{y}$. (Surprise!) The *complete information* log-likelihood

can then be expressed as follows :

$$\begin{aligned}\log \mathbf{Pr}(\boldsymbol{y}_k, \boldsymbol{\eta}_k|\boldsymbol{\theta}) &= -\frac{p+m}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Theta}| - \frac{1}{2}\mathbf{Tr}\{\boldsymbol{y}_k\boldsymbol{y}_k^T\} \\ &\quad + \mathbf{Tr}\left\{\boldsymbol{\eta}_k\boldsymbol{y}_k^T\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda}\right\} \\ &\quad - \frac{1}{2}\mathbf{Tr}\left\{\boldsymbol{\eta}_k\boldsymbol{\eta}_k^T(\boldsymbol{I} + \boldsymbol{\Lambda}^T\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda})\right\}.\end{aligned}$$

Now we let $\hat{Z}^t$ bet the set of posterior estimation of $\{\boldsymbol{\eta}_k^t\}_{k=1}^N$ given Y in the t^{th} step,

$$\begin{aligned}\boldsymbol{\Sigma}_{yy} &= \frac{1}{N}\sum_k \boldsymbol{y}_k\boldsymbol{y}_k^T, \\ \hat{\boldsymbol{\Sigma}}_{\eta y}^t &= \frac{1}{N}\sum_k E[\boldsymbol{\eta}_k|\boldsymbol{y_k}, \boldsymbol{\theta}]\boldsymbol{y}_k^T, \\ \hat{\boldsymbol{\Sigma}}_{y\eta}^t &= \frac{1}{N}\sum_k \boldsymbol{y}_k E[\boldsymbol{\eta}_k|\boldsymbol{y_k}, \boldsymbol{\theta}]^T, \\ \hat{\boldsymbol{\Sigma}}_{\eta\eta}^t &= \frac{1}{N}\sum_k E[\boldsymbol{\eta}_k\boldsymbol{\eta}_k^T|\boldsymbol{y_k}, \boldsymbol{\theta}].\end{aligned}$$

The superscript t in the matrices are with the same meaning as for $\hat{Z}^t$. The expected complete information log-likelihood can thus be obtained.

$$\begin{aligned}Q(\theta|Y, \hat{Z}^t) &= -\frac{(p+m)N}{2}\log 2\pi - \frac{N}{2}\log|\boldsymbol{\Theta}| - \frac{N}{2}\mathbf{Tr}\left\{\boldsymbol{\Sigma}_{yy}\right\} \\ &\quad + N\mathbf{Tr}\left\{\hat{\boldsymbol{\Sigma}}_{\eta y}^t\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda}\right\} - \frac{N}{2}\mathbf{Tr}\left\{\hat{\boldsymbol{\Sigma}}_{\eta\eta}^t(\boldsymbol{I} + \boldsymbol{\Lambda}^T\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda})\right\}. \qquad (56)\end{aligned}$$

Taking the gradient of $Q(\theta|Y, \hat{Z}^t)$ with respect to matrices Λ, and setting it to null, one can get that

$$\left(\hat{\boldsymbol{\Sigma}}_{\eta y}^t\boldsymbol{\Theta}^{-1}\right)^T - \boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda}\hat{\boldsymbol{\Sigma}}_{\eta\eta}^t = \mathbf{0}.$$

As $\boldsymbol{\Theta}$ is symmetric, its inverse is also a symmetric matrix. This implies that

$$\boldsymbol{\Lambda} = \hat{\boldsymbol{\Sigma}}_{y\eta}^t(\hat{\boldsymbol{\Sigma}}_{\eta\eta}^t)^{-1}. \qquad (57)$$

For the matrix $\boldsymbol{\Theta}$, which is a restricted to a diagonal matrix with positive elements. Using Equation (120) and Equation (57), and setting the gradient of $Q(\theta|Y, \hat{Z}^t)$ with respect to null, one can obtain that

$$\boldsymbol{\Theta} = \mathbf{diag}\left\{\boldsymbol{\Lambda}\hat{\boldsymbol{\Sigma}}_{\eta y}^t\right\}. \qquad (58)$$

As a result, the EM algorithm for maximizing the marginal log-likelihood, Equation (49), can be accomplished by running the following steps iteratively.

E-Step : Evaluate $\hat{\mathbf{\Sigma}}^t_{y\eta}$ and $\hat{\mathbf{\Sigma}}^t_{\eta\eta}$ conditioned on Y, $\mathbf{\Lambda}^t$ and $\mathbf{\Theta}^t$.

M-Step : Evaluate $\mathbf{\Lambda}^{t+1}$ and $\mathbf{\Theta}^{t+1}$ by

$$\begin{aligned}\mathbf{\Lambda}^{t+1} &= \hat{\mathbf{\Sigma}}^t_{y\eta}(\hat{\mathbf{\Sigma}}^t_{\eta\eta})^{-1}.\\ \mathbf{\Theta}^{t+1} &= \mathbf{diag}\left\{\mathbf{\Lambda}^{t+1}(\hat{\mathbf{\Sigma}}^t_{y\eta})^T\right\}.\end{aligned}$$

5.2 Confirmatory FA model

Above formulation applies for a factor loading matrix in which all elements are free parameters. For a CFA model, certain factor loadings are either 0 or set to a constants. In such case, modification on the M-Step will be needed for running above EM algorithm.

The simplest way to describe this modification is better from an example. Suppose a factor loading is defined as follows :

$$\Lambda = \begin{bmatrix} 1 & 0 & 0 \\ a_{21} & 0 & 0 \\ a_{31} & a_{32} & 0 \\ 0 & a_{42} & 0.5 \\ 0 & a_{52} & a_{53} \end{bmatrix}. \tag{59}$$

It corresponds to an FA model with $\boldsymbol{y} \in R^5$ and $\boldsymbol{\eta} \in R^3$. After the E-Step, we have the values for $\hat{\boldsymbol{\eta}}$ and $\hat{\mathbf{\Sigma}}$. So, what we need to do next is to determine the parameteric values for a_{21}, a_{31}, a_{32}, a_{42}, a_{52} and a_{53}.

Consider the output of y_3,

$$y_3 = a_{31}\eta_1 + a_{32}\eta_2 + \epsilon_3,$$

which is independent of other parameters. For notation simplicity, we denote the posterior expectation by $\langle\cdot\rangle$.

$$\begin{aligned}\langle(y_3 - a_{31}\eta_1 - a_{32}\eta_2)^2\rangle &= y_3^2 + a_{31}^2\langle\eta_1^2\rangle + a_{32}^2\langle\eta_2^2\rangle - 2a_{31}\langle y_3\eta_1\rangle \\ &\quad - 2a_{32}\langle y_3\eta_2\rangle + 2a_{31}a_{32}\langle\eta_1\eta_2\rangle\end{aligned}$$

The derivative of $Q(\theta|Y, \hat{Z}^t)$ with respect to a_{31} and a_{32} will be given by

$$\begin{aligned}\frac{\partial Q(\theta|Y,\hat{Z}^t)}{\partial a_{31}} &= (2a_{31}\langle\eta_1^2\rangle - 2\langle y_3\eta_1\rangle + 2a_{32}\langle\eta_1\eta_2\rangle)/\mathbf{\Theta}_{33},\\ \frac{\partial Q(\theta|Y,\hat{Z}^t)}{\partial a_{32}} &= (2a_{32}\langle\eta_2^2\rangle - 2\langle y_3\eta_2\rangle + 2a_{31}\langle\eta_1\eta_2\rangle)/\mathbf{\Theta}_{33}.\end{aligned}$$

Setting both gradients to zero, one will obtain two equations for two variables. Sum up for data $k = 1, 2, \cdots, N$, the update for a_{31} and a_{32}, can be realized by

$$\begin{bmatrix} a_{31} \\ a_{32} \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{\eta_1\eta_1} & \hat{\boldsymbol{\Sigma}}_{\eta_1\eta_2} \\ \hat{\boldsymbol{\Sigma}}_{\eta_2\eta_1} & \hat{\boldsymbol{\Sigma}}_{\eta_2\eta_2} \end{bmatrix}^{-1} \begin{bmatrix} N^{-1}\sum_{k=1}^{N} y_3(k)\hat{\eta}_1(k) \\ N^{-1}\sum_{k=1}^{N} y_3(k)\hat{\eta}_2(k) \end{bmatrix}. \tag{60}$$

So, the update of parameters in M-Steps for factor loading $\boldsymbol{\Lambda}$ has to be done row by row. For the i^{th} row, let $\boldsymbol{a}_i = (a_{i\pi_1}, a_{i\pi_2}, \cdots, a_{i\pi_i})$ be the parametric vector to be estimated. Construct a submatrix $\tilde{M}$ from $\hat{\boldsymbol{\Sigma}}_{\eta\eta}$ and a vector $\tilde{Y}$ such that

$$(\tilde{M})_{rs} = \left(\hat{\boldsymbol{\Sigma}}^t_{\eta\eta}\right)_{\pi_r\pi_s} \tag{61}$$

$$\tilde{Y}_r = N^{-1}\sum_{k=1}^{N} y_3(k)\hat{\boldsymbol{\eta}}^t_{\pi_r}(k). \tag{62}$$

Then, the estimation of $\hat{\boldsymbol{a}}_i$ can be accomplished by

$$\hat{\boldsymbol{a}}_i^{t+1} = \tilde{M}^{-1}\tilde{Y}. \tag{63}$$

The above procedure repeats until $i = p$. As a result, the EM algorithm for maximizing the marginal log-likelihood, Equation (49), can be accomplished by running the following steps repeatedly until converge.

E-Step : Evaluate $\hat{\boldsymbol{\Sigma}}^t_{y\eta}$ and $\hat{\boldsymbol{\Sigma}}^t_{\eta\eta}$ conditioned on Y, $\boldsymbol{\Lambda}^t$ and $\boldsymbol{\Theta}^t$.

M-Step : Evaluate the row vectors of $\boldsymbol{\Lambda}^{t+1}$ by Equation (61), Equation (62) and Equation (63), and then evaluate $\boldsymbol{\Theta}^{t+1}$ by

$$\boldsymbol{\Lambda}^{t+1} = \begin{bmatrix} \hat{\boldsymbol{a}}_1^{t+1} \\ \hat{\boldsymbol{a}}_2^{t+1} \\ \vdots \\ \hat{\boldsymbol{a}}_p^{t+1} \end{bmatrix},$$

$$\boldsymbol{\Theta}^{t+1} = \mathbf{diag}\left\{\boldsymbol{\Lambda}^{t+1}(\hat{\boldsymbol{\Sigma}}^t_{y\eta})^T\right\}.$$

5.3 Structural equation model

Recall that an SEM is defined as follows :

$$\begin{aligned} \boldsymbol{\eta} &= \boldsymbol{B\eta} + \boldsymbol{\Gamma\xi} + \boldsymbol{\zeta} \\ \boldsymbol{y} &= \boldsymbol{\Lambda_y\eta} + \boldsymbol{\epsilon} \\ \boldsymbol{x} &= \boldsymbol{\Lambda_x\xi} + \boldsymbol{\delta}, \end{aligned}$$

where $\boldsymbol{\eta} \in R^m$, $\boldsymbol{\xi} \in R^n$, $\boldsymbol{y} \in R^p$ and $\boldsymbol{x} \in R^q$. The logarithm of the joint probability $\mathbf{Pr}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{\xi}|\boldsymbol{\theta})$ for the k^{th} data is given by

$$\begin{aligned} \log \mathbf{Pr}(\boldsymbol{y}_k, \boldsymbol{x}_k, \boldsymbol{\eta}_k, \boldsymbol{\xi}_k|\boldsymbol{\theta}) &= \log \mathbf{Pr}(\boldsymbol{y}_k|\boldsymbol{\eta}_k, \boldsymbol{\xi}_k, \boldsymbol{\theta}) + \log \mathbf{Pr}(\boldsymbol{\eta}_k|\boldsymbol{\xi}_k, \boldsymbol{\theta}) \quad (64) \\ &\quad + \log \mathbf{Pr}(\boldsymbol{x}_k|\boldsymbol{\xi}_k, \boldsymbol{\theta}) + \log \mathbf{Pr}(\boldsymbol{\xi}_k|\boldsymbol{\theta}). \quad (65) \end{aligned}$$

As all the probabilities follow Gaussian distribution and let $\boldsymbol{A} = (\boldsymbol{I} - \boldsymbol{B})^{-1}$, the log-probability terms in the above equations can be written in either sum of square errors form or in matrix trace form. In the former case,

$$\begin{aligned} -2 \log \mathbf{Pr}(\boldsymbol{y}_k|\cdot, \cdot, \boldsymbol{\theta}) &= \log \{(2\pi)^m |\boldsymbol{\Theta}_\epsilon|\} + (\boldsymbol{y}_k - \boldsymbol{\Lambda}_y \boldsymbol{\eta}_k)^T \boldsymbol{\Theta}_\epsilon^{-1} (\boldsymbol{y}_k - \boldsymbol{\Lambda}_y \boldsymbol{\eta}_k) \quad (66) \\ -2 \log \mathbf{Pr}(\boldsymbol{\eta}_k|\cdot, \boldsymbol{\theta}). &= \log \{(2\pi)^p |\boldsymbol{\Psi}|\} + (\boldsymbol{\eta}_k - \boldsymbol{B}\boldsymbol{\eta}_k + \boldsymbol{\Gamma}\boldsymbol{\xi}_k)^T \boldsymbol{\Psi}^{-1} (\boldsymbol{\eta}_k - \boldsymbol{B}\boldsymbol{\eta}_k - \boldsymbol{\Gamma}\boldsymbol{\xi}_k) \quad (67) \\ -2 \log \mathbf{Pr}(\boldsymbol{x}_k|\cdot, \boldsymbol{\theta}). &= \log \{(2\pi)^q |\boldsymbol{\Theta}_\delta|\} + (\boldsymbol{x}_k - \boldsymbol{\Lambda}_x \boldsymbol{\xi}_k)^T \boldsymbol{\Theta}_\delta^{-1} (\boldsymbol{x}_k - \boldsymbol{\Lambda}_x \boldsymbol{\xi}_k) \quad (68) \\ -2 \log \mathbf{Pr}(\boldsymbol{\xi}_k|\boldsymbol{\theta}). &= \log \{(2\pi)^n |\boldsymbol{\Phi}|\} + \boldsymbol{\xi}_k^T \boldsymbol{\Phi}^{-1} \boldsymbol{\xi}_k. \quad (69) \end{aligned}$$

While in the latter case, Equation (65) will be expressed as follows :

$$\begin{aligned} \log \mathbf{Pr}(\cdot, \cdot, \cdot, \cdot|\boldsymbol{\theta}) &= -\frac{1}{2}(p + q + m + n) \log(2\pi) - \frac{1}{2} \log |\boldsymbol{\Theta}_\epsilon| |\boldsymbol{\Theta}_\delta| |\boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{A}^T| |\boldsymbol{\Phi}| \\ &\quad - \frac{1}{2} \mathbf{Tr} \Big\{ \boldsymbol{y}_k \boldsymbol{y}_k^T \boldsymbol{\Theta}_\epsilon^{-1} - 2\boldsymbol{\eta}_k \boldsymbol{y}_k^T \boldsymbol{\Theta}_\epsilon^{-1} \boldsymbol{\Lambda}_y + \boldsymbol{x}_k \boldsymbol{x}_k^T \boldsymbol{\Theta}_\delta^{-1} \\ &\quad - 2\boldsymbol{\xi}_k \boldsymbol{x}_k^T \boldsymbol{\Theta}_\delta^{-1} \boldsymbol{\Lambda}_x + \boldsymbol{\eta}_k \boldsymbol{\eta}_k^T \left(\boldsymbol{\Lambda}_y^T \boldsymbol{\Theta}_\epsilon^{-1} \boldsymbol{\Lambda}_y + (\boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{A}^T)^{-1} \right) \\ &\quad - 2\boldsymbol{\xi}_k \boldsymbol{\eta}_k^T (\boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{A}^T)^{-1} \boldsymbol{A}\boldsymbol{\Gamma} \\ &\quad + \boldsymbol{\xi}_k \boldsymbol{\xi}_k^T \left(\boldsymbol{\Gamma}^T \boldsymbol{A}^T (\boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{A}^T)^{-1} \boldsymbol{A}\boldsymbol{\Gamma} + \boldsymbol{\Lambda}_x^T \boldsymbol{\Theta}_\delta^{-1} \boldsymbol{\Lambda}_x + \boldsymbol{\Phi} \right) \Big\}. \quad (70) \end{aligned}$$

Let $\hat{\boldsymbol{\theta}}^t$ be the estimated model parametric vector, the posterior estimation of latent vectors $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$, and their related covariance matrices are given in the following.

$$\hat{\boldsymbol{\Sigma}}_{\eta y}^t = \frac{1}{N} \sum_{k=1}^N E \left[\boldsymbol{\eta}_k | \boldsymbol{y}_k, \boldsymbol{x}_k, \hat{\boldsymbol{\theta}}^t \right] \boldsymbol{y}_k^T. \quad (71)$$

$$\hat{\boldsymbol{\Sigma}}_{\xi x}^t = \frac{1}{N} \sum_{k=1}^N E \left[\boldsymbol{\xi}_k | \boldsymbol{y}_k, \boldsymbol{x}_k, \hat{\boldsymbol{\theta}}^t \right] \boldsymbol{x}_k^T. \quad (72)$$

$$\hat{\boldsymbol{\Sigma}}_{\eta\eta}^t = \frac{1}{N} \sum_{k=1}^N E \left[\boldsymbol{\eta}_k \boldsymbol{\eta}_k^T | \boldsymbol{y}_k, \boldsymbol{x}_k, \hat{\boldsymbol{\theta}}^t \right]. \quad (73)$$

$$\hat{\boldsymbol{\Sigma}}_{\xi\xi}^t = \frac{1}{N} \sum_{k=1}^N E \left[\boldsymbol{\xi}_k \boldsymbol{\xi}_k^T | \boldsymbol{y}_k, \boldsymbol{x}_k, \hat{\boldsymbol{\theta}}^t \right]. \quad (74)$$

$$\hat{\boldsymbol{\Sigma}}_{\xi\eta}^t = \frac{1}{N} \sum_{k=1}^N E \left[\boldsymbol{\xi}_k \boldsymbol{\eta}_k^T | \boldsymbol{y}_k, \boldsymbol{x}_k, \hat{\boldsymbol{\theta}}^t \right]. \quad (75)$$

The Q function can thus be expressed as follows :

$$\begin{aligned} Q(\boldsymbol{\theta}|\mathcal{D},\hat{\boldsymbol{\theta}}^t) &= -\frac{N}{2}(p+q+m+n)\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Theta}_\epsilon||\boldsymbol{\Theta}_\delta||\boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{A}^T||\boldsymbol{\Phi}| \\ &\quad - \frac{N}{2}\mathbf{Tr}\Big\{ \boldsymbol{\Sigma}_{yy}\boldsymbol{\Theta}_\epsilon^{-1} - 2\hat{\boldsymbol{\Sigma}}^t_{\eta y}\boldsymbol{\Theta}_\epsilon^{-1}\boldsymbol{\Lambda}_y + \boldsymbol{\Sigma}_{xx}\boldsymbol{\Theta}_\delta^{-1} - 2\hat{\boldsymbol{\Sigma}}^t_{\xi x}\boldsymbol{\Theta}_\delta^{-1}\boldsymbol{\Lambda}_x \\ &\quad + \hat{\boldsymbol{\Sigma}}^t_{\eta\eta}\left(\boldsymbol{\Lambda}_y^T\boldsymbol{\Theta}_\epsilon^{-1}\boldsymbol{\Lambda}_y + (\boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{A}^T)^{-1}\right) - 2\hat{\boldsymbol{\Sigma}}^t_{\xi\eta}\boldsymbol{A}^{-T}\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma} \\ &\quad + \hat{\boldsymbol{\Sigma}}^t_{\xi\xi}\left(\boldsymbol{\Gamma}^T\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{A}^T)^{-1}\boldsymbol{A}\boldsymbol{\Gamma} + \boldsymbol{\Lambda}_x^T\boldsymbol{\Theta}_\delta^{-1}\boldsymbol{\Lambda}_x + \boldsymbol{\Phi}\right)\Big\}. \end{aligned} \tag{76}$$

Maximizing the above equation is thus equivalent to minimizing the following $\mathcal{L}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^t)$ function.

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^t) &= \log|\boldsymbol{\Theta}_\epsilon||\boldsymbol{\Theta}_\delta||\boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{A}^T||\boldsymbol{\Phi}| \\ &\quad + \mathbf{Tr}\Big\{ \boldsymbol{\Sigma}_{yy}\boldsymbol{\Theta}_\epsilon^{-1} - 2\hat{\boldsymbol{\Sigma}}^t_{\eta y}\boldsymbol{\Theta}_\epsilon^{-1}\boldsymbol{\Lambda}_y + \boldsymbol{\Sigma}_{xx}\boldsymbol{\Theta}_\delta^{-1} - 2\hat{\boldsymbol{\Sigma}}^t_{\xi x}\boldsymbol{\Theta}_\delta^{-1}\boldsymbol{\Lambda}_x \\ &\quad + \hat{\boldsymbol{\Sigma}}^t_{\eta\eta}\left(\boldsymbol{\Lambda}_y^T\boldsymbol{\Theta}_\epsilon^{-1}\boldsymbol{\Lambda}_y + (\boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{A}^T)^{-1}\right) - 2\hat{\boldsymbol{\Sigma}}^t_{\xi\eta}\boldsymbol{A}^{-T}\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma} \\ &\quad + \hat{\boldsymbol{\Sigma}}^t_{\xi\xi}\left(\boldsymbol{\Gamma}^T\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{A}^T)^{-1}\boldsymbol{A}\boldsymbol{\Gamma} + \boldsymbol{\Lambda}_x^T\boldsymbol{\Theta}_\delta^{-1}\boldsymbol{\Lambda}_x + \boldsymbol{\Phi}\right)\Big\}. \end{aligned} \tag{77}$$

As a result, the EM algorithm for obtaining the MLE estimate of $\boldsymbol{\theta}$ can be accomplished by the following iterative steps until converge.

E-Step : Evaluate $\hat{\boldsymbol{\Sigma}}^t_{\eta y}$, $\hat{\boldsymbol{\Sigma}}^t_{\xi x}$, $\hat{\boldsymbol{\Sigma}}^t_{\xi\eta}$, $\hat{\boldsymbol{\Sigma}}^t_{\xi\xi}$ and $\hat{\boldsymbol{\Sigma}}^t_{\eta\eta}$.

M-Step : Obtain $\hat{\boldsymbol{\theta}}^{t+1}$ which is

$$\hat{\boldsymbol{\theta}}^{t+1} = \arg\min_{\hat{\boldsymbol{\theta}}}\left\{\mathcal{L}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^t)\right\}.$$

After the M-Step, set $t = t+1$ and then goto E-Step again.

To derive analytical equations for the estimation of the conditional expectation of the latent vectors in the E-Step and their covariance matrices, we let

$$\boldsymbol{w}_1 = \begin{bmatrix} \boldsymbol{\eta} \\ \boldsymbol{\xi} \end{bmatrix} \quad \text{and} \quad \boldsymbol{w}_2 = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix}.$$

Besides, we also let

$$\hat{\boldsymbol{\Sigma}}^t_H = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}^t_{\eta\eta} & \hat{\boldsymbol{\Sigma}}^t_{\eta\xi} \\ \hat{\boldsymbol{\Sigma}}^t_{\xi\eta} & \hat{\boldsymbol{\Sigma}}^t_{\xi\xi} \end{bmatrix}, \quad \hat{\boldsymbol{\Lambda}}^t = \begin{bmatrix} \hat{\boldsymbol{\Lambda}}^t_y & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Lambda}}^t_x \end{bmatrix} \quad \text{and} \quad \hat{\boldsymbol{\Theta}}^t = \begin{bmatrix} \hat{\boldsymbol{\Theta}}^t_\epsilon & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Theta}}^t_\delta \end{bmatrix}.$$

For $\hat{\boldsymbol{\Sigma}}_H^t$, one can readily obtain that

$$\hat{\boldsymbol{\Sigma}}_H^t = \begin{bmatrix} \hat{\boldsymbol{A}}^t\hat{\boldsymbol{\Gamma}}^t\hat{\boldsymbol{\Phi}}^t(\hat{\boldsymbol{A}}^t\hat{\boldsymbol{\Gamma}}^t)^T + \hat{\boldsymbol{A}}^t\hat{\boldsymbol{\Psi}}^t(\hat{\boldsymbol{A}}^t)^T & \hat{\boldsymbol{A}}^t\hat{\boldsymbol{\Gamma}}^t\hat{\boldsymbol{\Phi}}^t \\ \hat{\boldsymbol{\Phi}}^t(\hat{\boldsymbol{A}}^t\hat{\boldsymbol{\Gamma}}^t)^T & \hat{\boldsymbol{\Phi}}^t \end{bmatrix}. \tag{78}$$

Thus, the covariance matrix for the $(\boldsymbol{w}_1^T, \boldsymbol{w}_2^T)^T$ can thus be obtained.

$$\mathrm{Cov}\left(\begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \end{bmatrix}\right) = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_H^t & \hat{\boldsymbol{\Sigma}}_H^t(\hat{\boldsymbol{\Lambda}}^t)^T \\ \hat{\boldsymbol{\Lambda}}^t\hat{\boldsymbol{\Sigma}}_H^t & \hat{\boldsymbol{\Lambda}}^t\hat{\boldsymbol{\Sigma}}_H^t(\hat{\boldsymbol{\Lambda}}^t)^T + \hat{\boldsymbol{\Theta}} \end{bmatrix}. \tag{79}$$

The conditional expectation of $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ can be obtained by the following equations.

$$\begin{bmatrix} \hat{\boldsymbol{\eta}}_k^t \\ \hat{\boldsymbol{\xi}}_k^t \end{bmatrix} = \hat{\boldsymbol{\Sigma}}_H^t(\hat{\boldsymbol{\Lambda}}^t)^T(\hat{\boldsymbol{\Lambda}}^t\hat{\boldsymbol{\Sigma}}_H^t(\hat{\boldsymbol{\Lambda}}^t)^T + \hat{\boldsymbol{\Theta}}^t)^{-1} \begin{bmatrix} \boldsymbol{y}_k \\ \boldsymbol{x}_k \end{bmatrix}. \tag{80}$$

In sequel, $\hat{\boldsymbol{\Sigma}}_{\eta y}^t$ and $\hat{\boldsymbol{\Sigma}}_{\xi x}^t$ can be obtained by the following equations.

$$\hat{\boldsymbol{\Sigma}}_{\eta y}^t = \frac{1}{N}\sum_{k=1}^{N} \hat{\boldsymbol{\eta}}_k^t \boldsymbol{y}_k^T. \tag{81}$$

$$\hat{\boldsymbol{\Sigma}}_{\xi x}^t = \frac{1}{N}\sum_{k=1}^{N} \hat{\boldsymbol{\xi}}_k^t \boldsymbol{x}_k^T.. \tag{82}$$

The conditional covariance matrix for $\boldsymbol{w}_1$ can thus be obtained by the following equations.

$$\begin{aligned} \mathrm{Cov}\left(\boldsymbol{w}_1\boldsymbol{w}_1^T|\mathcal{D}, \boldsymbol{\theta}^t\right) &= \hat{\boldsymbol{\Sigma}}_H^t - \hat{\boldsymbol{\Sigma}}_H^t(\hat{\boldsymbol{\Lambda}}^t)^T\left(\hat{\boldsymbol{\Lambda}}^t\hat{\boldsymbol{\Sigma}}_H^t(\hat{\boldsymbol{\Lambda}}^t)^T + \hat{\boldsymbol{\Theta}}\right)^{-1}\hat{\boldsymbol{\Lambda}}^t\hat{\boldsymbol{\Sigma}}_H^t \\ &\quad + \frac{1}{N}\sum_{k=1}^{N}\begin{bmatrix} \hat{\boldsymbol{\eta}}_k^t \\ \hat{\boldsymbol{\xi}}_k^t \end{bmatrix}\begin{bmatrix} \hat{\boldsymbol{\eta}}_k^t \\ \hat{\boldsymbol{\xi}}_k^t \end{bmatrix}^T, \end{aligned} \tag{83}$$

Equation (78), (80), (81), (82) and (83) are applied to the evaluation of $\hat{\boldsymbol{\Sigma}}_{\eta y}^t$, $\hat{\boldsymbol{\Sigma}}_{\xi x}^t$, $\hat{\boldsymbol{\Sigma}}_{\xi\eta}^t$, $\hat{\boldsymbol{\Sigma}}_{\xi\xi}^t$ and $\hat{\boldsymbol{\Sigma}}_{\eta\eta}^t$ in the E-Step.

For the M-Step, we need to evaluate the matrices $\boldsymbol{\Theta}_\epsilon$, $\boldsymbol{\Theta}_\delta$, $\boldsymbol{\Psi}$, $\boldsymbol{\Phi}$, $\boldsymbol{B}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Lambda}_x$ and $\boldsymbol{\Lambda}_y$ that minimize the value of $\mathcal{L}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^t)$. By convention, we set $\boldsymbol{\Phi}$ to be an $(m \times m)$ identity matrix, i.e.

$$\boldsymbol{\Phi} = \boldsymbol{I}_{(m\times m)}. \tag{84}$$

To obtain $\boldsymbol{\Lambda}_y$ $\boldsymbol{\Theta}_\epsilon$, $\boldsymbol{\Lambda}_x$ and $\boldsymbol{\Theta}_\delta$, the same approach as for FA model is applied. Using the technique in matrix derivative (see Appendix) and setting $\partial\mathcal{L}/\partial X$ be zero matrix, it can

obtain that

$$\hat{\mathbf{\Lambda}}_y^{t+1} = (\hat{\mathbf{\Sigma}}_{\eta y}^t)^T(\hat{\mathbf{\Sigma}}_{\eta\eta}^t)^{-1}. \quad (85)$$

$$\hat{\mathbf{\Theta}}_\epsilon^{t+1} = \mathbf{diag}\left\{\hat{\mathbf{\Sigma}}_{yy}^t - (\hat{\mathbf{\Sigma}}_{\eta y}^t)^T(\hat{\mathbf{\Sigma}}_{\eta\eta}^t)^{-1}\hat{\mathbf{\Sigma}}_{\eta y}^t\right\}. \quad (86)$$

$$\hat{\mathbf{\Lambda}}_x^{t+1} = (\hat{\mathbf{\Sigma}}_{\xi x}^t)^T(\hat{\mathbf{\Sigma}}_{\xi\xi}^t)^{-1}. \quad (87)$$

$$\hat{\mathbf{\Theta}}_\delta^{t+1} = \mathbf{diag}\left\{\hat{\mathbf{\Sigma}}_{xx}^t - (\hat{\mathbf{\Sigma}}_{\xi x}^t)^T(\hat{\mathbf{\Sigma}}_{\xi\xi}^t)^{-1}\hat{\mathbf{\Sigma}}_{\xi x}^t\right\}. \quad (88)$$

For the factor loading matrices, if there are elements that are fixed to a constant, the method described in Equation (61) and (62) will be needed. Note that the negative log probability of $\boldsymbol{\eta}_k$ given $\boldsymbol{\xi}_k$ is also given as follows :

$$-2\log\mathbf{Pr}(\boldsymbol{\eta}_k|\boldsymbol{\xi}_k,\cdot) = m\log(2\pi) + \log|\mathbf{\Psi}| + \sum_{i=1}^m(\eta_i - \sum_{r\neq i}\beta_{ir}\eta_i - \sum_s \gamma_{is}\xi_s)^2/\Psi_{ii}.$$

This is the only likelihood term depended on $\boldsymbol{B}$ and $\mathbf{\Gamma}$. Let

$$e_i = \left(\eta_i - \sum_{r\neq i}\beta_{ir}\eta_r - \sum_s \gamma_{is}\xi_s\right)^2/\Psi_{ii}.$$

It is clear that

$$\frac{1}{N}\sum_{k=1}^N\sum_{i=1}^m\langle e_i(k)\rangle = \mathbf{Tr}\left\{\hat{\mathbf{\Sigma}}_{\eta\eta}^t(\boldsymbol{A}\mathbf{\Psi}\boldsymbol{A}^T)^{-1} - 2\hat{\mathbf{\Sigma}}_{\xi\eta}^t\boldsymbol{A}^{-T}\mathbf{\Psi}^{-1}\mathbf{\Gamma} + \hat{\mathbf{\Sigma}}_{\xi\xi}^t\left(\mathbf{\Gamma}^T\boldsymbol{A}^T(\boldsymbol{A}\mathbf{\Psi}\boldsymbol{A}^T)^{-1}\boldsymbol{A}\mathbf{\Gamma}\right)\right\},$$

where $\langle e_i(k)\rangle$ is the conditional expectation of $e_i(k)$ given $\boldsymbol{y}_k$, $\boldsymbol{x}_k$ and $\hat{\boldsymbol{\theta}}^t$.

Similar to the situation in confirmatory FA model,

$$\langle e_i\rangle = \langle\eta_i^2\rangle + \sum_{r\neq i}\beta_{ir}^2\langle\eta_r^2\rangle + \sum_s\gamma_{is}^2\langle\xi_s^2\rangle - 2\sum_{r\neq i}\beta_{ir}\langle\eta_i\eta_r\rangle - 2\gamma_{is}\langle\eta_i\xi_s\rangle + 2\sum_{r\neq i}\sum_s\beta_{ir}\gamma_{is}\langle\eta_r\xi_s\rangle.$$

Taking the derivative of the above equation with respect to all β_{ir} $(r\neq i)$ and γ_{is}, one will obtain $(m+n-1)$ linear equations for solving $(m+n-1)$ variables. Hence, the matrix $\hat{\boldsymbol{B}}^{t+1}$ and $\hat{\mathbf{\Gamma}}^{t+1}$ can be determined accordingly.

Taking the matrix derivative of $\mathcal{L}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^t)$ with respect to Ψ, it is readily shown that

$$\mathbf{\Psi} = \mathbf{diag}\left\{\boldsymbol{A}\hat{\mathbf{\Sigma}}_{\eta\eta}^t\boldsymbol{A}^T - \boldsymbol{A}\hat{\mathbf{\Sigma}}_{\xi\eta}^t\mathbf{\Gamma}^T - \mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}_{\eta\xi}^t\boldsymbol{A}^T + \mathbf{\Gamma}\hat{\mathbf{\Sigma}}_{\xi\xi}^t\mathbf{\Gamma}^T\right\}. \quad (89)$$

Putting the matrices $\boldsymbol{B}$ and $\mathbf{\Gamma}$ obtained in the previous step, $\hat{\mathbf{\Psi}}^{t+1}$ can thus be obtained. A sample Matlab code for this EM algorithm is added in the Appendix for reference.

6 Predictive Inference

Once the true (or the estimated model) parameters have been given (obtained), the prediction of $\boldsymbol{y}$ and $\boldsymbol{x}$ can be found.

6.1 Prediction of $\boldsymbol{y}$ given $\boldsymbol{x}$

Since the joint probability of $(\boldsymbol{y}, \boldsymbol{x})$ is Normal distribution with mean vector equals to null, the conditional probability of $\boldsymbol{y}$ given $\boldsymbol{x}$ is also a Normal distribution. The *posterior* estimation of $\boldsymbol{y}$ given $\boldsymbol{x}$ (denoted by $\hat{\boldsymbol{y}}$) will be equal to the mode of the conditional distribution. That is,

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} \{\mathbf{Pr}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))\}. \tag{90}$$

In Normal distribution, the mean and mode are in the same location. So,

$$\begin{aligned} \hat{\boldsymbol{y}} &= \boldsymbol{\Sigma}_{\boldsymbol{yx}} \boldsymbol{\Sigma}_{\boldsymbol{xx}}^{-1} \boldsymbol{x}. & (91) \\ &= \boldsymbol{\Lambda}_{\boldsymbol{y}} (\boldsymbol{I} - \boldsymbol{B})^{-1} \boldsymbol{\Phi} \boldsymbol{\Lambda}_{\boldsymbol{x}}^T \left(\boldsymbol{\Lambda}_{\boldsymbol{x}} \boldsymbol{\Phi} \boldsymbol{\Lambda}_{\boldsymbol{x}}^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}} \right)^{-1} \boldsymbol{x}. & (92) \end{aligned}$$

Please refer to the Appendix for the equations as well.

It should be noted that the prediction $\hat{\boldsymbol{y}}$ is valid only for $\mathbf{Pr}(\cdot|\cdot)$ is Normal distribution. For other distribution, the maximum *a posterior* estimation of $\hat{\boldsymbol{y}}$ might not be equal to the mean vector of the conditional probability. In such case, other techniques will be needed for the prediction.

6.2 Prediction of factor scores $(\boldsymbol{\eta}, \boldsymbol{\xi})$

Let $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{x}}$ be the *posterior* estimation (i.e. prediction) of $\boldsymbol{y}$ and $\boldsymbol{x}$. Moreover, we let

$$\boldsymbol{w}_1 = \begin{bmatrix} \boldsymbol{\eta} \\ \boldsymbol{\xi} \end{bmatrix} \text{ and } \boldsymbol{w}_2 = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix}.$$

As the mean vectors of $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are null, the prediction can then be conducted by the following equation.

$$\hat{\boldsymbol{w}}_1 = \arg\max_{\boldsymbol{w}_1} \{\mathbf{Pr}(\boldsymbol{w}_1|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))\}. \tag{93}$$

By the same technique being applied for the prediction of $\boldsymbol{y}$,

$$\hat{\boldsymbol{w}}_1 = \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \boldsymbol{w}_2, \tag{94}$$

where

$$\boldsymbol{\Sigma}_{12} = \begin{bmatrix} (\boldsymbol{I} - \boldsymbol{B})^{-1} \left(\boldsymbol{\Gamma} \boldsymbol{\Phi} \boldsymbol{\Gamma}^T + \boldsymbol{\Psi} \right) (\boldsymbol{I} - \boldsymbol{B})^{-T} \boldsymbol{\Lambda}_{\boldsymbol{y}}^T & (\boldsymbol{I} - \boldsymbol{B})^{-1} \boldsymbol{\Gamma} \boldsymbol{\Phi} \boldsymbol{\Lambda}_{\boldsymbol{x}}^T \\ \boldsymbol{\Phi} \boldsymbol{\Gamma}^T (\boldsymbol{I} - \boldsymbol{B})^{-T} \boldsymbol{\Lambda}_{\boldsymbol{y}}^T & \boldsymbol{\Phi} \boldsymbol{\Lambda}_{\boldsymbol{x}}^T \end{bmatrix}$$

and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is given by Equation (27).

Table 1: List of assessment indices.

Index	Defnition
GFI_{ML}	$1 - \mathbf{Tr}\{(\boldsymbol{\Sigma}^{-1}\boldsymbol{S} - \boldsymbol{I})^2\}/\mathbf{Tr}\{(\boldsymbol{\Sigma}^{-1}\boldsymbol{S})^2\}$
GFI_{LS}	$1 - \mathbf{Tr}\{(\boldsymbol{S} - \boldsymbol{\Sigma})^2\}/\mathbf{Tr}\{\boldsymbol{S}^2\}$
AGFI	$1 - \frac{d'}{d}(1 - \text{GFI})$
RMR	$\sqrt{\sum_{i=1}^{(p+q)}\sum_{j=1}^{i}(\boldsymbol{S}_{ij} - \boldsymbol{\Sigma}_{ij})^2 / d'}$
LL	$-\frac{N(p+q)}{2}\log(2\pi) - \frac{N}{2}\log\lvert\boldsymbol{\Sigma}(\boldsymbol{\theta})\rvert - \frac{N}{2}\mathbf{Tr}\{\boldsymbol{S}\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\}$
LLR	$-\frac{N}{2}\{\log\lvert\boldsymbol{\Sigma}(\boldsymbol{\theta})\rvert - \log\lvert\boldsymbol{S}\rvert\} - \frac{N}{2}\mathbf{Tr}\{\boldsymbol{S}\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\} + \frac{N(p+q)}{2}$
LPO	$\log\mathbf{Pr}(\mathcal{D}\lvert\boldsymbol{\theta}_1) - \log\mathbf{Pr}(\mathcal{D}\lvert\boldsymbol{\theta}_2) + \log\mathbf{Pr}(\boldsymbol{\theta}_1) - \log\mathbf{Pr}(\boldsymbol{\theta}_2)$
PE(train)	$\sum_k (\boldsymbol{y}_k - \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{x}_k)^2 / \lvert\mathcal{D}_{\text{train}}\rvert$
PE(test)	$\sum_l (\boldsymbol{y'}_l - \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{x'}_l)^2 / \lvert\mathcal{D}_{\text{test}}\rvert$

7 Model Assessment

Once a model $\hat{\boldsymbol{\theta}}$ has been obtained, the next step is to evaluate its *goodness of fit*. As the true underlying model is an unknown, we can make comparison only with the observable dataset $\mathcal{D}\{(\boldsymbol{y}_k, \boldsymbol{x}_k)_{k=1}^N\}$. Two pieces of information available are the *sample mean* that is assumed to be a null vector, and the *sample covariance* matrix,

$$\boldsymbol{S} = \begin{bmatrix} \frac{1}{N}\sum_{k=1}^N \boldsymbol{y}_k\boldsymbol{y}_k^T & \frac{1}{N}\sum_{k=1}^N \boldsymbol{y}_k\boldsymbol{x}_k^T \\ \frac{1}{N}\sum_{k=1}^N \boldsymbol{x}_k\boldsymbol{y}_k^T & \frac{1}{N}\sum_{k=1}^N \boldsymbol{x}_k\boldsymbol{x}_k^T \end{bmatrix}. \tag{95}$$

Before proceed, one should understand that the purposes of an objective function and the fit-index are very different. The former is used for the search an estimated model which is what we call *an optimal solution* in the previous section. The latter is used for assessing the *goodness* of that estimated model. In many occasions, one would like to use an objective function that has direct connection to the fit-index. For examples, F_{LS} and F_{LR} are two of them which have such direct links. While the objective functions, F_{AP} will have no such linkage. As a preview, a list of common assessment measures are depicted in Table 1.

7.1 Covariance matrix-based

To compare the fitness based on the covariance, Goodness-of-fit index (GFI), the adjusted goodness-of-fit index (AGFI) and the root mean square residual (RMR) are three

common measures.

$$\begin{aligned}
\mathrm{GFI}_{ML} &= 1 - \frac{\mathbf{Tr}\{(\boldsymbol{\Sigma}^{-1}\boldsymbol{S} - \boldsymbol{I})^2\}}{\mathbf{Tr}\{(\boldsymbol{\Sigma}^{-1}\boldsymbol{S})^2\}}. && (96)\\
\mathrm{GFI}_{LS} &= 1 - \frac{\mathbf{Tr}\{(\boldsymbol{S} - \boldsymbol{\Sigma})^2\}}{\mathbf{Tr}\{\boldsymbol{S}^2\}}. && (97)\\
\mathrm{AGFI} &= 1 - \frac{d'}{d}\left(1 - \mathrm{GFI}\right). && (98)\\
\mathrm{RMR} &= \sqrt{\frac{1}{d'}\sum_{i=1}^{(p+q)}\sum_{j=1}^{i}(\boldsymbol{S}_{ij} - \boldsymbol{\Sigma}_{ij})^2}. && (99)
\end{aligned}$$

Here d' is the total number of elements in the lower (or upper) triangle of the matrix $\boldsymbol{S}$. While d is the total number of free parameters in the model $\boldsymbol{\theta}$, i.e.

$$\begin{aligned}
d' &= \frac{(\dim\{(\boldsymbol{y},\boldsymbol{x})\})(\dim\{(\boldsymbol{y},\boldsymbol{x})\}+1)}{2}\\
&= \frac{(p+q)(p+q+1)}{2}.\\
d &= \dim\{\boldsymbol{\theta}\}.
\end{aligned}$$

7.2 Likelihood-based

Another approach to assess the goodness of a model or compare different models is based on the *log likelihood* or the *log likelihood ratio*, which have been defined in Equation (35) and Equation (30).

$$\begin{aligned}
\mathcal{L}(\mathcal{D}|\boldsymbol{\theta}) &= -\frac{N(p+q)}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \frac{N}{2}\mathbf{Tr}\left\{\boldsymbol{S}\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\right\}.\\
\mathcal{L}(\boldsymbol{\theta},\boldsymbol{S}|\mathcal{D}) &= -\frac{N}{2}\left\{\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \log|\boldsymbol{S}|\right\} - \frac{N}{2}\mathbf{Tr}\left\{\boldsymbol{S}\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\right\} + \frac{N(p+q)}{2}.
\end{aligned}$$

To explain their differences, let us define $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are two models to be compared. The true underlying model is with covariance matrix $\boldsymbol{S}_0$. The first likelihood tells us how likely (in probability sense) does a model $\boldsymbol{\theta}$ generate the data set $\mathcal{D}$. In other words, *what is the probability that $\mathcal{D}$ is generated if model $\boldsymbol{\theta}$ is the true model*. Therefore, we prefer $\boldsymbol{\theta}_1$ to $\boldsymbol{\theta}_2$ if

$$\mathcal{L}(\mathcal{D}|\boldsymbol{\theta}_1) > \mathcal{L}(\mathcal{D}|\boldsymbol{\theta}_2).$$

Without loss of generality, we replace $\boldsymbol{\theta}_1$ by $\boldsymbol{\Sigma}(\boldsymbol{\theta}_1)$ and $\boldsymbol{\theta}_2$ by $\boldsymbol{S}$. $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{S}|\mathcal{D})$ tells us the difference between how likely does a model $\boldsymbol{\Sigma}(\boldsymbol{\theta}_1)$ generate the data set $\mathcal{D}$ and how likely does a model $\boldsymbol{S}$ generate the data set $\mathcal{D}$. Moreover, it is assumed that the true model is specified by $\boldsymbol{\Sigma}$, i.e. $\boldsymbol{S} = \boldsymbol{S}_0$. We then *accept* $\boldsymbol{\theta}_1$ to $\boldsymbol{S}$ if $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{S}|\mathcal{D}) > 0$.

7.3 Bayesian decision-based

Likelihood-based assessment assumes no *a prior* information about the models to be compared. The only information we have is $\mathcal{D}$. For N is large, we usually believe that the information from $\mathcal{D}$ is rich for us to obtain a good model. How about if (1) the sample size is small, (2) the accepted model is very complex in compared with the rejected model, and (3) the parametric values of an accepted model are of large magnitudes ?

Example 1 : Let us have a simple example to show that idea. Suppose we have to estimate the parametric values for two models, namely $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, which correspond to two different model structures. Before estimation starts, we would need to ask ourself how much information we know about the model. If we do not have any information, we could simply assume that these two structures are equally likely. After parametric estimation, we evaluate that

$$\mathbf{Pr}(\mathcal{D}|\boldsymbol{\theta}_1) = 0.3. \quad \text{and} \quad \mathbf{Pr}(\mathcal{D}|\boldsymbol{\theta}_2) = 0.2.$$

By comparing their likelihoods (or by likelihood ratio test), one can conclude that $\boldsymbol{\theta}_1$ is better than $\boldsymbol{\theta}_2$.

Example 2 : Certainly, analyst normally will a *belief* on the structure of a model. This belief can be built upon an extensive literature survey, or simply a gut feeling. In any case, this belief turns out to be a weighting factor affecting the judgement how good a model is. In Bayesian statistics, this belief is called *a prior* probability. With the same results and assuming that the *a prior* probabilities for the models are 0.3 and 0.7. The probabilities (formerly called *posterior* probabilities) will be given by

$$\begin{aligned}
\mathbf{Pr}(\boldsymbol{\theta}_1|\mathcal{D}) &= \frac{\mathbf{Pr}(\mathcal{D}|\boldsymbol{\theta}_1) \times \mathbf{Pr}(\boldsymbol{\theta}_1)}{\mathbf{Pr}(\mathcal{D})} = \frac{0.3 \times 0.3}{\mathbf{Pr}(\mathcal{D})}, \\
\mathbf{Pr}(\boldsymbol{\theta}_2|\mathcal{D}) &= \frac{\mathbf{Pr}(\mathcal{D}|\boldsymbol{\theta}_2) \times \mathbf{Pr}(\boldsymbol{\theta}_2)}{\mathbf{Pr}(\mathcal{D})} = \frac{0.2 \times 0.7}{\mathbf{Pr}(\mathcal{D})}.
\end{aligned}$$

As the factor $\mathbf{Pr}(\mathcal{D})$ is common to both, $\boldsymbol{\theta}_2$ will be preferred to $\boldsymbol{\theta}_1$.

Therefore, the comparison between two models can be made by considering the following ratio.

$$\underbrace{\frac{\mathbf{Pr}(\boldsymbol{\theta}_1|\mathcal{D})}{\mathbf{Pr}(\boldsymbol{\theta}_2|\mathcal{D})}}_{\text{Posterior Odds}} = \underbrace{\frac{\mathbf{Pr}(\mathcal{D}|\boldsymbol{\theta}_1)}{\mathbf{Pr}(\mathcal{D}|\boldsymbol{\theta}_2)}}_{\text{Bayes Factor}} \times \underbrace{\frac{\mathbf{Pr}(\boldsymbol{\theta}_1)}{\mathbf{Pr}(\boldsymbol{\theta}_2)}}_{\text{Prior Odds}} . \tag{100}$$

Taking log on both sides, above equation can be rewritten as follows :

$$\begin{aligned} \mathrm{LPO}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= \log \mathbf{Pr}(\mathcal{D}|\boldsymbol{\theta}_1) - \log \mathbf{Pr}(\mathcal{D}|\boldsymbol{\theta}_2) \\ &+ \log \mathbf{Pr}(\boldsymbol{\theta}_1) - \log \mathbf{Pr}(\boldsymbol{\theta}_2). \end{aligned} \tag{101}$$

Thus, model $\boldsymbol{\theta}_1$ is preferred to model $\boldsymbol{\theta}_2$ if $\mathrm{LPO}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) > 0$.

7.4 Prediction error-based

Prediction error is a popular technique being used in assessing a nonlinear model and it has a tight relationship with another assessment method called cross-validation. Prediction error is defined as the square error between the predicted $\hat{\boldsymbol{y}}$ and the actual $\boldsymbol{y}$. Here $\hat{\boldsymbol{y}}$ is under the condition that the input $\boldsymbol{x}$ is given, Equation (92).

$$\hat{\boldsymbol{y}} = \boldsymbol{\Lambda_y}(\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\Phi}\boldsymbol{\Lambda_x}^T \left(\boldsymbol{\Lambda_x}\boldsymbol{\Phi}\boldsymbol{\Lambda_x}^T + \boldsymbol{\Theta_\delta}\right)^{-1} \boldsymbol{x}.$$

Let

$$\boldsymbol{A}(\boldsymbol{\theta}) = \boldsymbol{\Lambda_y}(\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\Phi}\boldsymbol{\Lambda_x}^T \left(\boldsymbol{\Lambda_x}\boldsymbol{\Phi}\boldsymbol{\Lambda_x}^T + \boldsymbol{\Theta_\delta}\right)^{-1}.$$

The relation between the prediction $\hat{\boldsymbol{y}}$ given input $\boldsymbol{x}$ is simply be a linear equation. Then, the prediction error of a model $\boldsymbol{\theta}$ can be defined as follows :

$$\mathrm{PE} = \frac{1}{N}\sum_{k=1}^{N} \left(\boldsymbol{y}_k - \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{x}_k\right)^2. \tag{102}$$

Clearly, this value can be very small if the model over-fits the sample data. A better criteria is based on cross-validation. Let us partition the data set $\mathcal{D}$ into two subsets, namely $\mathcal{D}_{\mathrm{train}}$ and $\mathcal{D}_{\mathrm{test}}$. Only the data in $\mathcal{D}_{\mathrm{train}}$ will take part in the estimation of the model parameters.

To differentiate the training data and the testing data, we denote $(\boldsymbol{y}_k, \boldsymbol{x}_k)$ be a data in $\mathcal{D}_{\mathrm{train}}$ and $(\boldsymbol{y'}_l, \boldsymbol{x'}_l)$ be a data in $\mathcal{D}_{\mathrm{test}}$. Then the prediction errors with respect to the training and testing data sets will be given by

$$\mathrm{PE}_{\mathrm{train}} = \frac{1}{|\mathcal{D}_{\mathrm{train}}|}\sum_{k} \left(\boldsymbol{y}_k - \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{x}_k\right)^2, \tag{103}$$

$$\mathrm{PE}_{\mathrm{test}} = \frac{1}{|\mathcal{D}_{\mathrm{test}}|}\sum_{l} \left(\boldsymbol{y'}_l - \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{x'}_l\right)^2, \tag{104}$$

where $|\mathcal{D}_{\mathrm{train}}|$ and $|\mathcal{D}_{\mathrm{test}}|$ correspond to the number of training data and the number of testing data respectively. The value $\mathrm{PE}_{\mathrm{test}}$ is also called the cross-validation error.

Prediction error and cross-validation error are not limited to be used for SEM. For other models that are not suitable for GFI assessment and log likelihood ratio test, prediction error can be applicable.

7.5 Parameter significance

To remove insignificant parameters is a way to reduce the complexity of a model. For SEM, the significant of a parameter can be determined by its value and its standard error (t-value in SAS programming) which is defined by the following formulae.

$$\tau_i = \hat{\boldsymbol{\theta}}_i / \sqrt{s_{ii}}, \tag{105}$$

where s_{ii} is the i^{th} diagonal element of the Hessian matrix of the negative log-likelihood function at the MLE $\hat{\boldsymbol{\theta}}$.

$$s_{ii} = -\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \mathcal{L}(\hat{\boldsymbol{\theta}}). \tag{106}$$

Note that this Hessian matrix is an approximation of the Information Matrix for MLEs.

A Useful Mathematics

A.1 Matrix

Let A, B and D are non-singular square matrix. Matrix Inversion Lemma is useful for derivation of conditional probability.

$$(D - AB^{-1}A^T)^{-1} = D^{-1} + D^{-1}A(B - A^T D^{-1}A)^{-1}A^T D^{-1}. \tag{107}$$

To compute the log of the determinant of a matrix, $\log(|\cdot|)$, the following equation can be employed.

$$\log|D + AB^{-1}A^T| = \log|D| - \log|B| + \log|B + A^T D^{-1}A|. \tag{108}$$

Let us denote

$$\left(\frac{\partial f(A)}{\partial A}\right)_{ij} = \frac{\partial f(A)}{\partial A_{ij}},$$

The following equations are useful for dealing with matrix derivative.

$$\frac{\partial}{\partial A}\log|A| = A^{-T}. \tag{109}$$

$$\frac{\partial}{\partial A}\mathbf{Tr}\{B^T A\} = B. \tag{110}$$

$$\frac{\partial}{\partial A}\mathbf{Tr}\{BA^T CA\} = 2CAB. \tag{111}$$

For X is not square and A is symmetric, then

$$\frac{\partial|X^T AX|}{\partial X} = 2|X^T AX|X^{-T}. \tag{112}$$

$$\frac{\partial}{\partial X}\mathrm{Tr}(AXB) = BA. \tag{113}$$

$$\frac{\partial}{\partial X}\mathrm{Tr}(X^T A) = A. \tag{114}$$

$$\frac{\partial}{\partial X}\mathrm{Tr}(X^T BX) = BX + B^T X. \tag{115}$$

$$\frac{\partial}{\partial X}\mathrm{Tr}(AXBX) = A^T X^T B^T + B^T X^T A^T. \tag{116}$$

$$\frac{\partial}{\partial X}\mathrm{Tr}(AXBX^T C) = A^T C^T XB^T + CAXB. \tag{117}$$

$$\frac{\partial}{\partial X}\mathrm{Tr}\left[(X^T CX)^{-1}A\right] = -CX(X^T CX)^{-1}(A + A^T)(X^T CX)^{-1}. \tag{118}$$

$$\frac{\partial}{\partial X}\mathrm{Tr}(AX^{-1}B) = -(X^{-1}BAX^{-1})^T. \tag{119}$$

Besides, we let B be a $(n \times n)$ diagonal matrix with positive elements $\lambda_1, \lambda_2, \ldots, \lambda_n$. A is a $(n \times n)$ square matrix defined as $(a_{ij})_{n \times n}$. Define a scalar function $f(B)$ as follows :

$$f(B) = \log |B| + \mathbf{Tr}\left\{AB^{-1}\right\}.$$

It is readily shown that

$$\frac{\partial}{\partial \lambda_i} f(\mathbf{diag}\{A\}) = 0, \tag{120}$$

for all $i = 1, 2, \ldots, n$.

A.2 Conditional Probability

Let $(\boldsymbol{w}_1^T, \boldsymbol{w}_2^T)^T$ be a random vector from a multi-dimension Normal distribution, i.e.

$$\left[\begin{array}{c} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \end{array}\right] \sim \mathcal{N}\left(\left[\begin{array}{c} \bar{\boldsymbol{w}}_1 \\ \bar{\boldsymbol{w}}_2 \end{array}\right], \left[\begin{array}{cc} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right]\right). \tag{121}$$

The conditional probability of $\boldsymbol{w}_1$ given $\boldsymbol{w}_2 = \boldsymbol{a}$ is also a Normal distribution $\mathcal{N}(\hat{\boldsymbol{w}}_1, \hat{\boldsymbol{\Sigma}}_{11})$, where

$$\hat{\boldsymbol{w}}_1 = \bar{\boldsymbol{w}}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{a} - \bar{\boldsymbol{w}}_2). \tag{122}$$

$$\hat{\boldsymbol{\Sigma}}_{11} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \tag{123}$$

By Matrix Inversion Lemma, the inversion of the covariance matrix $\hat{\boldsymbol{\Sigma}}_{11}$ can also be expressed as follows :

$$\hat{\boldsymbol{\Sigma}}_{11} = \boldsymbol{\Sigma}_{11}^{-1} + \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22}^{-1} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}.$$

B Matlab Codes for EM Algorithms

B.1 Maximum Likelihood Confirmatory FA

```
% ========================================================
% In this program, you need to assign the number for RUN
% ========================================================
QQ = -1;
while (QQ < 0)
    % ========================================================
    % Sample Data Generation
    %
    % Te = 0.25, Td = 1; Psi = 1, Phi = 1;
    % ========================================================
    %
    N = 200;
    LX0 = [[1 1 1]' zeros(3,2);
        zeros(3,1) [2 2 2]' zeros(3,1);
        zeros(3,2) [0.8 0.8 0.8]'];
    XI0 = randn(N, 3);
    X = XI0*LX0' + 0.5*randn(N, 9);
    SS = X'*X/N;
    q = 9;
    LiHmax =  (q*log(2*pi) + log(det(SS)) + q);
    % ========================================================

    % ========================================================
    % Model Structural Definition
    %
    % XX-Mask: Element '1' corresponds to a parameter
    % m: Number of latent variable \eta
    % n: Number of latent variable \xi
    % ========================================================
    %
    n = 3;
    LXMask = sign(abs(LX0));
    %
    % ========================================================

    % ========================================================
```

```
% Load Data File & Initialize X and Y
%
% The matrices SIG_Y_Y and SIG_X_X are defined for
% the M-Step.
% ========================================================
%
p = 9; q = 9;
Xbar = mean(X);
X = X - kron(Xbar,ones(N,1));
SIG_X_X = (X'*X)/N;
%
% ========================================================
% ========================================================
% Initialize Parametric Matrices
% ========================================================
%
Phi = eye(n);
Td = 0.1*eye(q);
LX = 0.01*abs(randn(q,n)).*LXMask;
%
% ========================================================
Q = zeros(RUN,1); LiH = zeros(RUN,1);
it = 1; QQ = 1;  u = 1;
while((QQ>0)&&(it<RUN+1))
    % ========================================================
    % E-Step
    % ========================================================
    %
    % Posterior Estimation of \xi
    XIH = X*inv(LX*LX'+Td)'*LX;

    % Posterior Estimation of \Sigma_xi_x
    PSIG_XI_X  = (XIH'*X)/N;

    % Posterior Estimation of the Covariance Matrix \Sigma_xi_xi
    PSIG = eye(3) - LX'*inv(LX*LX' + Td)*LX;
    PSIG_XI_XI = PSIG + XIH'*XIH/N;
    %
    % ========================================================
```

```
        % ========================================================
        % M-Step
        % ========================================================
        %
        for r=1:q,
            [Value Loc] = setdiff(LXMask(r,:).*[1:1:n], 0);
            Sig_tmp = PSIG_XI_XI(Loc,Loc);
            Y_tmp = PSIG_XI_X(Loc,r);
            LX(r,Loc) = Y_tmp'*inv(Sig_tmp)';
            Td(r,r) = X(:,r)'*X(:,r)/N - 2*X(:,r)'*(XIH*LX(r,:)')/N;
            Td(r,r) = Td(r,r) + LX(r,:)*PSIG_XI_XI*LX(r,:)';
        end
        %
        % ========================================================

        % ========================================================
        % Likelihood Calculation
        % ========================================================
        %
        COV_EST = LX*LX'+ Td;
        LiH(it)=q*log(2*pi)+log(det(COV_EST))+trace(SS*inv(COV_EST));
        LiH(it)=LiH(it)-LiHmax;

        Qtmp1 = (q+n)*log(2*pi) + log(det(Td));
        Qtmp2 = trace(SIG_X_X*inv(Td) -PSIG_XI_X'*LX'*inv(Td));
        Qtmp2 = Qtmp2 - trace(LX*PSIG_XI_X*inv(Td));
        Qtmp3 = trace(PSIG_XI_XI*LX'*inv(Td)*LX + PSIG_XI_XI);
        Q(it) = (Qtmp1 + Qtmp2 + Qtmp3) - LiHmax;
        QQ = Q(it);
        %
        % ========================================================
        it = it + 1;
    end
end
figure(1); semilogy([Q LiH]); legend('Q', 'LiH')
```

B.2 Maximum Likelihood SEM

```
% ========================================================
% Model Structural Definition
%
% XX-Mask: Element '1' corresponds to a parameter
% m: Number of latent variable \eta
% n: Number of latent variable \xi
% ========================================================
%
m = 3; n = 3;
BMask = [0 1 0; 0 0 0; 1 1 0];
TMask = [1 0 1; 0 1 1; 1 0 1];
LYMask = kron(eye(m), [1;1;1]);
LXMask = kron(eye(n), [1;1;1]);
%
% ========================================================

% ========================================================
% Load Data File & Initialize X and Y
%
% The matrices SIG_Y_Y and SIG_X_X are defined for
% the M-Step.
% ========================================================
%
load CPI_12_2008.txt;
PI = CPI_12_2008;
clear CPI_12_2008;

X_index = [2 4 5 12 13 15 8 10 11];
Y_index = [24 25 26 18 20 21 30 31 32];
X = PI(:,X_index); Y = PI(:,Y_index);
[N Q] = size(PI);
X = X - kron(mean(X),ones(N,1));
Y = Y - kron(mean(Y),ones(N,1));
p = length(Y_index); q = length(X_index);
SIG_Y_Y = (Y'*Y)/N;
SIG_X_X = (X'*X)/N;
%
% ========================================================
```

```
% =========================================================
% Initialize Parametric Matrices
% =========================================================
%
ETA = zeros(N, m); XI = zeros(N, n);
B = randn(m,m).*BMask; B00 = B;
T = randn(m,n).*TMask; T00 = T;
LY = randn(p,m).*LYMask; LY00 = LY;
LX = randn(q,n).*LXMask; LX00 = LX;
Psi = 0.1*eye(m);
Phi = eye(n);
Te = 0.1*eye(p);
Td = 0.1*eye(q);

%
% =========================================================

for run = 1:100,
    % ========================================================
    % E-Step
    % ========================================================
    %
    A = eye(m)-B;
    LAM = [LY zeros(p,n); zeros(q,m) LX];
    THETA = [Te zeros(p,p); zeros(q,q) Td];
    SIG_ETA_ETA = A*T*Phi*T'*A'+A*Psi*A';
    SIG_ETA_XI  = A*T*Phi;
    SIG_XI_ETA  = Phi*T'*A';
    SIG_XI_XI   = Phi;
    SIG_H = [SIG_ETA_ETA SIG_ETA_XI; SIG_XI_ETA SIG_XI_XI];

    % Posterior Estimation of \eta and \xi
    ETA_XI = [Y X]*inv(LAM*SIG_H*LAM'+THETA)'*LAM*SIG_H';
    ETAH = ETA_XI(:,[1 2 3]);
    XIH  = ETA_XI(:,[4 5 6]);

    % Posterior Estimation of \Sigma_eta_y and \Sigma_xi_x
    PSIG_ETA_Y = (ETAH'*Y)/N;
    PSIG_XI_X  = (XIH'*X)/N;
```

```
% Posterior Estimation of the Covariance Matrices
% \Sigma_eta_eta, \Sigma_eta_xi, \Sigma_xi_xi and \Sigma_xi_eta
PSIG = SIG_H - SIG_H*LAM'*inv(LAM*SIG_H*LAM' + THETA)*LAM*SIG_H';
PSIG = PSIG + ETA_XI'*ETA_XI/N;
PSIG_ETA_ETA = PSIG(1:m,1:m);
PSIG_ETA_XI  = PSIG(1:m,m+1:m+n);
PSIG_XI_ETA  = PSIG(m+1:m+n,1:m);
PSIG_XI_XI   = PSIG(m+1:m+n,m+1:m+n);

%
% ========================================================

% ========================================================
% M-Step
% ========================================================
%

% (For the Y Equation)
for r=1:p,
    [Value Loc] = setdiff(LYMask(r,:).*[1:1:m], 0);
    Sig_tmp = PSIG_ETA_ETA(Loc,Loc);
    Y_tmp = PSIG_ETA_Y(Loc,r);
    LY(r,Loc) = Y_tmp'*inv(Sig_tmp)';
end

% (For the X equation)
for r=1:q,
    [Value Loc] = setdiff(LXMask(r,:).*[1:1:n], 0);
    Sig_tmp = PSIG_XI_XI(Loc,Loc);
    Y_tmp = PSIG_XI_X(Loc,r);
    LX(r,Loc) = Y_tmp'*inv(Sig_tmp)';
end

% (For the latent ETA equation)
for r=1:m,
    [Value Loc_eta] = setdiff(BMask(r,:).*[1:1:m], 0);
    [Value Loc_xi] = setdiff(TMask(r,:).*[1:1:n], 0);
    Sig_11 = PSIG_ETA_ETA(Loc_eta, Loc_eta);
    Sig_12 = PSIG_ETA_XI(Loc_eta,Loc_xi);
```

```
        Sig_22 = PSIG_XI_XI(Loc_xi,Loc_xi);
        Sig_21 = Sig_12';
        Sig_tmp = [Sig_11 Sig_12; Sig_21 Sig_22];
        Y_tmp = [PSIG_ETA_ETA(Loc_eta,r); PSIG_XI_ETA(Loc_xi, r)];
        BT_tmp = Y_tmp'*inv(Sig_tmp)';
        le = length(Loc_eta);
        lx = length(Loc_xi);
        B(r,Loc_eta) = BT_tmp(1,1:le);
        T(r,Loc_xi) = BT_tmp(1,le+1:le+lx);
    end

    Te = diag(diag(LY*PSIG_ETA_Y));
    Td = diag(diag(LX*PSIG_XI_X));
    A = eye(m)-B;
    TMP = A*PSIG_ETA_ETA*A'-T*PSIG_XI_ETA*A'-A*PSIG_ETA_XI*T'+T*PSIG_XI_XI*T';
    Psi = diag(diag(TMP));

    %
    % =======================================================

end
```

References

[1] Anderson T.W., *An Introduction to Multivariate Statistical Analysis*, 2nd Ed. Wiley, 1984.

[2] Bentler P.M. and D.G. Weeks, Linear structural equations with latent variables, *Psychometrika*, Vol.45, 289-308, 1980.

[3] Joreskog K.G., Some contributions to maximum likelihood factor analysis, *Psychometrika*, Vol.32, 443-482, 1967.

[4] Joreskog K.G. and D. Sorbom, Recent developments in structural equation modeling, *Journal of Marketing Research*, Vol.19(4), 404-416, 1982.

[5] Lee S.Y. *Structural Equation Modeling: A Bayesian approach*, Wiley, 2007.

[6] Rao S.S., *Optimization Theory and Applications*, 2nd Edition, Wiley, 1984.