1

Introduction

The origins of structural equation modeling (SEM) stem from factor analysis (Spearman, 1904; Tucker, 1955) and path analysis (or simultaneous equations) (Wright,1918, 1921, 1934). By integrating the measurement (factor analysis) and structural (path analysis) approaches, a more generalized analytical framework is produced, called SEM (Jöreskog, 1967, 1969, 1973; Keesling, 1972; Wiley, 1973). In SEM, unobservable latent variables (constructs or factors) are estimated from observed indicator variables, and the focus is on estimation of the relations among the latent variables free of the influence of measurement errors (Jöreskog, 1973; Jöreskog and Sörbom, 1979; Bentler,1980, 1983; Bollen, 1989a).

SEM provides a mechanism for taking into account measurement error in the observed variables involved in a model. In social sciences, some constructs, such as intelligence, ability, trust, self-esteem, motivation, success, ambition, prejudice, alienation, and conservatism, cannot be directly observed. They are essentially hypothetical constructs or concepts, for which there exists no operational method for direct measurement. Researchers can only find some observed measures that are indicators of a latent variable. The observed indicators of a latent variable usually contain sizable measurement errors. Even for variables, which can be directly measured, measurement errors are always a concern in statistical analysis. Traditional statistical methods [e.g., multiple regressions, analysis of variance (ANOVA), path analysis, simultaneous equations] ignore the potential measurement error of variables included in a model. If an independent variable in a multiple regression model has measurement error, then the model residuals would be correlated with this independent variable, leading to violation of the basic statistical assumption. As a result, the parameter estimates of the regression model would be biased and result in incorrect conclusions. SEM provides a flexible and powerful means of simultaneously assessing the quality of measurement and examining causal relationships among constructs. That is, it offers an opportunity of constructing the unobserved

Structural Equation Modeling: Applications Using Mplus, First Edition. Jichuan Wang and Xiaoqian Wang.

latent variables and estimating the relationships among the latent variables that are uncontaminated by measurement errors.

Other advantages of SEM include, but are not limited to, the ability to model multiple dependent variables simultaneously; the ability to test overall model fit, direct and indirect effects, complex and specific hypotheses, and parameter invariance across multiple between-subjects groups; the ability to handle difficult data (e.g., time series with autocorrelated error, non-normal, censored, count and categorical outcomes), and to combine person-centered and variable-centered analytical approaches. The related topics on these model features will be discussed in the following chapters.

This chapter gives a brief introduction to SEM through five steps that characterize most SEM applications (Bollen and Long, 1993):

1. *Model formulation.* It refers to correctly specifying the SEM model that the researcher wants to test. The model may be formulated on the basis of theory or empirical findings. A general SEM model is composed of two parts: the measurement model and the structural model.
2. *Model identification.* It determines whether there is a unique solution for all the free parameters in the specified model. Model estimation cannot be implemented if a model is not identified, and model estimation may not converge or reach a solution if the model is misspecified.
3. *Model estimation.* It is to estimate model parameters and generate fitting function. Various estimation methods are available for SEM. The most common method for SEM model estimation is maximum likelihood.
4. *Model evaluation.* After meaningful model parameter estimates are obtained, the researcher needs to assess whether the model fits the data. If the model fits data well and results are interpretable, then the modeling process can stop after this step.
5. *Model modification.* If the model does not fit the data, re-specification or modification of the model is needed. In this instance, the researcher makes a decision regarding how to delete, add, or modify parameters in the model. The fit of the model could be improved through parameter re-specification. Once a model is re-specified, steps 1 through 4 may be carried out again. The model modification may be repeated more than once in real research. In the following sections we will introduce the SEM process step by step.

1.1 Model formulation

In SEM, researchers begin with the specification of a model to be estimated. There are different approaches to specify a model of interest. The most intuitive way of doing this is to describe one's model by path diagrams first suggested by Wright (1934). Path diagrams are fundamental to SEM since it allows researchers to formulate the model of interest in a direct and appealing fashion. The diagram provides a useful guide for clarifying a researcher's ideas about the relationships among variables and they can be directly translated into corresponding equations for modeling. Several conventions are used in developing a SEM model path diagram, in which the observed variables (also known as measured variables, manifest variables, or indicators) are presented in boxes, and latent variables or factors are in circles or ovals. Relationships between variables are indicated by lines; lack of line connecting variables implies that no direct relationship has been hypothesized between the corresponding variables. A line with a single arrow represents a hypothesized direct relationship between two variables, with the head of the arrow pointing toward the variable being influenced by another variable. The bidirectional arrows refer to relationships or associations, instead of effects, between variables.

An example of a hypothesized general structural equation model is specified in the path diagram shown in Figure 1.1. As mentioned above, the latent variables are enclosed in ovals and the observed variables are in boxes in the path diagram. The measurement of a latent variable or a factor is accomplished through one or more observable indicators, such as responses to questionnaire items that are assumed to represent the latent variable. In our example two observed variables (x_1 and x_2) are used as indicators of the latent variable ξ_1, three indicators ($x_1 - x_3$) for latent variable ξ_2, and three ($y_1 - y_3$) for latent variable η_1. Note that η_2 has a single indicator, indicating that the latent variable is directly measured by a single observed variable. This special case will be discussed later.

The latent variables or factors that are determined by variables within the model are called endogenous latent variables, denoted by η; the latent variables, whose causes lie outside the model, are called exogenous latent variables, denoted by ξ. In the example model, there are two exogenous latent variables (ξ_1 and ξ_2) and two endogenous latent variables (η_1 and η_2). Indicators of the exogenous latent variables are called exogenous indicators (e.g., $x_1 - x_5$), and indicators of the endogenous latent variables are endogenous indicators (e.g., $y_1 - y_4$). The former has a

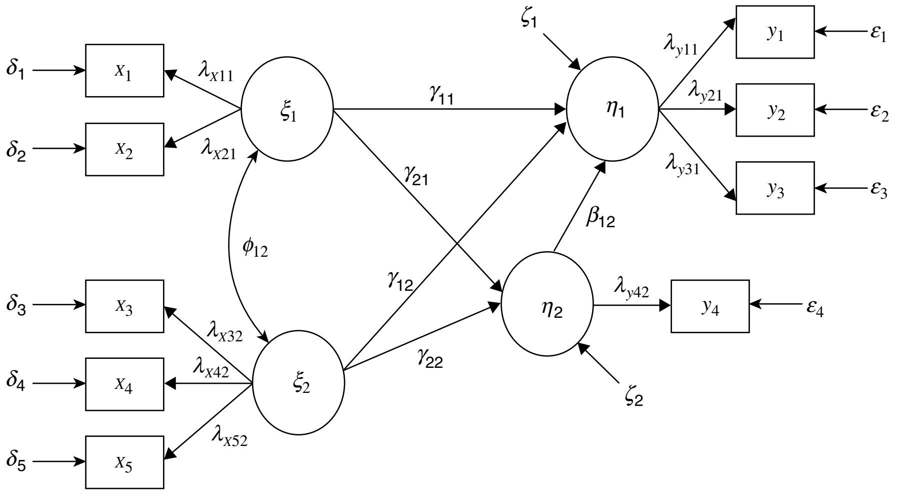

Figure 1.1 A hypothesized general structural equation model.

measurement error term symbolized as δ, and the latter has measurement errors symbolized as ε (Figure 1.1).

The coefficients β and γ in the path diagram are path coefficients. The first subscript notation of a path coefficient indexes the dependent endogenous variable, and the second subscript notation indexes the causal variable (either endogenous or exogenous). If the causal variable is exogenous (ξ), the path coefficient is a γ; if the causal variable is another endogenous variable (η), the path coefficient is a β. For example, β_{12} is the effect of endogenous variable η_2 on the endogenous variable η_1; γ_{12} is the effect of the second exogenous variable ξ_2 on the first endogenous variable η_1. As in multiple regressions, nothing is predicted perfectly; there are always residuals or errors. The ζs in the model, pointing toward the endogenous variables, are structural equation residual terms.

Different from the traditional statistical methods, such as multiple regressions, ANOVA, and path analysis, SEM focuses on latent variables/factors rather than on the observed variables. The basic objectives of SEM are to provide a means of estimating the structural relations among the unobserved latent variables of a hypothesized model free of the effects of measurement errors. These objectives are fulfilled through integrating a measurement model (confirmatory factor analysis, CFA) and structural model (structural equations or latent variable model) into the framework of a structural equation model. It can be claimed that a general structural equation model consists of two parts: (1) the measurement model that links observed variables to unobserved latent variables (factors); and (2) structural equations that link the latent variables to each other via a system of simultaneous equations (Jöreskog, 1973).

1.1.1 Measurement model

A measurement model is the measurement component of a structural equation model. The main purpose of a measurement model is to describe how well the observed indicator variables serve as a measurement instrument for the underlying latent variables or factors. Measurement models are usually carried out and evaluated by CFA. As a measurement model, CFA proposes links or relations between the observed indicator variables and the underlying latent variables/factors that they are designed to measure; then, it tests them against the data to 'confirm' the proposed factorial structure.

In the structural equation model specified in Figure 1.1, three measurement models can be considered (Figure 1.2a–c). In each measurement model, the λ coefficients, which are called factor loadings in the terminology of factor analysis, are the links between the observed variables and latent variables. For example, in Figure 1.2a the observed variables $x_1 - x_5$ are linked through $\lambda_{x11} - \lambda_{x52}$ to latent variables ξ_1 and ξ_2, respectively. In Figure 1.2b the observed variables $y_1 - y_3$ are linked through $\lambda_{y11} - \lambda_{y31}$ to latent variable η_1. Note that Figure 1.2c can be considered as a special CFA model with a single factor η_2 and a single indicator y_4. Of course this model cannot be estimated separately because it is unidentified. We will discuss this issue later.

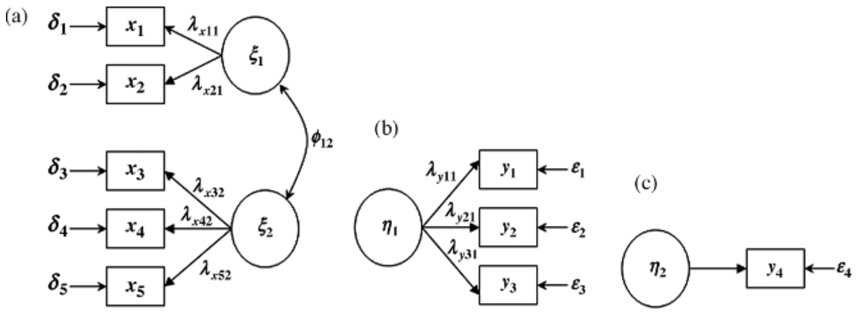

Figure 1.2 (a) Measurement model 1. (b) Measurement model 2. (c) Measurement model 3.

Factor loadings in CFA models are usually denoted by the Greek letter λ. The first subscript notation of a factor loading indexes the indicator, and the second subscript notation indexes the corresponding latent variable. For example, λ_{x21} represents the factor loading linking indicator x_2 to exogenous latent variable ξ_1; and λ_{y31} represents the factor loading linking indicator y_3 to endogenous latent variable η_1.

In the measurement model shown in Figure 1.2a, there are two latent variables/factors, ξ_1 and ξ_2, each of which is measured by a set of observed indicators. Observed variables x_1 and x_2, are indicators of the latent variable ξ_1, and $x_3 - x_5$ are indicators of ξ_2. The two latent variables, ξ_1 and ξ_2, in this measurement mode are correlated with each other (ϕ_{12} in Figure 1.2a stands for the covariance between ξ_1 and ξ_2), but no directional or causal relationship is assumed between the two latent variables. If these two latent variables were not correlated with each other (i.e., $\phi_{12} = 0$) there would be a separate measurement model for ξ_1 and ξ_2, respectively, where the measurement model for ξ_1 would have only two observed indicators, thus it would not be identified.

For a one-factor solution CFA model, a minimum of three indicators is required for model identification. If no errors are correlated, a one-factor CFA model with three indicators (e.g., the measurement model shown in Figure 1.2b) is just identified (i.e., the number of observed variances/covariances equals the number of free parameters).[1] In such a case, model fit cannot be assessed although model parameters can be estimated. In order to assess model fit, the model must be over-identified (i.e., the observed pieces of information are more than model parameters that need to be estimated). Without specifying error covariances, a one-factor solution CFA model needs at least four indicators in order to be over-identified. However, a factor with only two indicators may be acceptable if the factor is specified to be correlated with at least one of the other factors in a CFA model and no error terms are

[1] For a one-factor CFA model with three indicators, there are $3(3+1)/2 = 6$ observed variances/covariances. When covariance structure (COVS) is analyzed, six free parameters: two factor loadings (one loading is fixed to 1.0), one variance of the factor, and three variances of the error terms; thus degrees of freedom (df) $= 0$.

correlated with each other (Bollen, 1989a; Brown, 2006). The measurement model shown in Figure 1.2a is over-identified though factor ξ_1 has only two indicators. Nonetheless, multiple indicators need to be considered to represent the underlying construct more completely since different indicators can reflect nonoverlapping aspects of the underlying construct.

Figure 1.2c shows a simple measurement model. For some single observed indicator variables (e.g., gender, ethnicity) that are less likely to have measurement errors, the simple measurement model would become like $y_4 = \eta_2$, where factor loading λ_{y42} is set to 1.0 and measurement error ε_4 is 0.0. That is, the observed variable y_4 is a 'perfect' measure of construct η_2. If the single indicator is not a perfect measure, measurement error cannot be modeled but rather one must specify a fixed measurement error variance based on a known reliability of the indicator (Hayduk, 1987; Wang *et al.*, 1995). This issue will be discussed in Chapter 3.

1.1.2 Structural model

Once latent variables/factors have been assessed in the measurement models, the potential relationships among the latent variables are hypothesized and assessed in the structural model (structural equations or latent variable model) (Figure 1.3), in which path coefficients γ_{11}, γ_{12}, γ_{21}, and γ_{22} specify the effects of the exogenous latent variables ξ_1 and ξ_2 on the endogenous latent variables η_1 and η_2, while β_{12} specifies the effect of η_2 on η_1; that is, the structural model defines the relationships among the latent variables, and it is estimated simultaneously with the measurement models. Note, if the variables in a structural model were all observed variables, rather than latent variables, the structural model would become a modeling system of structural relationships among a set of observed variables; thus, the model reduces to the traditional path analysis in sociology or simultaneous equation model in econometrics.

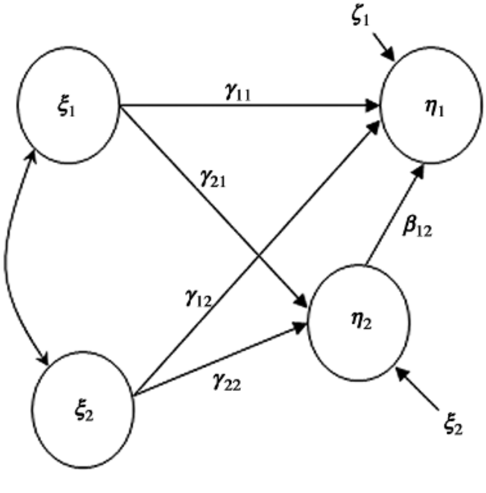

Figure 1.3 Structural model.

The model shown in Figure 1.3 is a recursive model. If the model allows for reciprocal or feedback effects (e.g., η_1 and η_2 influence each other), then the model is called a nonrecursive model. Applications of only recursive models will be discussed in this book. Readers who are interested in nonrecursive models are referred to Berry (1984) and Bollen (1989a).

1.1.3 Model formulation in equations

When the covariance structure is analyzed, the general structural equation model can be expressed by three basic equations:

$$\begin{aligned} \eta &= \mathbf{B}\eta + \mathbf{\Gamma}\xi + \zeta \\ Y &= \mathbf{\Lambda}_y \eta + \varepsilon \\ X &= \mathbf{\Lambda}_x \xi + \delta \end{aligned} \tag{1.1}$$

These three equations are expressed in matrix format. Definitions of the variable matrices involved in the three equations are shown in Table 1.1.

The first equation in Equation (1.1) represents the structural model which establishes the relationships or structural equations among latent variables. The components of η are endogenous latent variables; and the components of ξ are exogenous latent variables. The endogenous and exogenous latent variables are connected by a system of linear equations with coefficient matrices **B** (beta) and $\mathbf{\Gamma}$ (gamma), as well as a residual vector ζ (zeta), where $\mathbf{\Gamma}$ represents effects of exogenous latent variables on endogenous latent variables, **B** represents effects of some endogenous latent variables on other endogenous latent variables, and ζ represents the regression residual terms.

The second and third equations in Equation (1.1) represent measurement models which define the latent variables from the observed variables. The second equation links the endogenous indicators – the observed y variables – to endogenous latent variables (i.e., ηs), while the third equation links the exogenous indicators – the observed x variables – to the exogenous latent variables (i.e., ξs).

Table 1.1 Definitions of the variable matrices in the three basic equations of the general structural equation model.

Variable	Definition	Dimension
η (eta)	Latent endogenous variable	$m \times 1$
ξ (xi)	Latent exogenous variable	$n \times 1$
ζ (zeta)	Residual term in equations	$m \times 1$
y	Endogenous indicators	$p \times 1$
x	Exogenous indicators	$q \times 1$
ε (epsilon)	Measurement errors of y	$p \times 1$
δ (delta)	Measurement errors of x	$q \times 1$

Note: m and n represent the number of latent endogenous and exogenous latent variables, respectively; p and q are the number of endogenous and exogenous indicators, respectively, in the sample.

Table 1.2 Eight fundamental parameter matrices for the general structural equation model.

Matrix	Definition	Dimension
Coefficient matrices		
Λ_{y} (lambda y)	Factor loadings relating y to η	$p \times m$
Λ_{x} (lambda x)	Factor loadings relating x to ξ	$q \times n$
$\boldsymbol{B}$ (beta)	Coefficient matrix relating η to η	$m \times m$
Γ (gamma)	Coefficient matrix relating ξ to η	$m \times n$
Variance/covariance matrices		
Φ (phi)	Variance/covariance matrices of ξ	$n \times n$
Ψ (psi)	Variance/covariance matrices of ζ	$m \times m$
Θ_ε (theta-epsilon)	Variance/covariance matrices of ε	$p \times p$
Θ_δ (theta-delta)	Variance/covariance matrices of δ	$q \times q$

Note: p is the number of y variables, q is the number of x variables, n is the number of ξ variables, and m is the number of η variables.

The observed variables y and x are related to the corresponding latent variables η and ξ by factor loadings Λ_y (lambda y) and Λ_x. The ε and δ are the measurement errors associated with the observed variables y and x, respectively. It is assumed that $\mathrm{E}(\varepsilon) = 0$, $\mathrm{E}(\delta) = 0$, $\mathrm{Cov}\,(\varepsilon,\ \xi) = 0$, $\mathrm{Cov}\,(\varepsilon,\ \eta) = 0$, $\mathrm{Cov}\,(\delta,\ \eta) = 0$, $\mathrm{Cov}\,(\delta,\ \xi) = 0$, and $\mathrm{Cov}\,(\varepsilon,\ \delta) = 0$, but $\mathrm{Cov}(\ \varepsilon_i,\ \varepsilon_j)$ and $\mathrm{Cov}\,(\eta_i,\ \eta_j)$ $(\mathrm{i} \neq \mathrm{j})$ might not be zero.

Note that no intercepts are specified in the above SEM equations. This is because the deviations from means of the original observed variables are usually used in structural equation model specification for simplicity. The original observed variables will be used for model estimation when estimates of intercepts, the means, and thresholds of variables are involved in a model. We will discuss this issue in later chapters on modeling categorical outcomes and multi-group modeling.

In the three basic equations shown in Equation (1.1), there are a total of eight parameter matrices in LISREL notation:[2] Λ_x, Λ_y, Γ, $\mathbf{B}$, Φ,Ψ, Θ_δ and Θ_δ (Jöreskog and Sörbom, 1981). A SEM model is fully defined by the specification of the structure of the eight matrices. In the early stages of SEM, a SEM model was specified in matrix format using the eight-parameter matrix. Although this is no longer the case in current SEM programs/software, information about parameter estimates in the parameter matrices are reported in the output of M*plus* and other SEM computer programs. Understanding these notations is helpful for researchers to check the estimates of specific parameters in the output.

A summary of these matrices is presented in Table 1.2. The first two matrices, Λ_{y} and Λ_{x}, are factor loading matrices that link the observed indicators to latent variables η and ξ, respectively. The next two matrices, $\mathbf{B}$ (beta) and Γ (gamma), are structural coefficient matrices. The B matrix is an $m \times m$ coefficient matrix representing the relationships among latent endogenous variables. The model assumes

[2] LISREL, standing for linear structural relationship, was the first computer software for SEM, written by Drs Karl Jöreskog and Dag Sörbom from the University of Uppsala, Sweden.

that (I – B) must be nonsingular, thus, $(\mathrm{I}-\mathrm{B})^{-1}$ exists so that model estimation can be done. A zero in the B matrix indicates the absence of an effect of one latent endogenous variable on another. For example, $\eta_{12}=0$ indicates that the latent variable η_2 does not have an effect on η_1. Note that the main diagonal of matrix B is always zero; that is, a latent variable η cannot be a predictor of itself. The Γ matrix is an $m \times n$ coefficient matrix that relates latent exogenous variables to latent endogenous variables.

There are four parameter variance/covariance matrices for a general structural equation model: Φ (phi), Ψ (psi), Θ_ε (theta-epsilon), and Θ_δ (theta-delta).[3] All four variance/covariance matrices are symmetric square matrices; that is, the number of rows equals the number of columns in each of the matrices. The elements in the main diagonal of each of the matrices are the variances that should always be positive; the elements in the off-diagonal are covariances of all pairs of variables in the matrices. When all the variables, both observed variables (i.e., indicators of latent variables) and latent variables are standardized, each of the variance/covariance matrices would become a correlation matrix in which the diagonal values would all become 1, and the off-diagonal values would become correlations. The n × n matrix Φ is the variance/covariance matrix for the latent exogenous variable ξs. Its off-diagonal element ϕ_{ij} (i.e., the element in the i^{th} row and j^{th} column in matrix Φ) is the covariance between the latent exogenous variables ξ_i and ξ_j $(i \neq i)$. If ξ_i and ξ_j were not hypothesized to be correlated with each other in the model, $\phi_{\mathrm{ij}}=0$ should be set up when specifying the model. The m × m matrix Ψ is the variance/covariance matrix of the residual terms ζ of the structural equations. In simultaneous equations of econometrics, the disturbance terms in different equations are often assumed to be correlated with each other. This kind of correlation can be readily set up in matrix Ψ and estimated in SEM. The last two variance/coviances matrices (i.e., the $p \times p$ Θ_ε and $q \times q$ Θ_δ) are variance/covariance matrices of the measurement errors for the observed variables y and x, respectively. In longitudinal studies, the autocorrelations can be easily handled by correlating specific error terms with each other.

SEM model specification is actually to formulate a set of model parameters contained in the eight matrices. Those parameters can be specified as either fixed or free. Fixed parameters are not estimated from the model and their values are typically fixed at zero (e.g., zero covariance or zero slope indicating no relationship or no effect) or 1.0 (e.g., fixing one of the factor loadings to 1.0 for the purpose of model identification). Free parameters are estimated from the model.

The hypothesized model shown in Figure 1.1 can be specified in matrix notation based on the three basic equations. First, the equation $\eta = \mathrm{B}\eta + \Gamma\xi + \zeta$ can be expressed as:

$$\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix} = \begin{bmatrix}0 & \beta_{12}\\ 0 & 0\end{bmatrix}\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix} + \begin{bmatrix}\gamma_{11} & \gamma_{12}\\ \gamma_{21} & \gamma_{22}\end{bmatrix}\begin{bmatrix}\xi_1\\ \xi_2\end{bmatrix} + \begin{bmatrix}\zeta_1\\ \zeta_2\end{bmatrix} \tag{1.2}$$

[3] The variance/covariance matrix for the latent endogenous variables η need not be estimated from modeling since it can be calculated as: $\mathrm{Var}(\eta) = \mathrm{Var}\,[(\Gamma\xi + \zeta)/(\mathrm{I}-\mathrm{B})]$.

where the free parameters are represented by symbols (e.g., Greek letters). The fixed parameters (e.g., whose values are fixed) represent restrictions on the parameters, according to the model. For example, $\beta_{21}$is fixed to zero, indicating that η_2 is not specified to be influenced by η_1 in the hypothetical model. The diagonal elements in the matrix ***B*** are all fixed to zero as a variable is not supposed to influence itself. The elements in matrix ***B*** are the structural coefficients that express endogenous latent variable η as a linear function of other endogenous latent variables; elements in matrix Γ are the structural coefficients that express endogenous variable η as a linear function of exogenous latent variables. From Equation (1.2), we have the following two structural equations:

$$\begin{aligned} \eta_1 &= \beta_{12}\,\eta_2 + \gamma_{11}\,\xi_1 + \gamma_{12}\,\xi_2 + \zeta_1 \\ \eta_2 &= \gamma_{21}\,\xi_1 + \gamma_{22}\,\xi_2 + \zeta_2 \end{aligned} \tag{1.3}$$

The measurement equation $\mathrm{Y} = \Lambda_\mathrm{y}\,\eta + \varepsilon$ can be expressed as:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \lambda_{y21} & 0 \\ \lambda_{y31} & 0 \\ 0 & \lambda_{y42} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix} \tag{1.4}$$

where the Λ_y matrix decides which observed endogenous y indicators are loaded onto which endogenous η latent variables. The fixed value of 0 indicates the corresponding indicators are not loaded onto the corresponding latent variables, while the fixed value of 1 is used for the purpose of model identification and defining the scale of the latent variable. We will discuss this issue in detail later in Chapter 2.

From Equation (1.4) we have the following four measurement structural equations:

$$\begin{aligned} y_1 &= \eta_1 + \varepsilon_1 \\ y_2 &= \lambda_{y21}\,\eta_1 + \varepsilon_2 \\ y_3 &= \lambda_{y31}\,\eta_1 + \varepsilon_3 \\ y_4 &= \lambda_{y42}\,\eta_2 + \varepsilon_4 \end{aligned} \tag{1.5}$$

As the second endogenous latent variable η_2 has only one indicator (i.e., y_4), thus λ_{y42} should be set to 1.0, thus $y_4 = \eta_2 + \varepsilon_4$. As it is hard to estimate the measurement error in such an equation in SEM, the equation is usually set to $y_4 = \eta_2$, assuming that the latent variable η_2 is perfectly measuring the single indicator y_4. However, if the reliability of y_4 is known, based on empirical finding or estimated from item reliability study, the variance of ε_4 in the equation $y_4 = \eta_2 + \varepsilon_4$ can be estimated and specified in the model to take into consideration the effect of measurement errors in y_4. We will demonstrate how to do this in Chapter 3.

Another measurement equation $X = \Lambda_x \xi + \delta$ can be expressed as:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \lambda_{x21} & 0 \\ 0 & 1 \\ 0 & \lambda_{x42} \\ 0 & \lambda_{x52} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \end{bmatrix} \tag{1.6}$$

Thus,

$$\begin{aligned} x_1 &= \xi_1 + \delta_1 \\ x_2 &= \lambda_{x21}\,\xi_1 + \delta_2 \\ x_3 &= \xi_1 + \delta_3 \\ x_4 &= \lambda_{x42}\,\xi_2 + \delta_4 \\ x_5 &= \lambda_{x52}\,\xi_2 + \delta_5 \end{aligned} \tag{1.7}$$

Among the seven random variable vectors (δ, ε, ζ, x, y, ξ, and η), x, y, ξ, and η are usually used together with the eight-parameter matrices to define a SEM model; the others are error terms or model residuals. It is assumed that $E(\zeta) = 0$, $E(\varepsilon) = 0$, and $E(\delta) = 0$, $\text{Cov}(\zeta,\xi) = 0$, $\text{Cov}(\varepsilon,\eta) = 0$, and $\text{Cov}(\delta,\xi) = 0$. In addition, multivariate normality is assumed for the observed and latent variables.

1.2 Model identification

A fundamental consideration when specifying a SEM model is model identification. Essentially, model identification concerns whether a unique value for each and every unknown parameter can be estimated from the observed data. For a given free (i.e., unknown) parameter that needs to be model estimated, if it is not possible to express the parameter algebraically as a function of sample variances/covariances, then that parameter is defined to be unidentified. We can get a sense of the problem by considering the example equation $\text{Var}(y) = \text{Var}(\eta) + \text{Var}(\varepsilon)$, where $\text{Var}(y)$ is the variance of the observed variable y, $\text{Var}(\eta)$ is the variance of the latent variable η, and $\text{Var}(\varepsilon)$ is the variance of the measurement error. There are one known [i.e., $\text{Var}(y)$] and two unknowns [i.e., $\text{Var}(\eta)$ and $\text{Var}(\varepsilon)$] in the equation; therefore, there is no unique solution for either $\text{Var}(\eta)$ or $\text{Var}(\varepsilon)$ in this equation. That is, there are an infinite number of combinations of values of $\text{Var}(\eta)$ and $\text{Var}(\varepsilon)$ that would sum to $\text{Var}(y)$, thus rendering this single equation model unidentified. If we wish to solve the problem, we need to impose some constrains in the equation. One such constraint might be to fix the value of $\text{Var}(\varepsilon)$ to a constant by adding one more equation $\text{Var}(\varepsilon) = C$ (where C is a constant). Then, $\text{Var}(\eta)$ would be ensured to have a unique estimate, that is, $\text{Var}(\eta) = \text{Var}(y) - C$. In other words, the parameter $\text{Var}(\eta)$ in the equation is identified. The same general principles hold for more complicated SEM models. If an unknown parameter can be expressed by at least one algebraic function of one or more elements of the variance/covariance matrix of the observed variables, that parameter is identified. If all the unknown parameters are identified, then the model is identified. Very often, parameters can be expressed by more than one distinct function. In this case, the parameter is over-identified. Over-identification means there is more than one way of estimating a parameter because there is more than enough information for estimating the parameter. However, parameter estimates obtained from different functions should have an identical value in the population when the model is correct (Bollen, 1989a). A model is over-identified when each parameter is identified and at least one parameter is over-identified. A model is just-indentified when each parameter is identified and none is over-identified. The term identified models refers to both just-identified and over-identified models.

A not identified (under-identified or unidentified) model has one or more unidentified parameters. If a model is under-identified, consistent estimates of all the parameters will not be attainable. Since identification is not an issue of sample size, no matter how big the sample size, an under-identified model remains under-identified. For any model to be estimated it must be either just identified or over-identified.

Over-identified SEM models are of primary interest in SEM applications. It refers to a situation where there are fewer parameters in the model than data points.[4] However, an over-identified model may not necessarily fit the data, thus creating the possibility of finding whether a model fits the observed data. The difference between the number of observed variances and covariances and the number of free parameters is called the degrees of freedom (*df*) associated with the model fit. By contrast, a just-identified model has a zero *df*, therefore goodness-of-fit cannot be tested for the model.

There is no simple set of necessary and sufficient conditions that provide a means for verification of identification of parameters in SEM models. However, two necessary conditions should always be checked. First, the number of data points must not be less than the number of free parameters. The number of data points is the number of distinct elements in the observed variance/covariance matrix, which equals $(p+q)(p+q+1)/2$ where $(p+q)$ is the total number of observed variables (i.e., p endogenous indicators and q exogenous indicators). That is, only the diagonal elements and one set of the off-diagonal elements in the observed variance/covariance matrix, either above or below the diagonal, are counted. When variance/covariance is analyzed, the free parameters in a SEM model are usually the factor loadings, factor variances/covariances, path coefficients, residual variances/covariances, and error variances that are to be estimated in the model. If there are more data points than free parameters, the model is said to be over-identified. If the data points are less than the number of free parameters, the model is said to be under-identified and parameters cannot

[4] Data points usually refer to the number of variances and covariances among the observed variables; however, when mean and covariance structures (*MACS*) are analyzed, the means of the observed variables will be counted in the data points.

be estimated because it is never possible to estimate more unknowns than there are knowns. Secondly, a measurement scale must be established for every latent variable in the model. To establish the measurement scale of a latent variable, one may (1) fix one of the factor loadings (λs) that link a latent variable to its observed indicators;[5] or (2) fix the variance of the latent variable to 1 (by doing so, the latent variable is standardized). If the variance of the latent variable is free and if all the factor loadings (λs) are free, the factor loadings and the variance of the latent variable are not identified. If one or more parameters were unidentified, specifically, for an independent latent variable, the variance of the latent variable, coefficients associated with all paths emitted by the latent variable would be unidentified; for a dependent latent variable, the residual variance and coefficients associated with all paths leading to or from the latent variable would be unidentified.

These two conditions are necessary but not sufficient. Identification problems can still arise even if these two conditions are satisfied. Although a rigorous verification of model identification can be achieved algebraically, existing SEM software/programs generally provide a check for identification during model estimation. When a model is not identified, error messages will be printed in the program output, pointing to the parameters that are involved in the identification problem. Using this information, one can modify the model in a meaningful way to eliminate the problem.

The best way to solve the identification problem is to avoid it. Usually, one can add more indicators of latent variables so that there would be more data points. However, the primary prevention strategy is to emphasize correct parameter specification. Model identification depends on the specification of parameters as free, fixed, or constrained. A free parameter is a parameter that is unknown and needs to be model estimated. A fixed parameter is a parameter that is fixed to a specified value. A constrained parameter is a parameter that is unknown but is constrained to equal one or more other parameters. Supposing that previous research shows variables x_1 and x_2 have the same effect on a dependent measure, one may constrain their path coefficients equal in the SEM model. By fixing or constraining some of the parameters, the number of free parameters can be reduced; as such, an under-identified model may become identified. In addition, reciprocal or nonrecursive SEM is another common source of identification problem. A structural model is nonrecursive when a reciprocal or bidirectional relationship is specified so that there are feedback loops between two dependent variables in the model (e.g., y_1 affects y_2 on the one hand; and y_2 affects y_1 on the other hand). Such models are generally unidentified. For the y_1 (y_2) equation to be identified, one or more instrumental variables are needed to directly affect y_1 (y_2), but not y_2 (y_1) (Berry, 1984). Nonrecursive models are not discussed in this book.

[5] Most of the existing SEM software/programs set the factor loading of the first observed indicator of a latent variable to 1.0 by default.

1.3 Model estimation

Estimation of SEM models is different from that of multiple regressions. Instead of minimizing the discrepancies between the fitted and observed values of the response variable [i.e., $\Sigma(y - \hat{y})$], SEM estimation procedures minimize the residuals that are differences between the sample variances/covariances and the variances/covariances estimated from the model.

Let use Σ to denote the population covariance matrix of observed variables y and x; Σ can be expressed as a function of free parameters θ in a hypothesized model (Appendix 1.A). The basic hypothesis in SEM is:

$$\Sigma = \Sigma(\theta) \tag{1.8}$$

where $\Sigma(\theta)$ is the model implied variance/covariance matrix; that is, the variance/covariance matrix implied by the population parameters for the hypothesized model. The purpose of model estimation or model fit is to find a set of model parameters θ to produce $\Sigma(\theta)$ so that $[\Sigma - \Sigma(\theta)]$ can be minimized. The discrepancy between Σ and $\Sigma(\theta)$ indicates how well the model fits the data.

Because both Σ and $\Sigma(\theta)$ are unknown, $[\mathrm{S} - \Sigma(\hat{\theta})]$ or $(\mathrm{S} - \hat{\Sigma})$ is actually minimized in SEM where S is the sample variance/covariance matrix, $\hat{\theta}$ are the model parameter estimates, and $\Sigma(\hat{\theta})$ or $\hat{\Sigma}$ is the model estimated/implied variance/covariance matrix. As aforementioned, a given theoretical SEM model is represented by specifying a pattern of fixed and free (estimated) elements in each of the eight model parameter matrices. The matrix of observed covariances (S) is used to estimate values for the free parameters in the matrices that best reproduce the data. Given any set of specific numerical values of the eight model parameter matrices (Table 1.2), one and only one $\hat{\Sigma}$ would be reproduced. If the model is correct, $\hat{\Sigma}$ would be very close to S. This estimation process involves the use of a particular fitting function to minimize the difference between S and $\hat{\Sigma}$. There are many fitting functions or estimation procedures available for model estimation. The most commonly employed fitting function for SEM is the maximum likelihood (ML) function (see Appendix 1.B):

$$F_{ML}(\hat{\theta}) = \ln\left|\hat{\Sigma}\right| + tr\left(S\hat{\Sigma}^{-1}\right) - \ln|S| - (p+q) \tag{1.9}$$

where S and $\hat{\Sigma}$ are the sample and model estimated variance/covariance matrices, respectively, and $(p+q)$ is the number of observed variables involved in the model [yielding $(p+q)(p+q+1)/2$ unique variances and covariances].

The goal in SEM estimation is to estimate model parameters such that a function of the discrepancy between S and $\hat{\Sigma}$ is minimized. When a model fits data perfectly, the model estimated variance/covariance equals the sample variance/covariance matrix (i.e., $\hat{\Sigma} = \mathrm{S}$), then $\ln|\hat{\Sigma}| = \ln|S|$ and $\mathrm{tr}(\hat{\Sigma}^{-1}\mathrm{S}) = \mathrm{tr}(\mathrm{I}) = (p+q)$, therefore $F_{ML}(\hat{\theta}) = 0$. That is, a perfect model fit is indicated by a zero value of the fitting function.

The ML estimator has several important properties. First, ML estimates are unbiased – they on average, in large samples, neither overestimate nor underestimate the corresponding population parameters. Secondly, ML estimates are consistent – they converge in probability to the true value of the population parameters being estimated as sample size increases. Thirdly, ML estimates are efficient – they have minimum variance when sample size is large. Fourthly, the distribution of the parameter estimate approximates normal distribution as sample size increases (i.e., they are asymptotically normally distributed). Fifthly, ML function is usually scale free – change in variable scale does not yield different solutions. Finally, the ML fitting function $F_{ML}(\hat{\theta})$ multiplied by $(n-1)$ approximates a χ^2 distribution under the assumption of multivariate normality and large sample size, and the model χ^2 can be used for testing overall model fit.

ML is carried out for continuous outcome measures under normality assumption. Under conditions of severe non-normality, ML parameter estimates are less likely to be biased but the standard errors of parameter estimates may be biased, and the model χ^2 statistic may be enlarged, leading to an inflated Type I error for model rejection. When non-normality threatens the validity of the ML significance tests, several remedies are possible. First, researchers may consider transformations of non-normal variables that lead them to better approximate multi-normality. Secondly, remove outliers from data. Thirdly, bootstrap procedures may be applied to estimate variances of parameter estimates for significance tests (Bollen and Stine, 1993; Efron and Tibshirani, 1993; Shipley, 2000). Finally, alternative robust estimators that allow for non-normality may be applied.

A well-known asymptotically distribution free (ADF) estimator developed by Browne 1982, 1984) does not assume multivariate normality of the observed variables. ADF is a weighted least square estimator, where the weight matrix is a consistent estimate of the asymptotic covariance matrix of the sample variances and covariances (or correlations) (Browne, 1984; see also Kaplan, 2000). In Joreskog and Sorbom's notations (1988), the weight matrix in ADF is a $(k \times k)$ matrix with $k = p(p+1)/2$ and p is the total number of observed variables. Thus, the size of the weight matrix increases dramatically with increase of the number of observed variables. As a result, ADF is very computationally demanding even with a limited number of observed variables. In addition, ADF requires a large sample size in order to obtain consistent and efficient estimates (Muthén and Kaplan,1985, 1992; Jöreskog and Sörbom, 1989; Bentler and Yuan, 1999). According to Jöreskog and Sörbom (1988), the required sample size for estimating the weight of ADF should be at least 200 if $p \leq 12$, and at least $1.5p\,(p+1)$ if $p > 12$, where p is the number of observed variables. For example, our CFA model demonstrated in Chapter 2 has 18 observed indicators, the sample size needed for ADF would be $\mathrm{N} = 513$. The needed sample size would increase substantially when we expand the CFA into SEM models. It is usually hard to have a large enough sample size to utilize ADF even with a moderate number of observed variables. Because of these disadvantages, the application of ADF in real research is limited.

Another approach proposed by Satorra and Bentler (1988; Bentler 1995, 2005) is to adjust the ML estimator to account for non-normality. This method provides a rescaled χ^2 statistic which is robust under non-normality (Hoogland, 1999; Boomsma and Hoogland, 2001). Bentler and Yuan (1999) also proposed an adjusted ADF χ^2 and found it performed well at small sample size.

Several robust estimators are available in M*plus* for dealing with non-normality (Muthén and Muthén, 1998–2010). For example, MLM is a ML estimator which provides robust standard errors and mean adjusted χ^2 statistic. The MLM χ^2 statistic is referred to as the Satorra–Bentler χ^2. Another estimator, MLR, is a sandwich estimator with robust standard errors. The MLR χ^2 statistic is referred to as the Yuan–Bentler T_2^* test statistic (Muthén and Muthén, 1998–2010; Yuan and Bentler, 2000). MLR is recommended for small and medium sample size (Muthén, 2002). In the current version of M*plus*, MLM cannot handle missing values, while MLR allows missing completely at random (MCAR) and missing at random (MAR).

Using a numerical integration algorithm, M*plus* allows ML estimators (e.g., ML, MLM, and MLR) to estimate SEM models with categorical outcomes and continuous latent variables. However the estimation is computationally demanding, particularly as the number of factors and the sample size increase (Muthén and Muthén, 1998–2010). When a ML estimator is used for modeling categorical outcomes, the link function is Logit link by default in M*plus*.

To estimate SEM models with categorical outcome measures or a combination of binary, ordered categorical, and continuous outcome measures, the more generalized weighted least square based robust estimators, such as mean-adjusted WLS estimator (WLSM) and the mean and variance-adjusted WLS (WLSMV), are available in M*plus* (Muthén and Muthén, 1998–2010). The link function for WLS estimators is Probit link by default. Both WLSM and WLSMV are robust estimators, and they provide identical parameter estimates and standard errors. The difference between the two estimators is the model χ^2 statistic is adjusted differently. When categorical outcomes are modeled, the default estimator is WLSMV.

In model estimation, the *full information maximum likelihood* (*FIML*) approach is used by default in M*plus* to deal with missing data. *FIML* uses every piece of information in the observed data for analysis (Finkbeiner, 1979). Importantly, it not only assumes MCAR, but also MAR. In the case of MAR, where missingness is allowed to be related to both observed covariates and observed outcomes, *FIML* is more efficient and less biased than the traditional approaches (e.g., *LISTWISE* deletion, *PAIRWISE* deletion, or mean imputation methods) (Little and Rubin, 1987; Arbuckle, 1996; Wothke, 2000). Note, for modeling censored and categorical outcomes using WLS estimators, missingness is allowed to be a function of the observed covariates but not the observed outcomes (Muthén and Muthén, 1998–2010). As such, missingness allowed for WLS estimators is less restrictive than MCAR, but more restrictive than MAR.

1.4 Model evaluation

A key feature of SEM is to conduct an overall model fit test on the basic hypothesis $S = \hat{\Sigma}$. That is, to assess the degree to which the model estimated variance/covariance matrix $\hat{\Sigma}$ differs from the observed sample variance/covariance matrix S (Hoelter, 1983; Bollen, 1989a; Jöreskog and Sörbom, 1989; Bentler, 1990). If the model estimated variance/covariance matrix, $\hat{\Sigma}$, is not statistically different from the observed data covariance matrix, S, then we say the model fits data well, and we accept the null hypothesis or we say the model supports the plausibility of postulated relations among the variables; otherwise the model does not fit the data, and the null hypothesis should be rejected. The overall model fit evaluation should be done before interpreting the parameter estimates. Without evaluating the model fit any conclusion from the model estimation could be misleading.

To assess the closeness of S to $\hat{\Sigma}$, numerous model fit indices have been developed. For detail information on model fit testing and model fit indices, the readers are referred to Marsh, Balla, and McDonald (1988), Bollen (1989a), Gerbing and Anderson (1993), Tanaka (1993), Hu and Bentler (1995, 1998, 1999). Most of the SEM software/programs (e.g., LISREL, EQS, AMOS) provide a long list of model fit indices. However, only a few model fit indices are actually reported in real studies. In the following we focus on the model fit indices that M*plus* provides and that are commonly reported in SEM applications.

The model χ^2 statistic: The χ^2 statistic is the original fit index for structural models, which is defined as:[6]

$$\chi^2 = f_{\mathrm{ML}}(N - 1) \tag{1.10}$$

where $f_{\mathrm{ML}} = \mathrm{F}(\mathrm{S}, \hat{\Sigma})$ is the minimum value of the fitting function for the specified model (Appendix 1.B), and N is the sample size. This product is distributed asχ^2if the data are multivariate normal, and the specified model is correct. The χ^2 statistic assesses the magnitude of the discrepancy between the sample and the model estimated variance/covariance matrices. Different from the traditional statistical testing, instead of a significant χ^2 statistical test, a nonsignificant χ^2 is desired. That is, we expect the test not to reject the null hypothesis (H_0: the residual matrix is zero or there is no difference between the model estimated variances/covariances and the observed sample variances/covariances). As a matter of fact, this χ^2 is a *badness-of-fit* measure in the sense that a large χ^2 corresponds to bad fit, a small χ^2 to good fit, and a χ^2 value of zero indicates a perfect fit.

The model χ^2 statistic is a conventional overall test of fit in SEM. Before the model χ^2 statistic was developed by Jöreskog (1969), factor analysis was simply based on subjective decisions. The χ^2 statistic provides, for the first time, a means of evaluating factor analysis models with more objective criteria. However, the χ^2

[6] In most SEM computer programs, model χ^2 is defined as $\chi^2 = f_{\mathrm{ML}}(N - 1)$, but it is defined as $\chi^2 = f_{\mathrm{ML}}(N)$ in M*plus*.

statistic has some explicit limitations. First, χ^2 is defined as $N-1$ times the fitting function; thus, it is highly sensitive to sample size. The larger the sample size, the more likely to reject the model, thus the more likely to have a Type I error (rejecting the correct hypothesis). The probability of rejecting a model would substantially increase when sample size increases even when the difference between the observed and the model estimated variance/covariance matrices are trivial. Secondly, when sample size is small, the fitting function may not follow a χ^2 distribution. Thirdly, χ^2 is very sensitive to violations of the assumption of multivariate normality. The χ^2 value increases when variables have highly skewed and kurtotic distributions. Finally, χ^2 increases when the number of variables in a model increases. As such, the significance of the χ^2 test should not be a reason by itself to reject a model. To address the limitations of the χ^2 test, a number of model fit indices have been proposed for the model fit test.

Comparative fit index (CFI): As the name implies, Bentler's (1990) CFI compares the specified model with the null model which assumes zero covariances among the observed variables. This measure is based on the noncentrality parameter $d=(\chi^2 - df)$ where df is the degrees of freedom of the model.[7] The CFI is defined as:

$$CFI = \frac{d_{null} - d_{specified}}{d_{null}} \tag{1.11}$$

where d_{null} and $d_{specified}$ are the noncentrality parameters for the null model and the specified model, respectively. The CFI is defined as the ratio of improvement in noncentrality (moving from the null to the specified model) to the noncentrality of the null model. As the null model has the worst fit, it has considerably higher noncentrality (larger d) than a specified model. The values of CFI range from 0 to 1 (if outside this range it is reset to 0 or 1). CFI is an incremental fit index or relative fit index. Analogous to R^2, CFI = 0 indicates the worst fit and CFI = 1 indicates the best fit. Traditionally, the rule of thumb reasonable cutoff for the fit index is 0.90. However, Hu and Bentler 1998, 1999) suggest increasing this minimum rule of thumb from 0.90 to 0.95. The CFI is a good fit index even in small samples (Bentler, 1995). However, the CFI depends on the average size of the correlations in the data. If the average correlation between variables is not high, then the CFI will not be very high.

Tucker–Lewis index (TLI) or *non-normed fit index* (NNFI): The TLI (Tucker and Lewis, 1973) is also called the NNFI by Bentler and Bonett (1980). The TLI is

[7] The noncentrality parameter is estimated as $(\chi^2 - df)$ [if $\chi^2 < df$, then set $(\chi^2 - df) = 0$]. When model χ^2 equals the df, the model fit is considered perfect. When a model is incorrectly specified, the model χ^2 statistic would follow a noncentral χ^2 distribution that can be approximately considered as a result of the central χ^2 being shifted to the right by $(\chi^2 - df)$ units. As such, the noncentrality parameter can be considered as an index that reflects the degree to which a model fails to fit data. The larger the $(\chi^2 - df)$, the worse the model fit; the smaller the $(\chi^2 - df)$, the better the model fit.

another way to compare the lack of fit of a specified model to the lack of fit of the null model. TLI is defined as:

$$TLI = \frac{\left(\frac{\chi^2_{null}}{df_{null}} - \frac{\chi^2_{specified}}{df_{specified}}\right)}{\left(\frac{\chi^2_{null}}{df_{null}} - 1\right)} \qquad (1.12)$$

where χ^2_{null}/df_{null} and $\chi^2_{specified}/df_{specified}$ are ratios of χ^2 statistics to the degrees of freedom of the null model and the specified model, respectively. As such, TLI has a penalty for model complexity because the more free parameters the smaller $df_{specified}$, thus the larger $\chi^2_{specified}/df_{specified}$, leading to a smaller TLI.

Like CFI, TLI is an incremental fit index, and its values are not guaranteed to vary from 0 to 1. If its value is outside the 0–1 range, then reset it to 0 or 1. A negative TLI indicates that the χ^2/df ratio for the null model is less than the ratio for the specified model. This situation might occur if the specified model has too few degrees of freedom and correlations among the observed variables are low. Though TLI tends to run lower than CFI, the recommended cut-off value for TLI is the same for CFI. A TLI value lower than 0.90 indicates a need to respecify the model.

TLI also depends on the average size of the correlations in the data. If the average correlation between variables is not high, then the TLI will not be high. Unlike CFI, the TLI is moderately corrected for parsimony: its value estimates the relative model fit improvement per degree of freedom over the null model (Hoyle and Panter, 1995). TLI is often reported along with the CFI; this is akin to reporting the AGFI along with the GFI when LISREL or other SEM programs are used for modeling.

Root mean square error of approximation (RMSEA): RMSEA is one of the most recently proposed tests of model fit. The error of approximation means the lack of fit of the specified model to the population. This measure is based on the noncentrality parameter as:

$$RMSEA = \sqrt{\frac{(\chi^2_S - df_S)/N}{df_S}} = \sqrt{\frac{(\chi^2_S/df_S) - 1}{N}} \qquad (1.13)$$

where $(\chi^2_s - df_s)/\mathrm{N}$ is the rescaled non-centrality parameter to adjust for sample size. By adjusting for the model degrees of freedom, RMSEA measures average lack of fit per degree of freedom. The values of RMSEA are often interpreted as: 0 = perfect fit; <0.05 = close fit; 0.05–0.08 = fair fit; 0.08–0.10 = mediocre fit; and >0.10 = poor fit (Browne and Cudeck, 1993; MacCallum, Browne, and Sugawara, 1996; Byrne, 1998). Hu and Bentler (1999) suggest RMSEA $\leq$ 0.06 as the cut-off for a good model fit.

Besides the model χ^2 statistic, RMSEA is the only model fit index so far that can provide a confidence interval (CI) around its calculated value. The CI of RMSEA is asymmetric around the point estimate and ranges from zero to positive infinity (Browne and Cudeck, 1993). Usually, RMSEA is reported with its 90 % CI In a well-fitting model, the lower 90 % confidence limit includes or is close to 0, while the upper limit is less than 0.08. In addition, a close-fit test for null hypothesis (H_0: RMSEA ≤ 0.05) can be conducted. The P-value examines the alternative hypothesis (H_A: RMSEA > 0.05). If $P > 0.05$, then we cannot reject the null hypothesis, therefore, the specified model has a 'close fit.' RMSEA has become an increasingly used model fit index in applications of SEM, and simulation studies have shown that RMSEA performs better than other fit indices (Steiger, 1990; Browne and Cudeck, 1993; Sugawara and MaCallum, 1993; Marsh and Balla, 1994; Browne and Arminger, 1995).

Root mean square residual (RMR): This is residual-based model fit index. The RMR is the square root of the average residual. As aforementioned, residuals in SEM are differences in the elements between the sample variance/covariance matrix (S) and the model implied variance/covariance matrix ($\hat{\Sigma}$). RMR is defined as (Jöreskog and Sörbom, 1981):

$$RMR = \left(\sum_j \sum_k (s_{jk} - \hat{\sigma}_{jk})^2 \Big/ e \right)^{1/2} \tag{1.14}$$

where s_{jk} and $\hat{\sigma}_{jk}$ are elements in the observed variance/covariance matrix S and the model estimated variance/covariance matrix $\hat{\Sigma}$, respectively, $e = p\,(p + 1)/2$, and p is the total number of observed indicator variables.

Standardized root mean square residual (SRMR): SRMR is a standardized version of RMR based on standardized residuals. It is defined as (Bentler, 1995; Muthén, 1998–2004):

$$SRMR = \left(\left(\sum_j \sum_k r_{jk}^2 \right) \Big/ e \right)^{1/2} \tag{1.15}$$

where r_{jk} is the difference in the elements between the observed correlation matrix and the model estimated correlation matrix (Muthén, 1998–2004):

$$r_{jk} = \left(\frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}} \right) - \left(\frac{\hat{\sigma}_{jk}}{\sqrt{\hat{\sigma}_{jj}}\sqrt{\hat{\sigma}_{kk}}} \right) \tag{1.16}$$

where s_{jk} is the sample covariance between the observed variables $\mathrm{y_j}$ and $\mathrm{y_k}$ and $\hat{\sigma}_{jk}$is the corresponding model estimated covariance; s_{jj} and s_{kk} are sample variances and $\hat{\sigma}_{jj}$ and $\hat{\sigma}_{kk}$ are model estimated variances, respectively. A value of SRMR less than 0.08 is considered a good fit (Hu and Bentler, 1999; Kline, 2005), and less than 0.10 is acceptable (Kline, 2005). The value of this index tends to be smaller when sample size and the number of parameters in the model increase.

Weighted root mean square residual (WRMR): WRMR is another variant of RMR, which is defined as (Muthén, 1998–2004):

$$WRMR = \left\{ \frac{\sum_j \sum_k \left(s_{jk} - \hat{\sigma}_{jk}\right)^2}{\upsilon_{jk}} \Bigg/ e \right\}^{1/2} \tag{1.17}$$

where $\left(s_{jk} - \hat{\sigma}_{jk}\right)$ is the residual, υ_{jk} is the estimated asymptotic variance of s_{jk}, and e is the total number of sample variances and covariances. WRMR is more suitable for models where sample statistics have large disparate variances, outcome measures have non-normal distributions, and when sample statistics are on different scales such as in models with mean and/or threshold structures (Muthén, 1998–2004). A WRMR value of 1.0 or lower is considered a good fit (Yu, 2002).

Information criteria indices: Information criterion statistics are relative model fit statistics that are commonly used for model comparisons, including comparing non-nested models. The general form of information criterion statistics is defined as (Sclove, 1987):

$$-2\ln(L) + a(n)m \tag{1.18}$$

where L is the model maximum likelihood. The possible values of $-2\ln(L)$ range from 0 to ∞ with smaller values indicating a better fit. The term $a(n)m$ in Equation (1.18) is considered a penalty added to $-2\ln(L)$ for model complexity, where n and m are sample size and model free parameters, respectively. M*plus* provides three types of information criterion statistics: Akaike's information criterion (AIC) (1973, 1983), Bayesian information criterion (BIC) or Schwarz criterion (Schwarz, 1978), and sample-size adjusted BIC (ABIC) (Sclove, 1987), defined, respectively, as:

$$\text{AIC} = -2\ln(L) + 2m \tag{1.19}$$

$$\text{BIC} = -2\ln(L) + \ln(n)m \tag{1.20}$$

$$\text{ABIC} = -2\ln(L) + \ln(n^*)m \tag{1.21}$$

The above equations are all special cases of Equation (1.18). For AIC, the penalty term $a(n)m$ in Equation (1.18) is replaced with $2m$ regardless of sample size, whereas *a(n)m* is replaced with $\ln(n)m$ for BIC. For ABIC, sample size n is replaced with $n^* = (n + 2)/24$ to somewhat reduce the penalty for larger sample sizes (Sclove, 1987; Muthén, 1998–2004). Clearly, BIC and ABIC impose more penalties than AIC for model complexity because the product of sample size and the number of free parameters is included in the penalty term; thus BIC and ABIC favor smaller models with fewer free parameters.

With so many model fit indices being proposed, no single index should be relied on exclusively for testing a hypothesized SEM model. Instead, it is recommended that multiple fit indices should be reported for model evaluation in order to avoid making an inaccurate conclusion of model fit (Bollen, 1989a; Bollen and Long, 1992; Tanaka, 1993; Bentler, 2007). The model χ^2 statistic, RMSEA, 90% CI of RMSEA, *P*-value of the close-fit test, CFI, TLI, and SRMR are commonly reported in applications.

Importantly, the model fit indices indicate the overall model fit on average. A model with excellent fit indices does not necessarily mean that the model is a correct model. First, other model components are also important for model evaluation. For example, coefficient estimates should be interpretable, R-squares of equations are acceptable, and there are no improper solutions (e.g., negative variance, correlation less than -1 or greater than 1). Problems in the model components indicate that some parts of the model may fit the data poorly. Secondly, there may be many models that fit data equally well as judged by model fit indices. Among these equivalent models, the parsimonious model should be accepted. In addition, the model evaluation is not entirely a statistical matter. It should also be based on sound theory and empirical findings. If a model makes no substantive sense, it is not justified even if it statistically fits the data very well.

Model comparison: In SEM, it is recommended to consider alternative models rather than to examine a single model so that the best fit model can be determined by model comparisons (Bollen and Long, 1993). The likelihood ratio (LR) test is often used for model comparison in SEM for two nested models estimated from the same data set. For two models to be nested, for example, Model B is nested within Model A, Model B must have fewer free parameters, therefore, a larger number of degrees of freedom than does Model A. In addition, the parameters in Model B cannot include new parameters that are not included in Model A. Once these two conditions are satisfied, difference in model χ^2 or likelihood function between the two models will follow a χ^2 distribution with *df* that is the difference in *df* between the two models.

It is noteworthy that when some robust estimators, such as MLM, MLMV, MLR, ULSMV, WLSM and WLSMV, are used for model estimation, the model χ^2 statistics cannot be used for the LR test in the regular way because the difference in the model χ^2 statistic between two nested models does not follow a χ^2 distribution (Muthén and Muthén, 1998–2010). Such difference testing will be discussed with examples in the next chapter.

For models that are not nested, the information measures, such as AIC, BIC, and ABIC, can be used for model comparison. The model with smaller information measures has a better fit. These information measures are important parsimony-corrected indices that can be used to compare both non-nested as well as nested models. Raftery (1996), based on Jeffreys (1961), suggests some guidelines for the strength of evidence favoring one model against another model based on a difference in absolute value of BIC: 0–2, weak evidence; 2–6, positive evidence; 6–10, strong evidence; and 10+, very strong evidence.

Finally, an important approach for checking lack of model fit is to examine the model residuals. Recall, unlike the residuals in multiple regressions, the residuals in SEM are the elements in the residual matrix $(\mathrm{S}-\hat{\Sigma})$ where S is the sample variance/covariance matrix and $\hat{\Sigma}$ is the model estimated variance/covariance matrix. The residuals are dependent upon the measurement scale of the observed variables, and thus are not quire meaningful as the observed variables often have various metric. To avoid this problem, the residuals are often standardized, that is, divided by their asymptotical (large sample) standard errors, which is a complicated function of the elements of the observed variance/covariance matrix S (Jöreskog and Sörbom, 1989). Though standardized residuals are not technically a model fit index, they provide useful information about how close the estimated variances/covariances are to those observed. A large standardized residual indicates a large discrepancy in a specific variance or covariance between S and $\hat{\Sigma}$. A standardized residual is considered large if it is larger than 2.58 in magnitude (Jöreskog and Sörbom, 1989, p. 32).

1.5 Model modification

In application of SEM one usually specifies a model based on theory or empirical findings then fit the model to the available data. Very often the tentative initial model may not fit data well. In other words, the initial model may be somewhat mis-specified. In such a case, the possible sources of lack of model fit need to be assessed to determine what is specifically wrong with the model specification, then modify the model and re-test it using the same data. This process is called 'model specification search.'

To improve the initial model that does not fit the data satisfactorily, most often the modification indices (MIs) (Sörbom, 1989) that are associated with the fixed parameters of the model are used as diagnostic statistics to capture model misspecfication. A MI indicates the decrease in model χ^2 statistic with 1 *df* indicating if a particular parameter is freed from a constraint in the preceding model.

A high MI value indicates the corresponding fixed parameter should be freed to improve model fit. Although a drop in χ^2 of 3.84 with 1 *df* indicates a significant model fit improvement at $P = 0.05$ level, no strict rules of thumb exist concerning how large MIs must be to warrant a meaningful model modification. In M*plus* output MIs are listed by default if a drop in a corresponding χ^2 is at least 10. If there are several parameters with high MIs, they should be freed one at a time, beginning with the largest MI because change in a single parameter in a model could affect other parts of the solution (MacCallum, Roznowski, and Necowitz, 1992). Freeing additional parameters may improve model fit, however, the model modification must be theoretically meaningful. Associated with MI is the expected parameter change (EPC) index for the expected change in the value of a parameter if that parameter was freed (Saris, Satorra and Sörbom, 1987). M*plus* provides MIs, EPC, and standardized EPC for all parameters in the model that are fixed or constrained to be equal to other parameters (Muthén and Muthén, 1998–2010).

It must be emphasized that the model modification or re-specification should be both statistics-driven and theory-driven. Any model modification must be justified on a theoretical basis and empirical findings. Blind use of MIs for model modification should be avoided. Parameters should not be added or removed solely for the purpose of model fit improvement. Our goal is to find a model that fits data well from a statistical point of view, and importantly all the parameters of the model must have substantively meaningful interpretation.

1.6 Computer programs for SEM

A wide variety of computer programs/software has been developed in the past two decades for SEM. The most popular computer programs include LISREL (Jöreskog and Sörbom, 2006), AMOS (Arbuckle, 2006), EQS (Bentler, 1995), M*plus* (Muthèn and Muthèn, 1998–2010), SAS PROC CALIS and SAS PROC TCALIS (SAS Institute Inc., 2008). Each computer program has its own strengths and weaknesses, and most structural equation models can be estimated with each of the programs. The choice of program is often down to personal preference.

In this book, the computer program M*plus* is used for model demonstration. M*plus* was developed on the basis of the computer program LISCOMP (Muthén, 1988). While retaining most of LISCOMP's features for SEM of categorical and continuous data, M*plus* comes with some important additions. It allows SEM models with all different types of outcome measures (e.g., continuous, censored, ordinal, nominal, and count variables, as well as a combination of different variable types); it can handle various incomplete data, non-normality, and complex survey data. Additionally, some recently developed advanced models, such as multilevel SEM, mixture models, multilevel mixture models, SEM with exploratory factor analysis, and SEM with Bayesian approach, as well as Monte Carlo simulation, can be readily implemented in M*plus*. Overall, M*plus* is a user-friendly program that is becoming increasingly popular in SEM.

The models demonstrated in this book are intended to show readers how to build SEM models in M*plus* using both cross-sectional and longitudinal data. M*plus* syntax for the models is provided in the corresponding chapters of the book. While data used for these examples are drawn from public health studies, the methods and analytical techniques are applicable to SEM practices in many other fields.

Appendix 1.A Expressing variances and covariances among observed variables as functions of model parameters

Let us denote Σ the population variance/covariance matrix of variables y and x, then

$$\Sigma = \begin{bmatrix} E(YY') & E(XY')' \\ E(XY') & E(XX') \end{bmatrix} \tag{1.22}$$

where the diagonal elements are variances of the variables y and x, respectively; and the off-diagonal elements are covariances among y and x. In SEM it is hypothesized that the population variance/covariance matrix of y and x can be expressed as a function of the model parameters θ, that is:

$$\Sigma = \Sigma(\theta) \tag{1.23}$$

where $\Sigma(\theta)$ is called the model implied variance/covariance matrix.

Based on the three basic SEM equations [Equation (1.1)], we can derive that $\Sigma(\theta)$ can be expressed as functions of the parameters in the eight fundamental SEM matrices. Let us start with the variance/covariance matrix of y, then the variance/covariance matrix of x and the variance/covariance matrix of y and x, and then finally assemble them together.

The variance/covariance matrix of y can be expressed as:

$$\begin{aligned} \mathrm{E}(\mathrm{Y}Y') &= E[(\Lambda_y\eta + \varepsilon)(\Lambda_y\eta + \varepsilon)'] \\ &= E[(\Lambda_y\eta + \varepsilon)(\eta'\Lambda_y' + \varepsilon')] \\ &= E[\Lambda_y\eta\eta'\Lambda_y'] + \Theta_\varepsilon \\ &= \Lambda_y E[\eta\eta']\Lambda_y' + \Theta_\varepsilon \end{aligned} \tag{1.24}$$

were Θ_ε is the variance/covariance matrix of the error term ε.

$$\begin{aligned} \text{as } \eta &= \mathrm{B}\eta + \Gamma\xi + \zeta \\ \text{then } \eta &= (\mathrm{I} - \mathrm{B})^{-1}(\Gamma\xi + \zeta) \\ \eta\eta' &= [(\mathrm{I} - \mathrm{B})^{-1}(\Gamma\xi + \zeta)][(\mathrm{I} - \mathrm{B})^{-1}(\Gamma\xi + \zeta)]' \\ &= [(\mathrm{I} - \mathrm{B})^{-1}(\Gamma\xi + \zeta)]\left\{(\Gamma\xi + \zeta)'[(\mathrm{I} - \mathrm{B})^{-1}]'\right\} \\ &= [(\mathrm{I} - \mathrm{B})^{-1}(\Gamma\xi + \zeta)]\left\{(\xi'\Gamma' + \zeta')[(\mathrm{I} - \mathrm{B})^{-1}]'\right\} \end{aligned} \tag{1.25}$$

Assuming that ζ is independent of ξ, then

$$E(\eta\eta') = (I - \mathrm{B})^{-1}(\Gamma\Phi\Gamma' + \Psi)[(\mathrm{I} - \mathrm{B})^{-1}]' \tag{1.26}$$

where Φ is the variance/covariance matrix of the latent variable ξ; Ψ is the variance/covariance matrix of the residual ζ. Substituting Equation (1.26) into Equation (1.24), we have:

$$\mathrm{E}(\mathrm{Y}Y') = \Lambda_y\left\{(I-\mathrm{B})^{-1}(\Gamma\Phi\Gamma' + \Psi)[(I-\mathrm{B})^{-1}]'\right\}\Lambda'_y + \Theta_\varepsilon \tag{1.27}$$

This equation implies that variances/covariances of the observed y variables are a function of model parameters such as factor loadings Λ_y, path coefficients B and Γ, the variances/covariances Φ of the exogenous latent variables, residual variances/covariances matrix Ψ, and the error variances/covariances Θ_ε.

The variance/covariance matrix of x can be expressed as:

$$\begin{aligned}\mathrm{E}(\mathrm{X}X') &= E[(\Lambda_x\xi + \delta)(\Lambda_x\xi + \delta)'] \\ &= E[(\Lambda_x\xi + \delta)(\xi'\Lambda'_x + \delta')]\end{aligned} \tag{1.28}$$

Assuming that δ is independent of ξ, then

$$\begin{aligned}\mathrm{E}(\mathrm{X}X') &= E[\Lambda_x\xi\xi'\Lambda'_x + \delta\delta'] \\ &= \Lambda_x\Phi\Lambda'_x + \Theta_\delta\end{aligned} \tag{1.29}$$

where Θ_δ is the variance/covariance matrix of the error term δ. Equation (1.29) implies that variances/covariances of the observed x variables are a function of model parameters, such as the loadings Λ_x, the variances/covariances Φ of the exogenous latent variables, and the error variances/covariances Θ_ε.

The covariance matrix among x and y can be expressed as:

$$\begin{aligned}\mathrm{E}(\mathrm{X}Y') &= \mathrm{E}[(\Lambda_x\xi + \delta)(\Lambda_y\eta + \varepsilon)'] \\ &= \mathrm{E}[(\Lambda_x\xi + \delta)(\eta'\Lambda'_y + \varepsilon')]\end{aligned} \tag{1.30}$$

Assuming that δ and ε are independent of each other and independent of the latent variables, then

$$\begin{aligned}\mathrm{E}(\mathrm{X}Y') &= \mathrm{E}(\Lambda_x\xi\eta'\Lambda'_y) \\ &= \Lambda_x\mathrm{E}(\xi\eta')\Lambda'_y \\ &= \Lambda_x\mathrm{E}\left\{\xi[(\mathrm{I}-\mathrm{B})^{-1}(\Gamma\xi + \zeta)]'\right\}\Lambda'_y \\ &= \Lambda_x\mathrm{E}\left\{\xi\left\{(\Gamma\xi + \zeta)'[(\mathrm{I}-\mathrm{B})^{-1}]'\right\}\right\}\Lambda'_y \\ &= \Lambda_x\mathrm{E}\left\{\xi\xi'\Gamma'[(\mathrm{I}-\mathrm{B})^{-1}]' + \xi\zeta'[(\mathrm{I}-\mathrm{B})^{-1}]'\right\}\Lambda'_y \\ &= \Lambda_x\Phi\Gamma'[(\mathrm{I}-\mathrm{B})^{-1}]'\Lambda'_y\end{aligned} \tag{1.31}$$

Thus, the variances and covariances among the observed variables x and y can be expressed as in terms of the model parameters:

$$\Sigma(\theta)=\begin{bmatrix}\Lambda_y\left\{(I-\mathrm{B})^{-1}(\Gamma\Phi\Gamma'+\Psi)[(I-\mathrm{B})^{-1}]'\right\}\Lambda_y'+\Theta_\varepsilon & \Lambda_y(I-\mathrm{B})^{-1}\Gamma\Phi\Lambda_x' \\ \Lambda_x\Phi\Gamma'[(I-\mathrm{B})^{-1}]'\Lambda_y' & \Lambda_x\Phi\Lambda_x'+\Theta_\delta\end{bmatrix} \tag{1.32}$$

where the upper right part of the matrix is the transpose of the covariance matrix among x and y. Each element in the model implied variance/covariance matrix $\Sigma(\theta)$ is a function of model parameters. For a set of specific model parameters from the eight SEM fundamental matrices that constitute a SEM model, there is one and only one corresponding model implied variance/covariance matrix $\Sigma(\theta)$ (Hayduk, 1987).

Appendix 1.B Maximum likelihood function for SEM

In SEM model estimation, attention is directed to the sample distribution of the observed variance/covariance matrix S. If a random sample is selected from a multivariate normal population, the likelihood of finding a sample with variance/covariance matrix S is given by the Wishart distribution (Wishart, 1928):

$$W(S,\Sigma,n)=\frac{e^{-\frac{1}{2}n\cdot tr(S\Sigma^{-1})}|nS|^{\frac{1}{2}(n-K-1)}}{|\Sigma|^{\frac{1}{2}n}2^{\frac{1}{2}nK}\pi^{\frac{1}{4}K(K-1)}\prod_{k=1}^{K}\Gamma\left(\frac{1}{2}(n+1-k)\right)} \tag{1.33}$$

where S is the sample variance/covariance matrix, Σ is the population variance/covariance matrix, $n=N-1$ (where N is sample size), K is the number of variables, and Γ is the gamma function. Note that all the terms in Equation (1.33), except those involving Σ, are constant. Since we are only interested in maximizing the function rather than calculating the precise value of the function, all the constant terms in Equation (1.33) can be combined into one constant term C, thus the equation can be simplified to:

$$\begin{aligned}W(S,\Sigma,n)&=\frac{e^{-\frac{1}{2}n\cdot tr(S\Sigma^{-1})}}{|\Sigma|^{\frac{1}{2}n}}\mathrm{C}\\&=e^{-\frac{1}{2}n\cdot tr(S\Sigma^{-1})}|\Sigma|^{-\frac{1}{2}n}\mathrm{C}\end{aligned} \tag{1.34}$$

For a model that fits data perfectly, $\hat{\Sigma} = \mathrm{S}$. As such, the ratio of the Wishart function of the specified model to that of the perfect model is:

$$
\begin{aligned}
\mathrm{LR} &= \frac{e^{-\frac{1}{2}n\cdot tr(S\Sigma^{-1})}|\Sigma|^{-\frac{1}{2}n}\mathrm{C}}{e^{-\frac{1}{2}n\cdot tr(SS^{-1})}|S|^{-\frac{1}{2}n}\mathrm{C}} \\
&= e^{-\frac{1}{2}n\cdot tr(S\Sigma^{-1})}|\Sigma|^{-\frac{1}{2}n}e^{\frac{1}{2}n\cdot tr(SS^{-1})}|S|^{\frac{1}{2}n}
\end{aligned}
\tag{1.35}
$$

Taking a natural logarithm, we have

$$
\begin{aligned}
\mathrm{Ln(LR)} &= -\frac{1}{2}n\cdot tr(S\Sigma^{-1}) - \frac{1}{2}n\cdot \ln|\Sigma| + \frac{1}{2}n\cdot tr(SS^{-1}) + \frac{1}{2}n\cdot \ln S \\
&= -\frac{1}{2}n\Big[tr(S\Sigma^{-1}) + \ln|\Sigma| - tr(SS^{-1}) - \ln S\Big] \\
&= -\frac{1}{2}n\Big[tr(S\Sigma^{-1}) + \ln|\Sigma| - (p+q) - \ln S\Big]
\end{aligned}
\tag{1.36}
$$

Since a minus sign precedes the right-hand side of Equation (1.36), maximizing Equation (1.36) is equivalent to minimizing the function in brackets:

$$
F_{ML}(\theta) = \ln\left|\hat{\Sigma}\right| + tr(S\hat{\Sigma}^{-1}) - \ln S - (p+q) \tag{1.37}
$$

where $F_{ML}(\theta)$ or $\mathrm{F_{ML}}$ is called the minimum discrepancy function, which is the value of the fitting function evaluated at the final estimates (Hayduk, 1987).