□ 应用统计学丛书

Structural Equation Modeling:

Applications Using M*plus*

结构方程模型：

M*plus* 与应用（英文版）

王济川 王小倩 著

高等教育出版社
HIGHER EDUCATION PRESS

Structural Equation Modeling:

Applications Using Mplus

用统计学丛书

Structural Equation Modeling:

Applications Using M*plus*

结构方程模型：

M*plus* 与应用（英文版）

王济川 王小倩 著

JIEGOU FANGCHENG MOXING

高等教育出版社·北京
HIGHER EDUCATION PRESS BEIJING

图书在版编目(CIP)数据

结构方程模型 : Mplus与应用 = Structural Equation Modeling: Applications Using Mplus : 英文 / 王济川, 王小倩著. — 北京 : 高等教育出版社, 2012.8

(应用统计学丛书)

ISBN 978-7-04-034828-6

Ⅰ. ①结… Ⅱ. ①王… ②王… Ⅲ. ①统计分析—统计程序—英文 Ⅳ. ①C819

中国版本图书馆CIP数据核字(2012)第183607号

策划编辑 王丽萍　　责任编辑 李华英　　封面设计 王凌波　　责任印制 朱学忠

出版发行	高等教育出版社	咨询电话	400-810-0598
社　　址	北京市西城区德外大街4号	网　　址	http://www.hep.edu.cn
邮政编码	100120		http://www.hep.com.cn
印　　刷	涿州市星河印刷有限公司	网上订购	http://www.landraco.com
开　　本	787mm×1092mm 1/16		http://www.landraco.com.cn
印　　张	29.25	版　　次	2012年8月第1版
字　　数	750千字	印　　次	2012年8月第1次印刷
购书热线	010-58581118	定　　价	79.00元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换

物 料 号　34828-00

Preface

In the past two decades structural equation modeling (SEM) has quickly pervaded various fields, such as psychiatry, psychology, sociology, economics, education, demography, political sciences, as well as biology and health studies. Compared with traditional statistical methods such as multiple regression, ANOVA, path analysis, and multilevel models, the advantages of SEM include, but are not limited to, the ability to take into account measurement errors; model multiple dependent variables simultaneously; test overall model fit; estimate direct, indirect and total effects; test complex and specific hypothesis; handle difficult data (time series with auto-correlated error, non-normal, censored, and categorical outcomes); test model parameter invariance across multiple populations/groups, and conduct mixture modeling to deal with population heterogeneity. However, SEM is still an under-utilized technique in social science studies and health studies. The intent of this book is to provide a resource for learning SEM, and a reference guide for some advanced SEM models.

The book emphasizes basic concepts, methods and applications of structural equation modeling. It covers the fundamentals of SEM, as well as some recently developed advanced SEM models. Written in non-mathematical terms, a variety of SEM models for studying both cross-section and longitudinal data are discussed. Examples of various SEM models are demonstrated using real-world research data. The internationally well-known computer program M*plus* (Muthén & Muthén, 1998–2010) is used for model demonstrations, and M*plus* program syntax is provided for each example model.

This book is divided into seven chapters. Chapter 1 gives an overview of SEM. The basic concepts of SEM, the methods and principles of SEM applications are discussed through five steps of model formulation, model identification, model estimation, model evaluation, and model modification.

Chapter 2 discusses confirmatory factor analysis (CFA) and its applications. Some advanced issues in CFA modeling, such as how to deal with violation of multivariate normality assumption, censored outcome measures, and categorical outcomes, are addressed in model demonstration. At the end of the chapter the first-order CFA model is extended to second-order CFA model.

Chapter 3 discusses SEM model and its applications, starting with the special case of SEM, called MIMIC (multiple indicators and multiple causes) model, different SEM models are discussed and demonstrated using real data. This chapter addresses some important practical issues that SEM practitioners often encounter, such as interactions between covariates, interactions involving latent variables, testing differential item functioning (DIF), testing indirect and total effects, and correcting for effect of measurement error in a single indicator variable.

Chapter 4 extends the application of SEM to longitudinal data analysis where subjects are followed up over time, with repeated measures of each variables of interest. A recently developed SEM model for longitudinal data analysis, called latent growth model (LGM), is discussed. Various LGM models such as linear LGM, non-linear LGM, multi-growth process LGM, two-part LGM, and LGM with categorical outcomes are demonstrated to assess features of outcome growth trajectories.

Chapter 5 extends the application of SEM from a single group to multiple groups to assess whether measuring instrument operates equivalently across different populations/groups (i.e., measurement invariance) or whether causal relationships are invariant across populations/groups. Model demonstrations in this chapter cover multi-group CFA models, including multi-group first-order and second-order CFA models, multi-group SEM, and multi-group LGM models.

In Chapter 6 we switch our topic to mixture models (or finite mixture models) that have increasingly gained in popularity as a framework of combination of variable-centered and person-centered analytic approach. Mixture modeling enables researchers to identify unknown *a priori* homogeneous groups/classes of individuals based on the measures of interest; examine the features of heterogeneity across the groups/classes; evaluate the effects of covariates on the group/class membership; assess the relationship between the group/class membership and other outcomes; and study transitions between the latent group/class memberships over time. Different mixture models including latent class analysis (LCA) model, latent transition analysis (LTA) model, growth mixture model (GMM) and factor mixture model (FMM) are discussed and demonstrated.

The last chapter discusses power analysis and sample size for structural equation modeling. After a brief review of the rule of thumbs, regarding appropriate sample size for SEM, different approaches to estimate the sample size needed for SEM are discussed. In terms of the ability to detect nonzero model parameters, both Satorra-Saris's method (1985) and Monte Carlo simulation are applied to conduct power analysis and sample size estimates for CFA and LGM models. And then we demonstrate how to use some newly developed methods of power analysis for SEM, such as the MacCallum, Browne, & Sugawara's method (1996) and the Kim's method (2005), to calculate statistical power given a sample size or to estimate an appropriate sample size to achieve a desired power (e.g., 0.80) based upon null hypothesis test about a model overall fit index.

Structural equation modeling is a generalized analytical framework that can deal with many sophisticated modeling situations. The recent development in structural equation modeling includes, but is not limited to, continuous time survival SEM (Larsen 2005; Asparouhov, Masyn & Muthen 2006), multilevel SEM (Muthèn 1994; Toland & De Ayala 2005), multilevel mixture SEM (Asparouhov & Muthèn 2008), and exploratory SEM (Asparouhov & Muthèn, 2009), as well as Bayesian structural equation modeling (BSEM) (Asparouhov & Muthèn 2010; Muthén & Asparouhov 2011b). These topics are beyond the scope of this book.

A wide variety of computer programs are now available for structural equation modeling. Most structural equation models can be set up and estimated with each of these programs. Which program should be used is often a matter of price, support, and personal preference. The computer program used in this book for model demonstration is M*plus* (http://www.statmodel.com/) and is becoming increasingly popular in the field of structural equation modeling. This program allows researchers to conduct various advanced SEMs without much complexity of programming. 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. The M*plus* syntax used for the example models are provided in the book. While data sets used for these example models in the book are drawn from public health studies. The methods and analytical techniques are applicable to all fields of quantitative social studies.

The target readership of the book is teachers, graduate students, and researchers in social sciences and health studies. This book can be used as a resource for learning SEM and a reference guide for conducting SEMs using M*plus*. Readers are encouraged to contact the author at jiwang@gwu.edu in regard to feedback, suggestions and questions.

本书由高等教育出版社和 Wiley 公司合作出版，由 Wiley 负责编辑加工和排版。故书中量和单位等难免有不符合我国编辑规范的地方，特此说明。

Contents

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

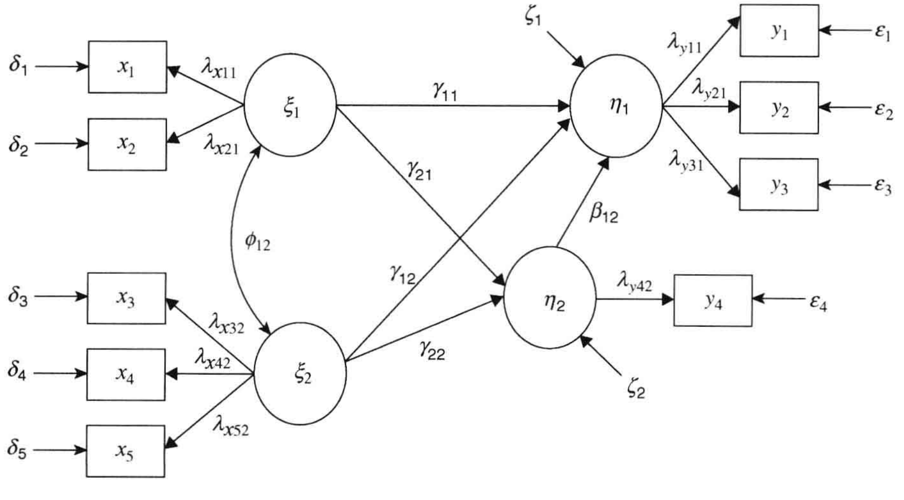

Figure 1.1 A hypothesized general structural equation model.

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.

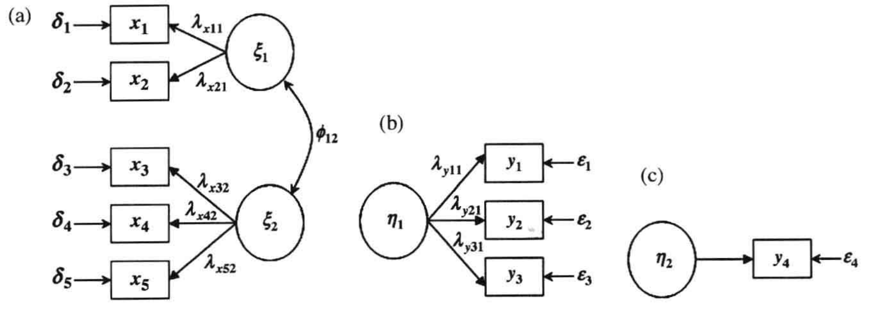

Figure 1.2 (a) Measurement model 1. (b) Measurement model 2. (c) Measurement model 3.

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 $3 \times (3+1)/2 = 6$ 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.

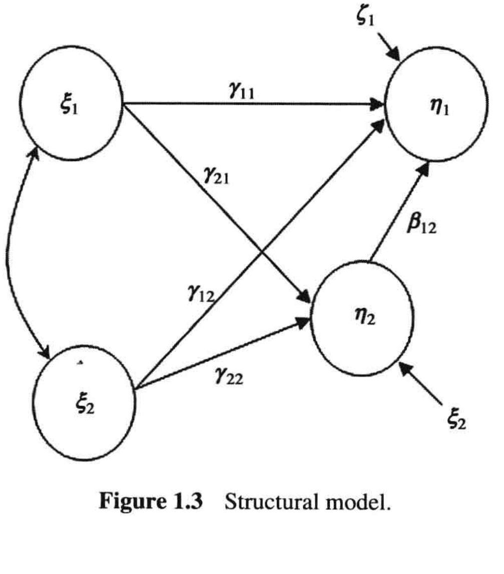

Figure 1.3 Structural model.

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:

$$\begin{aligned} \eta &= \mathrm{B}\eta + \Gamma\xi + \zeta \\ Y &= \Lambda_y \eta + \varepsilon \\ X &= \Lambda_x \xi + \delta \end{aligned} \tag{1.1}$$

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 Γ (gamma), as well as a residual vector ζ (zeta), where Γ 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.