Chapter1 Introduction

Observed variables are those that can be directly measured, such as age, weight, systolic blood pressure, diastolic blood pressure, waist hip ratio, body mass index, and heart rate.

In medical, social, and psychological research, it is common to encounter latent constructs that cannot be directly measured by a single observed variable. Examples in many fields:

Finance — behavioral tendency of insider trading
Management — job satisfaction, work attitude
Marketing — purchase habit
Education — intelligence, teaching performance, academic achievement
Psychology — personality, anxiety
Medical — quality of life, drug side effect, bone mineral density

To assess the nature of a latent construct, a combination of several observed variables is needed. For example,

- systolic blood pressure and diastolic blood pressure should be combined to evaluate blood pressure;
- waist hip ratio and body mass index should be combined to evaluate obesity.
- spine body mineral density (BMD) and hip BMD should be combined to evaluate total BMD.

In statistical inference, a latent construct is analyzed through a latent variable which is appropriately defined by a combination of several observed variables.

For practical research in social and biomedical sciences, it is often necessary to examine the relationships among the variables of interest. For example, in a study that focuses on kidney disease of type 2 diabetic patients (see Appendix 1.1), we have data from the following observed key variables:

- plasma creatine (PCr)
- urinary albumin creatinine ratio (ACR)
- systolic blood pressure (SBP)
- diastolic blood pressure (DBP)
- body mass index (BMI)
- waist hip ratio (WHR)
- glycated hemoglobin (HbAlc)
- fasting plasma glucose (FPG)

From the basic medical knowledge about kidney disease, we know that the severity of this disease is reflected by both PCr and ACR. In order to understand the effects of the explanatory (independent) variables such as SBP, BMI etc, on kidney disease, one possible approach is to apply the well-known regression model by treating PCr and ACR as outcome (dependent) variables and regressing them on the observed explanatory (independent) variables as follows:

$$\mathsf{PCr} = \alpha_1 \mathsf{SBP} + \alpha_2 \mathsf{DBP} + \alpha_3 \mathsf{BMI} + \alpha_4 \mathsf{WHR} + \alpha_5 \mathsf{HbA1c} + \alpha_6 \mathsf{FPG} + \epsilon_1, \tag{1}$$

$$\mathsf{ACR} = \beta_1 \mathsf{SBP} + \beta_2 \mathsf{DBP} + \beta_3 \mathsf{BMI} + \beta_4 \mathsf{WHR} + \beta_5 \mathsf{HbA1c} + \beta_6 \mathsf{FPG} + \epsilon_2. \tag{2}$$

From the estimates of α's and β's, we can assess the effects of the explanatory variables on PCr and ACR, respectively. However, the effects of observed explanatory variables on kidney disease cannot be directly assessed from results obtained from regression analysis of equations (1) and (2).

A better approach is to appropriately group observed variables to form latent variables. For instance,

- {PCr, ACR} — 'kidney disease (KD)'
- {SBP,DBP} — 'blood pressure (BP)'
- {BMI, WHR} — 'obesity (OB)'
- {HbA1c, FPG} — 'glycemic control (GC)'

Then, we consider a simple regression equation with latent variables:

$$\mathsf{KD} = \gamma_1 \mathsf{BP} + \gamma_2 \mathsf{OB} + \gamma_3 \mathsf{GC} + \delta. \tag{3}$$

Equation (3) can be extended to an equation with interaction terms:

$$\mathsf{KD} = \gamma_1 \mathsf{BP} + \gamma_2 \mathsf{OB} + \gamma_3 \mathsf{GC} + \gamma_4 (\mathsf{BP} \times \mathsf{OB}) + \gamma_5 (\mathsf{BP} \times \mathsf{GC}) + \gamma_6 (\mathsf{OB} \times \mathsf{GC}) + \delta. \tag{4}$$

Studying these interactive effects through the regression equations (1) and (2) with the observed variables is extremely tedious.

The advantages of incorporating latent variables in practical researches:

- It can reduce the number of variables in the key regression equation. Comparing Equation (3) with (1) and (2), the number of explanatory variables is reduced from six to three.
- As highly correlated observed variables are grouped into latent variables, the problem induced by multicollinearity is alleviated. For example, the multicollinearity induced by the highly correlated variables SBP and DBP in analyzing regression equation (1) or (2) does not exist in regression equation (3).
- It gives better assessments on the interrelationships of latent constructs. For instance, direct and interactive effects among the latent constructs blood pressure, obesity, and glycemic control can be assessed through the regression model (4).

The basic SEM is formulated by two components. The first component is a confirmatory factor analysis (CFA) model which groups the highly correlated observed variables into latent variables and takes the measurement error into account.

This component can be regarded as a regression model which regresses the observed variables on a smaller number of latent variables. As the covariance matrix of the latent variables is allowed to be non-diagonal, the correlations/covariances of the latent variables can be evaluated.

For example, for $j = 1, \cdots, p$,

$$y_j = \mu_j + \lambda_{j1}\omega_1 + \cdots + \lambda_{jq}\omega_q + \epsilon_j,$$

or

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\omega} + \boldsymbol{\epsilon}.$$

However, the effects of explanatory latent variables on key outcome latent variables of interest cannot be assessed by the CFA model of the first component. Hence, a second component is needed.

For example, for $j = 1, \cdots, q_1$,

$$\eta_j = \gamma_{j1}\xi_1 + \cdots + \gamma_{jq_2}\xi_{q_2} + \delta_j,$$

or

$$\boldsymbol{\eta} = \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\delta},$$

where $\boldsymbol{\omega} = (\boldsymbol{\eta}^T, \boldsymbol{\xi}^T)^T$.

This component is again a regression type model, in which the outcome latent variables are regressed on the explanatory latent variables. As a result, SEM is conceptually formulated by the familiar regression type model. However, because latent variables in the model are random, the standard technique in regression analysis cannot be applied to analyze SEMs.

Like most other statistical methods, the methodological developments of standard SEMs depend on crucial assumptions. The most basic assumptions are:

(I) The regression model in the second component is based on a simple linear regression equation in which higher-ordered product terms (such as quadratic terms or interaction terms) cannot be assessed.

(II) The observed random variables are assumed to be continuous, independently and identically distributed as a normal distribution.

As these assumptions may not be valid in substantive research, they induce limitations in applying SEMs to the analysis of real data in relation to complex situations.

New models and statistical methods have been developed to relax various aspects of the crucial assumptions for better analyses of complex data structure in practical research. These include but are not limited to:

(1) nonlinear SEMs with covariates,

(2) SEMs with mixed continuous and discrete data,

(3) multilevel SEMs,

(4) mixture SEMs,

(5) SEMs with missing data,

(6) longitudinal SEMs,

(7) semiparametric SEMs,

(8) transformation SEMs.

A traditional method for analyzing SEMs is the covariance structure approach which focuses on fitting the covariance structure under the proposed model to the sample covariance matrix computed from the observed data. However, some serious difficulties may be encountered in many complex situations in which deriving the covariance structure or obtaining an appropriate sample covariance matrix for statistical inferences is difficult.

In this book, we will use the Bayesian approach, together with various efficient Markov chain Monte Carlo (MCMC) algorithms, to analyze the advanced SEMs. In formulating and fitting models, we emphasize on the raw individual random observations rather than the sample covariance matrix.

The Bayesian approach has several advantages.

1. The development of statistical methods is based on the first moment properties of the raw individual observations which are simpler than the second moment properties of the sample covariance matrix. Hence, it has potential to be applied to more complex situations.
2. It produces a direct estimation of latent variables, which cannot be obtained with classical methods.
3. In addition to the information that is available in the observed data, the Bayesian approach allows the use of genuine prior information for producing better results.
4. The Bayesian approach provides more easily assessable statistics for goodness-of-fit and model comparison, and also other useful statistics such as the mean and percentiles of the posterior distribution.
5. It can give more reliable results for small samples.

We will use several real data sets for motivating the models and for illustrating the proposed Bayesian methodologies. These data sets are related to the following studies (see Appendix 1.1):

(1) job and life satisfaction, work attitude, and other related social-political issues;

(2) effects of some phenotype and genotype explanatory latent variables on kidney disease for type 2 diabetic patients;

(3) quality of life related to residents of several countries, and related to stroke patients;

(4) the development and findings from an AIDS preventative intervention for Filipino commercial sex workers;

(5) the longitudinal characteristics of cocaine and polydrug use;

(6) the functional relationships between bone mineral density (BMD) and its observed and latent determinants for old men;

(7) the academic achievement and its influential factors for American youth.

Some general notations are given as in Table 1.1.

Table 1.1 Typical Notations

Symbols	Meaning
$\boldsymbol{\omega}$	Latent vector in the measurement equation.
$\boldsymbol{\eta}$	Outcome (dependent) latent vector in the structural equation.
$\boldsymbol{\xi}$	Explanatory (independent) latent vector in the structural equation.
$\boldsymbol{\epsilon}$, $\boldsymbol{\delta}$	Random vectors of measurement errors.
$\boldsymbol{\Lambda}$	Factor loading matrix in the measurement equation.
$\mathbf{B}$, $\boldsymbol{\Pi}$, $\boldsymbol{\Gamma}, \boldsymbol{\Lambda}_\omega$	Matrices of regression coefficients in the structural equation.
$\boldsymbol{\Phi}$	Covariance matrix of explanatory latent variables.
$\boldsymbol{\Psi}_\epsilon$, $\boldsymbol{\Psi}_\delta$	Diagonal covariance matrices of measurement errors, with diagonal elements $\psi_{\epsilon k}$ and $\psi_{\delta k}$, respectively.
$\alpha_{0\epsilon k}, \beta_{0\epsilon k}, \alpha_{0\delta k}, \beta_{0\delta k}$	Hyperparameters in the Gamma distributions of $\psi_{\epsilon k}$ and $\psi_{\delta k}$.
$\mathbf{R}_0$, ρ_0	Hyperparameters in the Wishart distribution related to the prior distribution of $\boldsymbol{\Phi}$.
$\boldsymbol{\Lambda}_{0k}$, $\mathbf{H}_{0yk}$	Hyperparameters in the multivariate normal distribution related to the prior distribution of the kth row of $\boldsymbol{\Lambda}$ in the measurement equation.
$\boldsymbol{\Lambda}_{0\omega k}$, $\mathbf{H}_{0\omega k}$	Hyperparameters in the multivariate normal distribution related to the prior distribution of the kth row of $\boldsymbol{\Gamma}$ in the structural equation.
$\mathbf{I}_q$	A $q \times q$ identity matrix. Sometimes we just use $\mathbf{I}$ to denote an identity matrix if its dimension is clear.