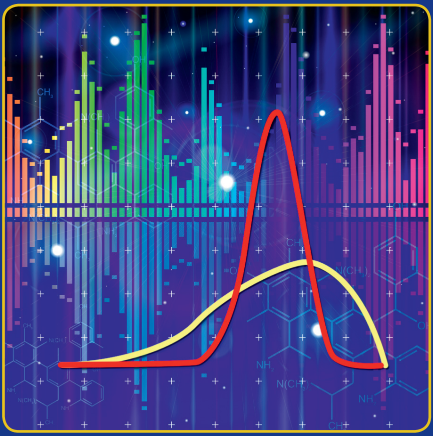
WILEY

Basic and Advanced Bayesian Structural Equation Modeling

With Applications in the Medical and Behavioral Sciences



Xin-Yuan Song • Sik-Yum Lee

WILEY SERIES IN PROBABILITY AND STATISTICS

Basic and Advanced Bayesian Structural Equation Modeling

WILEY SERIES IN PROBABILITY AND STATISTICS

Basic and Advanced Bayesian Structural Equation Modeling

With Applications in the Medical and Behavioral Sciences

Xin-Yuan Song and Sik-Yum Lee

Department of Statistics, The Chinese University of Hong Kong

WILEY
A John Wiley & Sons, Ltd., Publication

This edition first published 2012

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloging-in-Publication Data

Lee, Sik-Yum.
Basic and Advanced Bayesian Structural Equation Modeling: With Applications in the Medical and Behavioral Sciences / Sik-Yum Lee and Xin-Yuan Song.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-66952-5 (hardback)
1. Structural equation modeling. 2. Bayesian statistical decision theory. I. Song, Xin-Yuan. II. Title.
QA278.3.L439 2012
519.5′3–dc23

2012012199

A catalogue record for this book is available from the British Library.

ISBN: 978-0-470-66952-5

Typeset in 10/12pt Times by Aptara Inc., New Delhi, India

For our family members:

Yulin and Haotian Wu;
Mable, Anna, and Timothy Lee

Contents

About the authors

Xin-Yuan Song is an associate professor at the Department of Statistics of the Chinese University of Hong Kong. She earned her PhD in Statistics at the Chinese University of Hong Kong. She serves as an associate editor for *Psychometrika*, and as a member of the editorial board of *Frontiers in Quantitative Psychology and Measurement* and the *Open Journal of Statistics*. Her research interests are in structural equation models, latent variable models, Bayesian methods, survival analysis, and statistical computing. She has published over 80 papers in prestigious international journals.

Sik-Yum Lee is an emeritus professor of statistics at the Chinese University of Hong Kong. He earned his PhD in biostatistics at the University of California, Los Angeles. He received a distinguished service award from the International Chinese Statistical Association, is a former president of the Hong Kong Statistical Society, and is an elected member of the International Statistical Institute and a Fellow of the American Statistical Association. He had served as an associate editor for *Psychometrika* and *Computational Statistics & Data Analysis*, and as a member of the editorial board of the *British Journal of Mathematical and Statistical Psychology*, *Structural Equation Modeling*, and the *Chinese Journal of Medicine*. His research interests are in structural equation models, latent variable models, Bayesian methods, and statistical diagnostics. He is editor of the *Handbook on Structural Equation Models* and author of *Structural Equation Modeling: A Bayesian Approach*, and over 160 papers.

Preface

Latent variables that cannot be directly measured by a single observed variable are frequently encountered in substantive research. In establishing a model to reflect reality, it is often necessary to assess various interrelationships among observed and latent variables. Structural equation models (SEMs) are well recognized as the most useful statistical model to serve this purpose. In past years, even the standard SEMs were widely applied to behavioral, educational, medical, and social sciences through commercial software, such as AMOS, EQS, LISREL, and M*plus*. These programs basically use the classical covariance structure analysis approach. In this approach, the hypothesized covariance structure of the observed variables is fitted to the sample covariance matrix. Although this works well for many simple situations, its performance is not satisfactory in dealing with complex situations that involve complicated data and/or model structures.

Nowadays, the Bayesian approach is becoming more popular in the field of SEMs. Indeed, we find that when coupled with data augmentation and Markov chain Monte Carlo (MCMC) methods, this approach is very effective in dealing with complex SEMs and/or data structures. The Bayesian approach treats the unknown parameter vector $\boldsymbol{\theta}$ in the model as random and analyzes the posterior distribution of $\boldsymbol{\theta}$, which is essentially the conditional distribution of $\boldsymbol{\theta}$ given the observed data set. The basic strategy is to augment the crucial unknown quantities such as the latent variables to achieve a complete data set in the posterior analysis. MCMC methods are then implemented to obtain various statistical results. The book *Structural Equation Modeling: A Bayesian Approach*, written by one of us (Sik-Yum Lee) and published by Wiley in 2007, demonstrated several advantages of the Bayesian approach over the classical covariance structure analysis approach. In particular, the Bayesian approach can be applied to deal efficiently with nonlinear SEMs, SEMs with mixed discrete and continuous data, multilevel SEMs, finite mixture SEMs, SEMs with ignorable and/or nonignorable missing data, and SEMs with variables coming from an exponential family.

The recent growth of SEMs has been very rapid. Many important new results beyond the scope of *Structural Equation Modeling* have been achieved. As SEMs have wide applications in various fields, many new developments are published not only in journals in social and psychological methods, but also in biostatistics and statistical computing, among others. In order to introduce these useful developments to researchers in different fields, it is desirable to have a textbook or reference book that includes those new contributions. This is the main motivation for writing this book.

Similar to *Structural Equation Modeling*, the theme of this book is the Bayesian analysis of SEMs. Chapter 1 provides an introduction. Chapter 2 presents the basic concepts of standard SEMs and provides a detailed discussion on how to apply these models in practice. Materials in this chapter should be useful for applied researchers. Note that we regard the nonlinear SEM as a standard SEM because some statistical results for analyzing this model can be easily obtained through the Bayesian approach. Bayesian estimation and model comparison are discussed in Chapters 3 and 4, respectively. Chapter 5 discusses some practical SEMs, including models with mixed continuous and ordered categorical data, models with variables coming from an exponential family, and models with missing data. SEMs for analyzing heterogeneous data are presented in Chapters 6 and 7. Specifically, multilevel SEMs and multisample SEMs are discussed in Chapter 6, while finite mixture SEMs are discussed in Chapter 7. Although some of the topics in Chapters 3–7 have been covered by *Structural Equation Modeling*, we include them in this book for completeness. To the best of our knowledge, materials presented in Chapters 8–13 do not appear in other textbooks. Chapters 8 and 9 respectively discuss second-order growth curve SEMs and a dynamic two-level multilevel SEM for analyzing various kinds of longitudinal data. A Bayesian semiparametric SEM, in which the explanatory latent variables are modeled through a general truncated Dirichlet process, is introduced in Chapter 10. The purposes for introducing this model are to capture the true distribution of explanatory latent variables and to handle nonnormal data. Chapter 11 deals with SEMs with unordered categorical variables. The main aim is to provide SEM methodologies for analyzing genotype variables, which play an important role in developing useful models in medical research. Chapter 12 introduces an SEM with a general nonparametric structural equation. This model is particularly useful when researchers have no idea about the functional relationships among outcome and explanatory latent variables. In the statistical analysis of this model, the Bayesian P-splines approach is used to formulate the nonparametric structural equation. As we show in Chapter 13, the Bayesian P-splines approach is also effective in developing transformation SEMs for dealing with extremely nonnormal data. Here, the observed nonnormal random vector is transformed through the Bayesian P-splines into a random vector whose distribution is close to normal. Chapter 14 concludes the book with a discussion. In this book, concepts of the models and the Bayesian methodologies are illustrated through analyses of real data sets in various fields using the software WinBUGS, R, and/or our tailor-made C codes. Chapters 2–4 provide the basic concepts of SEMs and the Bayesian approach. The materials in the subsequent chapters are basically self-contained. To understand the material in this book, all that is required are some fundamental concepts of statistics, such as the concept of conditional distributions.

We are very grateful to organizations and individuals for their generous support in various respects. The Research Grant Council of the Hong Kong Special Administration Region has provided financial support such as GRF446609 and GRF404711 for our research and for writing this book. The World Value Study Group, World Values Survey, 1981–1984 and 1990–1993, the World Health Organization WHOQOL group, Drs. D. E. Morisky, J. A. Stein, J. C. N. Chan, Y. I. Hser, T. Kwok, H. S. Ip, M. Power and Y. T. Hao have been kind enough to let us have their data sets. The Department of Epidemiology and Public Health of the Imperial College, School of Medicine at St. Mary's Hospital (London, UK) provided their WinBUGS software. Many of our graduate students and research collaborators, in particular J. H. Cai, J. H. Pan, Z. H. Lu, P. F. Liu, J. Chen, D. Pan, H. J. He, K. H. Lam, X. N. Feng,

B. Lu, and Y. M. Xia, made very valuable comments which led to improvements to the book. We are grateful to all the wonderful people on the John Wiley editorial staff, particularly Richard Leigh, Richard Davies, Heather Kay, Prachi Sinha Sahay, and Baljinder Kaur for their continued assistance, encouragement and support of our work. Finally, we owe deepest thanks for our family members for their constant support and love over many years.

This book features an accompanying website:

www.wiley.com/go/medical_behavioral_sciences

1

Introduction

1.1 Observed and latent variables

Observed variables are those that can be directly measured, such as systolic blood pressure, diastolic blood pressure, waist–hip ratio, body mass index, and heart rate. Measurements from observed variables provide data as the basic source of information for statistical analysis. In medical, social, and psychological research, it is common to encounter latent constructs that cannot be directly measured by a single observed variable. Simple examples are intelligence, health condition, obesity, and blood pressure. 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; and waist–hip ratio and body mass index should be combined to evaluate obesity. 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), and 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 and BMI 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:

$$\text{PCr} = \alpha_1\text{SBP} + \alpha_2\text{DBP} + \alpha_3\text{BMI} + \alpha_4\text{WHR} + \alpha_5\text{HbA1c} + \alpha_6\text{FPG} + \epsilon_1, \tag{1.1}$$

$$\text{ACR} = \beta_1\text{SBP} + \beta_2\text{DBP} + \beta_3\text{BMI} + \beta_4\text{WHR} + \beta_5\text{HbA1c} + \beta_6\text{FPG} + \epsilon_2. \tag{1.2}$$

Basic and Advanced Bayesian Structural Equation Modeling: With Applications in the Medical and Behavioral Sciences,
First Edition. Xin-Yuan Song and Sik-Yum Lee.

From the estimates of the αs and βs, we can assess the effects of the explanatory variables on PCr and ACR. For example, based on the estimates of α_1 and β_1, we can evaluate the effects of SBP on PCr and ACR, respectively. However, this result cannot provide a clear and direct answer to the question about the effect of SBP on kidney disease. Similarly, the effects of other observed explanatory variables on kidney disease cannot be directly assessed from results obtained from regression analysis of equations (1.1) and (1.2). The deficiency of the regression model when applied to this study is due to the fact that kidney disease is a latent variable (construct) rather than an observed variable. A better approach is to appropriately combine PCr and ACR into a latent variable 'kidney disease (KD)' and regress this latent variable on the explanatory variables. Moreover, one may be interested in the effect of blood pressure rather than in the separate effects of SBP and DBP. Although the estimates of α_1 and α_2 can be used to examine the respective effects of SBP and DBP on PCr, they cannot provide a direct and clear assessment on the effect of blood pressure on PCr. Hence, it is desirable to group SBP and DBP together to form a latent variable that can be interpreted as 'blood pressure (BP)', and then use BP as an explanatory variable. Based on similar reasoning, {BMI, WHR} and {HbA1c, FPG} are appropriately grouped together to form latent variables that can be interpreted as 'obesity (OB)' and 'glycemic control (GC)', respectively. To study the effects of blood pressure, obesity, and glycemic control on kidney disease, we consider the following simple regression equation with latent variables:

$$\mathrm{KD} = \gamma_1 \mathrm{BP} + \gamma_2 \mathrm{OB} + \gamma_3 \mathrm{GC} + \delta. \tag{1.3}$$

This simple regression equation can be generalized to the multiple regression equation with product terms. For example, the following regression model can be used to assess the additional interactive effects among blood pressure, obesity, and glycemic control on kidney disease:

$$\begin{aligned} \mathrm{KD} = {} & \gamma_1 \mathrm{BP} + \gamma_2 \mathrm{OB} + \gamma_3 \mathrm{GC} + \gamma_4 (\mathrm{BP} \times \mathrm{OB}) + \gamma_5 (\mathrm{BP} \times \mathrm{GC}) \\ & + \gamma_6 (\mathrm{OB} \times \mathrm{GC}) + \delta. \end{aligned} \tag{1.4}$$

Note that studying these interactive effects by using the regression equations with the observed variables (see (1.1) and (1.2)) is extremely tedious.

It is obvious from the above simple example that incorporating latent variables in developing models for practical research is advantageous. First, it can reduce the number of variables in the key regression equation. Comparing equation (1.3) with (1.1) and (1.2), the number of explanatory variables is reduced from six to three. Second, 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 (1.1) or (1.2) does not exist in regression equation (1.3). Third, 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 (1.4). Hence, it is important to have a statistical method that simultaneously groups highly correlated observed variables into latent variables and assesses interrelationships among latent variables through a regression model of latent variables. This strong demand is the motivation for the development of structural equation models.

1.2 Structural equation model

The structural equation model (SEM) is a powerful multivariate tool for studying interrelationships among observed and latent variables. This statistical method is very popular in behavioral, educational, psychological, and social research. Recently, it has also received a great deal of attention in biomedical research; see, for example, Bentler and Stein (1992) and Pugesek *et al*. (2003).

The basic SEM, for example, the widely used LISREL model (Jöreskogand Sörbom, 1996), consists of 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 nondiagonal, the correlations/covariances of the latent variables can be evaluated. However, various effects of the explanatory latent variables on the key outcome latent variables of interest cannot be assessed by the CFA model of the first component. Hence, a second component is needed. This component is again a regression type model, in which the outcome latent variables are regressed on the explanatory latent variables. As a result, the SEM is conceptually formulated by the familiar regression type model. However, as latent variables in the model are random, the standard technique in regression analysis cannot be applied to analyze SEMs.

It is often important in substantive research to develop an appropriate model to evaluate a series of simultaneous hypotheses on the impacts of some explanatory observed and latent variables on the key outcome variables. Based on its particular formulation, the SEM is very useful for achieving the above objective. Furthermore, it is easy to appreciate the key idea of the SEM, and to apply it to substantive research; one only needs to understand the basic concepts of latent variables and the familiar regression model. As a result, this model has been extensively applied to behavioral, educational, psychological, and social research. Due to the strong demand, more than a dozen user-friendly SEM software packages have been developed; typical examples are AMOS, EQS6, LISREL, and Mplus. Recently, the SEM has become a popular statistical tool for biomedical and environmental research. For instance, it has been applied to the analysis of the effects of *in utero* methylmercury exposure on neurodevelopment (Sánchez *et al*., 2005), to the study of ecological and evolutionary biology (Pugesek *et al*., 2003), and to the evaluation of the interrelationships among latent domains in quality of life (e.g. Lee *et al*., 2005).

1.3 Objectives of the book

Like most other statistical methods, the methodological developments of standard SEMs depend on crucial assumptions. More specifically, the most basic assumptions are as follows: (i) The regression model in the second component is based on a simple linear regression equation in which higher-order product terms (such as quadratic terms or interaction terms) cannot be assessed. (ii) The observed random variables are assumed to be continuous, and independently and identically normally distributed. 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. Motivated by the need to overcome these limitations, the growth of SEMs has been very rapid in recent years. 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: nonlinear SEMs with covariates (e.g. Schumacker and Marcoulides, 1998; Lee and Song, 2003a); SEMs with mixed continuous, ordered and/or unordered categorical variables (e.g. Shi and Lee, 2000; Moustaki, 2003; Song and Lee, 2004; Song *et al*., 2007); multilevel SEMs (e.g. Lee and Shi, 2001; Rabe-Hesketh *et al*., 2004; Song and Lee, 2004; Lee and Song, 2005); mixture SEMs (e.g. Dolan and van der Maas, 1998; Zhu and Lee, 2001; Lee and Song, 2003b); SEMs with missing data (e.g. Jamshidian and Bentler, 1999; Lee and Tang, 2006; Song and Lee, 2006); SEMs with variables from exponential family distributions (e.g. Wedel and Kamakura, 2001; Song and Lee, 2007); longitudinal SEMs (Dunson, 2003; Song *et al*., 2008); semiparametric SEMs (Lee *et al*., 2008; Song *et al*., 2009; Yang and Dunson, 2010; Song and Lu, 2010); and transformation SEMs (van Montfort *et al*., 2009; Song and Lu, 2012). As the existing software packages in SEMs are developed on the basis of the covariance structure approach, and their primary goal is to analyze the standard SEM under usual assumptions, they cannot be effectively and efficiently applied to the analysis of the more complex models and/or data structures mentioned above. Blindly applying these software packages to complex situations has a very high chance of producing questionable results and drawing misleading conclusions.

In substantive research, data obtained for evaluating hypotheses of complex diseases are usually very complicated. In analyzing these complicated data, more subtle models and rigorous statistically methods are important for providing correct conclusions. In view of this, there is an urgent need to introduce into applied research statistically sound methods that have recently been developed. This is the main objective in writing this book. As we write, there has only been a limited amount of work on SEM. Bollen (1989) was devoted to standard SEMs and focused on the covariance structure approach. Compared to Bollen (1989), this book introduces more advanced SEMs and emphasizes the Bayesian approach which is more flexible than the covariance structure approach in handling complex data and models. Lee (2007) provides a Bayesian approach for analyzing the standard and more subtle SEMs. Compared to Lee (2007), the first four chapters of this book provide less technical discussions and explanations of the basic ideas in addition to the more involved, theoretical developments of the statistical methods, so that they can be understood without much difficulty by applied researchers. Another objective of this book is to introduce important models that have recently been developed and were not covered by Lee (2007), including innovative growth curve models and longitudinal SEMs for analyzing longitudinal data and for studying the dynamic changes of characteristics with respect to time; semiparametric SEMs for relaxing the normality assumption and for assessing the true distributions of explanatory latent variables; SEMs with a nonparametric structural equation for capturing the true general relationships among latent variables, and transformation SEMs for analyzing highly nonnormal data. We believe that these advanced SEMs are very useful in substantive research.

1.4 The Bayesian approach

A traditional method in 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. For simple SEMs, when the underlying distribution of the observed data is normal, this approach works fine with reasonably large sample sizes. 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.

Thanks to recent advances in statistical computing, such as the development of various efficient Markov chain Monte Carlo (MCMC) algorithms, the Bayesian approach has been extensively applied to analyze many complex statistical models. Inspired by its wide applications in statistics, we will use the Bayesian approach to analyze the advanced SEMs that are useful for medical and social-psychological research. Moreover, in formulating and fitting the model, we emphasize the raw individual random observations rather than the sample covariance matrix. The Bayesian approach coupled with the formulation based on raw individual observations has several advantages. First, 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 the potential to be applied to more complex situations. Second, it produces a direct estimation of latent variables, which cannot be obtained with classical methods. Third, it directly models observed variables with their latent variables through the familiar regression equations; hence, it gives a more direct interpretation and can utilize the common techniques in regression such as outlier and residual analyses in conducting statistical analysis. Fourth, 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. Fifth, 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. Sixth, it can give more reliable results for small samples (see Dunson, 2000; Lee and Song, 2004). For methodological researchers in SEMs, technical details that are necessary in developing the theory and the MCMC methods are given in the appendices to the chapters. Applied researchers who are not interested in the methodological developments can skip those appendices. For convenience, we will introduce the freely available software WinBUGS (Spiegelhalter, *et al.*, 2003) through analyses of simulated and real data sets. This software is able to produce reliable Bayesian statistics including the Bayesian estimates and their standard error estimates for a wide range of statistical models (Congdon, 2003) and for SEMs (Lee, 2007).

1.5 Real data sets and notation

We will use several real data sets for the purpose of motivating the models and illustrating the proposed Bayesian methodologies. These data sets are respectively related to the studies about: (i) job and life satisfaction, work attitude, and other related social-political issues; (ii) effects of some phenotype and genotype explanatory latent variables on kidney disease for type 2 diabetic patients; (iii) quality of life for residents of several countries, and for stroke patients; (iv) the development of and findings from an AIDS preventative intervention for Filipina commercial sex workers; (v) the longitudinal characteristics of cocaine and polydrug use; (vi) the functional relationships between bone mineral density (BMD) and its observed and latent determinants for old men; and (vii) academic achievement and its influential factors for American youth. Some information on these data sets is given in Appendix 1.1.

In the discussion of various models and their associated statistical methods, we will encounter different types of observations in relation to observable continuous and discrete variables or covariates; unobservable measurements in relation to missing data or continuous measurements underlying the discrete data; latent variables; as well as different types of

Table 1.1 Typical notation.

Symbol	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.

parameters, such as thresholds, structural parameters in the model, and hyperparameters in the prior distributions. Hence, we have a shortage of symbols. If the context is clear, some Greek letters may serve different purposes. For example, α has been used to denote an unknown threshold in defining an ordered categorical variable, and to denote a hyperparameter in some prior distributions. Nevertheless, some general notation is given in Table 1.1.

Appendix 1.1 Information on real data sets

Inter-university Consortium for Political and Social Research (ICPSR) data

The ICPSR data set was collected in the World Values Survey 1981–1984 and 1990–1993 project (World Values Study Group, ICPSR Version). The whole data set consists of answers to a questionnaire survey about work attitude, job and family life, religious belief, interest in politics, attitude towards competition, etc. The items that have been used in the illustrative examples in this book are given below.

Thinking about your reasons for doing voluntary work, please use the following five-point scale to indicate how important each of the reasons below have been in your own case (1 is unimportant and 5 is very important).

V 62 Religious beliefs 1 2 3 4 5

During the past few weeks, did you ever feel . . . (Yes: 1; No: 2)

V 89 Bored 1 2
V 91 Depressed or very unhappy 1 2
V 93 Upset because somebody criticized you 1 2
V 96 All things considered, how satisfied are you with your life as a whole these days?
1 2 3 4 5 6 7 8 9 10
Dissatisfied Satisfied

Here are some aspects of a job that people say are important. Please look at them and tell me which ones you personally think are important in a job. (Mentioned: 1; Not Mentioned: 2)

V 99 Good Pay 1 2
V 100 Pleasant people to work with 1 2
V 102 Good job security 1 2
V 103 Good chances for promotion 1 2
V 111 A responsible job 1 2
V 115 How much pride, if any, do you take in the work that you do?
1 A great deal 2 Some 3 Little 4 None
V 116 Overall, how satisfied or dissatisfied are you with your job?
1 2 3 4 5 6 7 8 9 10
Dissatisfied Satisfied
V 117 How free are you to make decisions in your job?
1 2 3 4 5 6 7 8 9 10
Not at all A great deal
V 129 When jobs are scarce, people should be forced to retire early,
1 Agree, 2 Neither, 3 Disagree
V 132 How satisfied are you with the financial situation of your household?
1 2 3 4 5 6 7 8 9 10
Dissatisfied Satisfied

V 176 How important is God in your life? 10 means very important and 1 means not at all important.

1 2 3 4 5 6 7 8 9 10

V 179 How often do you pray to God outside of religious services? Would you say . . .

1 Often 2 Sometimes
3 Hardly ever 4 Only in times of crisis 5 Never

V 180 Overall, how satisfied or dissatisfied are you with your home life?

1 2 3 4 5 6 7 8 9 10
Dissatisfied Satisfied

V 241 How interested would you say you are in politics?

1 Very interested 2 Somewhat interested
3 Not very interested 4 Not at all interested

Now I'd like you to tell me your views on various issues. How would you place your views on this scale? 1 means you agree completely with the statement on the left, 10 means you agree completely with the statement on the right, or you can choose any number in between.

V 252

1 2 3 4 5 6 7 8 9 10

Individual should take more responsibility for providing for themselves.

The state should take more responsibility to ensure that everyone is provided for.

V 253

1 2 3 4 5 6 7 8 9 10

People who are unemployed should have to take any job available or lose their unemployment benefits.

People who are unemployed should have the right to refuse a job they do not want.

V 254

1 2 3 4 5 6 7 8 9 10

Competition is good. It stimulates people to work hard and develop new ideas.

Competition is harmful. It brings out the worst in people.

V 255

1 2 3 4 5 6 7 8 9 10

In the long run, hard work usually brings a better life.

Hard work doesn't generally bring success – it's more a matter of luck and connections.

Please tell me for each of the following statements whether you think it can always be justified, never be justified, or something in between.

V 296 Claiming government benefits which you are not entitled to

1 2 3 4 5 6 7 8 9 10
Never Always

V 297 Avoiding a fare on public transport

1 2 3 4 5 6 7 8 9 10

Never Always

V 298 Cheating on tax if you have the chance

1 2 3 4 5 6 7 8 9 10

Never Always

V 314 Failing to report damage you've done accidentally to a parked vehicle

1 2 3 4 5 6 7 8 9 10

Never Always

I am going to read out some statements about the government and the economy. For each one, could you tell me how much you agree or disagree?

V 336 Our government should be made much more open to the public

1 2 3 4 5 6

Agree Completely Disagree Completely

V 337 We are more likely to have a healthy economy if the government allows more freedom for individuals to do as they wish

1 2 3 4 5 6

Agree Completely Disagree Completely

V 339 Political reform in this country is moving too rapidly

1 2 3 4 5 6

Agree Completely Disagree Completely

Type 2 diabetic patients data

The data set was collected from an applied genomics program conducted by the Institute of Diabetes, the Chinese University of Hong Kong. It aims to examine the clinical and molecular epidemiology of type 2 diabetes in Hong Kong Chinese, with particular emphasis on diabetic nephropathy. A consecutive cohort of 1188 type 2 diabetic patients was enrolled into the Hong Kong Diabetes Registry. All patients underwent a structured 4-hour clinical and biochemical assessment including renal function measured by plasma creatine (PCr) and urinary albumin creatinine ratio (ACR); continuous phenotype variables such as systolic blood pressure (SBP), diastolic blood pressure (DBP), body mass index (BMI), waist–hip ratio (WHR), glycated hemoglobin (HbA1c), fasting plasma glucose (FPG), non-high-density lipoprotein cholesterol (non-HDL-C), lower-density lipoprotein cholesterol (LDL-C), plasma triglyceride (TG); and multinomial genotype variables such as beta-3 adrenergic receptor (ADRβ3), beta-2 adrenergic receptor SNP1 (ADRβ21), beta-2 adrenergic receptor SNP2 (ADRβ22), angiotensin converting enzyme (DCP1 intro 16 del/ins (DCP1)), and angiotensin II receptor type 1 AgtR1 A1166C (AGTR1).

WHOQOL-BREF quality of life assessment data

The WHOQOL-100 assessment was developed by the WHOQOL group in 15 international field centers for the assessment of quality of life (QOL). The WHOQOL-BREF instrument is a short version of WHOQOL-100 consisting of 24 ordinal categorical items selected from the 100 items. This instrument was established to evaluate four domains: physical health, mental health, social relationships, and environment. The instrument also includes two ordinal categorical items for overall QOL and health-related QOL, giving a total of

26 items. All of the items are measured on a 5-point scale (1 = ‘not at all/very dissatisfied’; 2 = ‘a little/dissatisfied’; 3 y ‘moderate/neither’; 4 = ‘very much/satisfied’; 5 = ‘extremely/very satisfied’). The frequencies of the ordinal scores of the items are as follows:

	Ordinal score					Number of
WHOQOL item	1	2	3	4	5	incomplete obs.
Q1 Overall QOL	3	41	90	233	107	1
Q2 Overall health	32	127	104	154	58	0
Q3 Pain and discomfort	21	65	105	156	127	1
Q4 Medical treatment dependence	21	57	73	83	239	2
Q5 Energy and fatigue	15	57	166	111	118	8
Q6 Mobility	16	36	58	120	243	2
Q7 Sleep and rest	28	87	95	182	83	0
Q8 Daily activities	7	73	70	224	100	1
Q9 Work capacity	19	83	88	191	91	3
Q10 Positive feelings	2	30	141	241	59	2
Q11 Spirituality/personal beliefs	13	45	149	203	61	4
Q12 Memory and concentration	4	40	222	184	21	4
Q13 Bodily image and appearance	9	46	175	137	106	2
Q14 Self-esteem	13	72	130	210	50	0
Q15 Negative feelings	4	54	137	239	39	2
Q16 Personal relationships	8	46	68	218	134	1
Q17 Sexual activity	25	55	137	149	76	33
Q18 Social support	2	23	84	228	136	2
Q19 Physical safety and security	2	25	193	191	62	2
Q20 Physical environment	4	29	187	206	43	6
Q21 Financial resources	27	56	231	105	54	2
Q22 Daily life information	5	27	176	194	70	3
Q23 Participation in leisure activity	10	99	156	163	47	0
Q24 Living conditions	9	27	53	235	151	0
Q25 Health accessibility and quality	0	17	75	321	61	1
Q26 Transportation	8	38	61	253	113	2

AIDS preventative intervention data

The data set was collected from female commercial sex workers (CSWs) in 95 establishments (bars, night clubs, karaoke TV and massage parlours) in cities in the Philippines. The whole questionnaire consists of 134 items on areas of demographic knowledge, attitudes, beliefs, behaviors, self-efficacy for condom use, and social desirability. The primary concern is finding an AIDS preventative intervention for Filipina CSWs. Questions are as follows:

(1) How much of a threat do you think AIDS is to the health of people?
no threat at all/very small/moderate/strong/very great