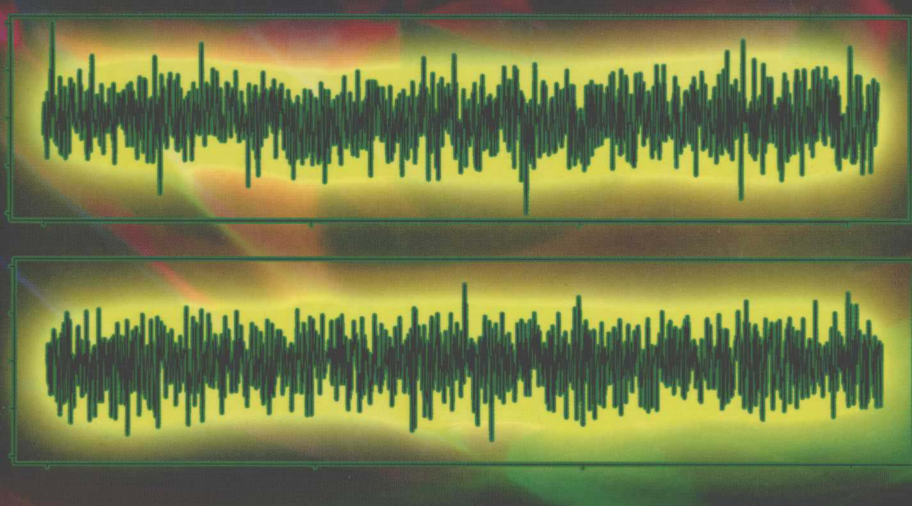
WILEY

Structural Equation Modeling

A Bayesian Approach



WILEY SERIES IN PROBABILITY AND STATISTICS

Structural Equation Modeling

A Bayesian Approach

Sik-Yum Lee
Department of Statistics
Chinese University of Hong Kong

BICENTENNIAL 1807 WILEY 2007 BICENTENNIAL

John Wiley & Sons, Ltd

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

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

Library of Congress Cataloging in Publication Data

British Library Cataloguing in Publication Data

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

ISBN-13 978-0-470-02423-2(HB)

Typeset in 10/12pt Galliard by Integra Software Services Pvt. Ltd, Pondicherry, India
Printed and bound in Great Britain by TJ International, Padstow, Cornwall
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

For Mable Lee and Timothy Lee

About the Author

Sik-Yum Lee is a professor of statistics at the Chinese University of Hong Kong. He earned his Ph.D. in biostatistics at the University of California, Los Angeles, USA. 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 serves as Associate Editor for *Psychometrika* and *Computational Statistics & Data Analysis*, and as a member of the Editorial Board of *British Journal of Mathematical and Statistical Psychology*, *Structural Equation Modeling*, *Handbook of Computing and Statistics with Applications* and *Chinese Journal of Medicine*. His research interests are in structural equation models, latent variable models, Bayesian methods and statistical diagnostics. He is editor of *Handbook of Latent Variable and Related Models* and author of over 140 papers.

Preface

Substantive theory usually involves observed and latent variables. An observed variable can be directly measured through a single measurement, such as income or systolic blood pressure. Latent variables are those that cannot be directly measured. Usually, one has to use several observed variables to assess the characteristic of a latent variable. It is very easy to give examples of latent variables in behavioral, biological, educational, medical, psychological and social sciences. For instance, in medical science, obesity is a latent variable that should be assessed by body mass, waist and hip indexes; lipid is a latent variable that is better assessed by non-high density lipoprotein, low density lipoprotein and triglyceride; blood pressure is another latent variable that should be measured by both systolic and diastolic blood pressures. Basically, structural equation models (SEMs) are regression models with observed and latent variables. For example, the influential LISREL model is composed of two simple regression equations: the measurement equation that relates the observed variables to the latent variables, and the structural equation that relates the endogenous latent variables to other latent variables. Due to the contributions of many psychometricians, including but not limited to Karl Jöreskog and Peter Bentler, and their LISREL and EQS6 programs, SEMs have been extensively applied not only to behavioral, educational and social science, but also to biological and medical sciences in the last quarter of a century.

The excellent book of Bollen (1989) on standard SEMs was published more than 15 years ago. Despite the widespread use of the standard SEMs, and the rapid growth in the new developments of nonlinear SEMs, two-level SEMs, and mixtures of SEMs, as well as SEMs for more complex data structures, such as dichotomous, ordered categorical, binary and missing data, there are very few new reference/textbooks in the field and there have been very few practical applications of the aforementioned recent developments to substantive research. This unexpected phenomenon is the motivation for writing this book, to provide a reference for researchers in various disciplines, and a textbook for graduate students in statistics, biostatistics and psychometrics. My main purpose is to introduce a Bayesian approach for developing efficient and rigorous statistical methodologies in SEMs and applying them to practical problems. Given the importance of latent variables and the popularity of the regression models, I hope that this book will help to promote SEMs to become a mainstream in statistics and psychometrics, and stimulate more applications.

The theme of this book is on the Bayesian analysis of SEMs. An introduction is given in Chapter 1, and some standard models are briefly discussed in Chapter 2. Chapter 3 presents the basic asymptotic theory for the traditional maximum likelihood (ML) and generalized least squares (GLS) approaches in analyzing the covariance structure of the model. The reason for including these nonBayesian materials is to provide a rigorous technical background related to the statistical results that are given in a large number of commercial software packages. For example, I present detailed proof on the consistency and asymptotic normality of the estimators, as well as the asymptotic chi-square distribution of the goodness-of-fit statistic. As far as I know, these kind of technical materials cannot be found in existing books on SEMs. Chapter 4 gives a detailed introduction to Bayesian estimation, whereas Chapter 5 treats Bayesian model comparison through the Bayes factor. Materials provided in these two core chapters can be applied to standard SEMs and their generalizations. Chapters 6 to 13 present the descriptions of the Bayesian approach as applied to SEMs with ordered categorical variables, SEMs with dichotomous variables, nonlinear SEMs, two-level nonlinear SEMs, multisample SEMs, mixtures of SEMs, SEMs with missing data and SEMs with variables from an exponential family of distributions. The Bayesian methodologies are illustrated by analyses of real examples in various fields via our tailor-made programs and/or the general software WinBUGS. The fundamental material in these chapters is self-contained. However, for generality and for enhancing wider applications, certain sections in the chapters discuss combinations of models. These sections depend on some of the material presented in previous chapters.

This book requires an understanding of some fundamental concepts in statistics. In particular, the concept of conditional distributions is necessary to understand the key ideas of the posterior distribution and the sampling-based computational methods. It does not need a knowledge of factor analysis or SEMs, but such a knowledge will enhance the appreciation of the Bayesian methodologies. It requires some basic knowledge of convergence in probability and convergence in distribution to understand the materials in Chapter 3. However, other chapters are completely independent of this chapter. Readers who do not have an interest in the asymptotic theory of the traditional ML and GLS approaches may skip Chapter 3. I am pleased to receive any comments about the book via email at sylee@sta.cuhk.edu.hk.

I owe a great debt to organizations and individuals for the use of their data sets. These include: World Value Study Group, World Values Survey, 1981–1984 and 1990–1993, for allowing the use of the Inter-university Consortium for Political and Social Research (ICPSR) data set; the Faculty of Education and Hong Kong Institute of Educational Research, the Chinese University of Hong Kong for providing part of the data set in the Accelerated Schools for Quality Education (ASQE) Project; D. E. Morisky, J. A. Stein and their colleagues for providing the AIDS data set; and Juliana C. N. Chan for providing the dataset about patient's non-adherence. I am deeply grateful to the Medical Research Council Biostatistics Unit (Cambridge, UK), and the Department of Epidemiology and Public Health of the Imperial College School of Medicine at St. Mary's Hospital (London, UK) for providing the powerful software WinBUGS as an alternative program in obtaining the Bayesian solutions. Without the kind support from the above mentioned organizations and individuals, this book might not have existed. The Research Grants Council of the Hong Kong Special Administration Region and the Chinese University of Hong Kong have provided financial support for my research and the writing of this book.

This book owes much to many people. It is certainly a great pleasure to thank them individually for their help and influence in producing this book, which represents most of my recent work in SEMs. My advisor, R. I. Jennrich, taught me computational methods and deeply inspired me with his serious attitude toward research. P. M. Bentler led me to the exciting field of SEM; he has had a distinctive influence on me during my early work in this field. I am grateful to colleagues in the Department of Statistics, the Chinese University of Hong Kong. In particular, W.Y Poon has given me generous support in research and administration work; W. H. Wong introduced the idea of data augmentation and Markov chain Monte Carlo methods during his 3 year stay in our department; and X. Y. Song gave numerous constructive suggestions in many chapters and essentially wrote all the programs for analyzing the artificial and real examples. I am very fortunate to have excellent students and research assistants. Those to whom I am greatly indebted include S. J. Wang, J. Q. Shi, W. Zhang, H. T. Zhu, X. Y. Song, J. S. Fu, L. Xu, B. Lu, Y. M. Xia, N. S. Tang, Y. Li, F. Chen, J. H. Cai, C. T. Poon and Y. Zhou. Most of them read the manuscript and gave helpful comments. WinBUGS results were mostly obtained by B. Lu. A number of people tackled the tedious task of typing the manuscript and drawing the path diagrams. For their excellent work and their patience with my handwriting, I would like to thank K. H. Leung and E. L. S. Tam. I am grateful to all the wonderful people on the John Wiley editorial staff, particularly Kelly Board, Lucy Bryan, Wendy Hunter, Simon Lightfoot, Kathryn Sharples, and Vidya Vijayan, and their design team for their continued assistance, encouragement and support of my work. Last and most important, I owe deepest thanks to Mable Lee and Timothy Lee for their constant understanding, support, encouragement and love that greatly release me from the pressure of the observed and latent variables.

SIK-YUM LEE
Department of Statistics
Chinese University of Hong Kong
Shatin, NT,
Hong Kong

Contents

1
Introduction

1.1 STANDARD STRUCTURAL EQUATION MODELS

In behavioural, educational, medical, and social sciences, substantive theory usually involves two kinds of variables, namely manifest (observed) and latent variables. Manifest variables are those that can be measured directly, such as income, test scores, systolic blood pressure, diastolic blood pressure, weight or heart rate. All data are records of measurements from observed data. Very often, it is necessary to deal with latent variables that cannot be directly measured by a single manifest variable. Examples are intelligence, personality, quantitative ability, anxiety, buying behavior, blood pressure and health condition. In practice, the characteristics of a latent variable can be partially measured by a linear combination of some manifest variables. For example, the quantitative ability of secondary school students can be reflected by their test scores in mathematics, physics and chemistry; the blood pressure of a patient can be measured by systolic and diastolic blood pressures. In most substantive research, it is important to establish an appropriate model to evaluate a series of simultaneous hypotheses about the impacts of latent variables and manifest variables on the other variables, and take the measurement errors into account. Structural equation models (SEMs) are well recognized as the most important statistical method to serve the above purpose. SEMs can be applied to many fields. For example, they can be applied to market research for establishing interrelationships between demand and supply, and the attitude and behaviour of the customers; to environmental science for investigating how health is affected by air and water pollution; to education for measuring the growth of intelligence and its relationship to personality and school environment; and to medicine for analyzing quality of life data and/or examining the impacts of

---

Structural Equation Modeling: A Bayesian Approach S-Y. Lee

physicians' concern, social influence and cognition on patients' adherence to medication.

The standard SEM, in particular the LISREL model (Jöreskog and Sörbom, 1996), is composed of two components. The first component is a confirmatory factor analysis model which relates the latent variables to all their corresponding manifest variables (indicators) and takes the measurement errors into account. This component can be regarded as a regression model which regresses the manifest variables with a small number of latent variables. The second component is again a regression type structural equation which regresses the endogenous (dependent) latent variables with the linear terms of some endogenous and exogenous (independent) latent variables. As latent variables are random, they cannot be directly analyzed by techniques in ordinary regression that are based on raw observations. However, conceptually, SEMs are formulated by the familiar regression type model, hence they are easy to apply in practice.

1.2 COVARIANCE STRUCTURE ANALYSIS

For standard SEMs, the covariance matrix of the manifest random vector $\mathbf{y}$ contains all the unknown parameters in the model. Hence, the classical methods for analyzing standard SEMs focused on the sample covariance matrix $\mathbf{S}$ and not the raw individual random vectors $\mathbf{y}_i$. This involves the formulation of the covariance structure $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, which is a matrix function of the unknown parameters vector $\boldsymbol{\theta}$; estimation of $\boldsymbol{\theta}$ by minimizing (or maximizing) some objective functions that measure the discrepancy between $\mathbf{S}$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, such as the maximum likelihood (ML) function or the generalized least square (GLS) function; and the derivation of asymptotic goodness-of-fit statistics for assessing whether $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ fits $\mathbf{S}$. As this kind of analysis emphasizes the population covariance matrix and the sample covariance matrix, it is often called covariance structure analysis. Today, more than a dozen user-friendly SEM software packages have been developed on the basis of covariance structure analysis approach with the sample covariance matrix. Typical examples are LISREL, EQS6 and AMOS. The covariance structure analysis approach depends heavily on the asymptotic normality of $\mathbf{S}$, either in defining the objective function or in deriving the asymptotic properties for statistical inferences. When the distribution of the manifest random vector $\mathbf{y}_i$ is multivariate normal and the sample size is reasonably large, the asymptotic distribution of $\mathbf{S}$ accurately approximates to the claimed multivariate normal distribution, and as a result this approach works fine. However, under slightly more complex situations that are common in substantive research, the covariance structure analysis approach on the basis of $\mathbf{S}$ is not effective and may encounter theoretical and computational problems.

It is well-recognized that estimating nonlinear terms (particularly the interaction term) among latent variables in the structural equation is an important issue in behavioral, social and psychological science (see Kenny and Judd, 1984; Bagozzi, Baumgartner and Yi, 1992). Due to the presence of the nonlinear terms of latent variables, the endogenous latent variables and the related manifest variables in $\mathbf{y}_i$ are not normally distributed. Hence, the sample covariance matrix of the raw sample observations is inadequate for modeling the nonlinear relationships. The product indicator approach (Jöreskog and Yang, 1996; Marsh, Wen and Hau, 2004) that artifically added products of indicators to $\mathbf{y}_i$ and used the sample covariance matrix of the enlarged manifest random vector for analysis, cannot provide a satisfactory method to cope with the problem (see Lee, Song, and Poon, 2004). For dichotomous or ordered categorical data, the sample covariance matrix of the raw sample data cannot be used. The multistage estimation procedures in LISREL or EQS produced estimates that are less optimal than the exact ML estimates and cannot be applied to analyze nonlinear terms of latent variables. For missing data that have a small number of observations within some missing patterns, the covariance structure analysis approach would also encounter serious difficulties because the sample covariance matrices corresponding to those patterns could be singular. The degree of difficulty is further compounded with a large number of missing patterns in the data set, or the missing data are missing with a nonignorable missing mechanism. For hierarchical data, the individual observations are correlated; this induces a problem in the covariance structure analysis with the sample covariance matrix.

In view of the above discussion, it is clear that although the covariance structure analysis approach based on the sample covariance matrix works well for the standard SEMs under the normality assumption, it cannot be applied to more complex models or data structures that are commonly encountered in substantive research. To develop sound statistical methods for those complex situations, it is necessary to develop better statistical methods which are based on the individual observations and their basic model, rather than the sample covariance matrix.

1.3 WHY A NEW BOOK?

In the past few years, the growth of SEM has been very rapid. New models and statistical methods have been developed for better analyses of more complex data structures in substantive research. These include but are not limited to: (i) SEMs with dichotomous and/or ordered categorical data (Shi and Lee, 1998, 2000; Moustaki, 2003; Rabe-Hesketh, Skrondal and Pickles, 2004; among others); (ii) nonlinear SEMs (Kenny and Judd, 1984; Klein and Moosbrugger, 2000; Lee and Song, 2003a; Wall and Amemiya, 2000; among others); (iii) linear or nonlinear SEMs with covariates (Lee and Shi, 2000; Moustaki, 2003; Song and Lee, 2006a; among others); (iv) two-level or multilevel SEMs (Lee and Shi, 2001; Ansari and Jedidi, 2000; Rabe-Hesketh, Skrondal and Pickles, 2004; Song and Lee, 2004; among others); (v) multisample SEMs (Song and Lee, 2001, 2002b); (vi) mixtures of SEMs (Jedidi, Jagpal and DeSarbo, 1997; Arminger, Stein and Wittenberg, 1999; Dolan and van der Maas, 1998; Zhu and Lee, 2001; Lee and Song, 2003b; among others); (vii) SEMs with missing data that are missing at random or with a nonignorable mechanism (Jamshidian and Bentler, 1999; Song and Lee, 2002a; Lee and Song, 2004a; Lee and Tang, 2006a; Song and Lee, 2006b; among others); and (viii) SEMs with variables from the exponential family distributions (Wedel and Kamakura, 2001; Rabe-Hesketh, Skrondal and Pickles, 2004; Lee and Tang, 2006b; Song and Lee, 2006b; among others). The above articles not only provide theoretical results, but also have significant practical value. For instance, it is very common in practice to encounter ordered categorical data with missing data, hierarchical data and/or heterogeneous data, and hence developments of sound statistical methods to cope with such practical situations are useful.

The primary goal of all the existing commercial software packages in SEMs is for analyzing the standard SEMs under the normal assumption. They cannot effectively and efficiently analyze the more complex models and/or data structures mentioned above. At the moment, there are only a limited number of reference/textbooks in SEM. Moreover, the emphasis of all the existing books, for example Bollen (1989), was devoted to the standard SEMs, and focused on the covariance structure analysis approach. Hence, despite the widespread use of SEMs, and the importance of the aforementioned complex models for sound and rigorous analyses of real data sets, there have been very few practical applications of the recent developments of SEMs to substantive research. The limited applications are not due to a lack of relevant substantive applications that required such models and their associative statistical methods. Rather, the main reasons are that the applied researchers are not familiar with these models and methods, and the existing commercial SEM software cannot produce satisfactory solutions for coping with the complex situations.

Therefore, there is a need for a new reference/textbook for the second generation of SEM which involves a much wider class of SEMs that include the standard SEMs and their useful generations. This book should provide a more appropriate approach than the covariance structure analysis approach in analyzing the general class of SEMs, together with a dependable software for obtaining reliable and rigorous results for statistical inference.

1.4 OBJECTIVES OF THE BOOK

One of the basic objectives of this book is to propose a Bayesian approach for analyzing some useful structural equation models and/or data structures that are commonly encountered in substantive research. This includes the treatments of the standard SEMs and their useful generalizations in various ways. More specifically, the generalizations are SEMs with ordered categorical variables, SEMs with dichotomous variables, nonlinear SEMs, two-level SEMs, multi-sample SEMs, mixtures of SEMs, SEMs with ignorable and/or nonignorable missing data, SEMs with variables from the exponential family distributions, and some of their combinations. In formulating various SEMs, and in developing the Bayesian methods, the emphasis is placed on the raw individual random observations rather than on the sample covariance matrix. This formulation has several advantages. First, the development of statistical methods is based on the first moment properties of the raw individual observations which is simpler than the second moment properties of the sample covariance matrix. Hence, it is easier to apply in more complex situations. Second, it leads to a direct estimation of the latent variables which is better than the classical regression method or the Bartlett's method for obtaining the factor score estimates. Third, as it directly models manifest variables with their latent variables through the familiar regression equations, it gives a more direct interpretation and can utilize the common techniques in regression such as outlier and residual analyses in conducting statistical analysis. The advantages of a Bayesian approach are that it allows the use of genuine prior information in addition to the information that is available in the observed data for producing better results, provides useful statistics such as the mean and percentiles of the posterior distribution of the unknown parameters, and gives more reliable results for small samples (see Dunson, 2000; Lee and Song; 2004b; Scheines, Hoijtink and Boomsma, 1999).

The next aim is to describe the technique of data augmentation (Tanner and Wong, 1987), and introduce some efficient tools in statistical computing for analyzing SEMs. The key idea of data augmentation is to augment the observed data with the latent quantities, which could be the latent variables or the unobservable data (missing data or the unobservable continuous measurements that underlie the dichotomous and/or ordered categorical data), so that the Bayesian analysis is feasible with the complete-data set. The introduced tools in statistical computing include some Markov chain Monte Carlo (MCMC) methods, such as the Gibbs sampler (Geman and Geman, 1984) and the Metropolis – Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970) and path sampling (Gelman and Meng, 1998). The strategy of data augmentation followed by MCMC methods will be used repeatedly to analyze complex SEMs and data structures throughout this book. The Bayesian methodologies will be demonstrated through real examples in the fields of education, management, medicine, psychology and sociology.

One of the main goals is the introduction of the freely available software WinBUGS (Spiegelhalter, Thomas, Best and Lunn, 2003) to the field of SEM. This software is able to produce reliable Bayesian statistics including the Bayesian