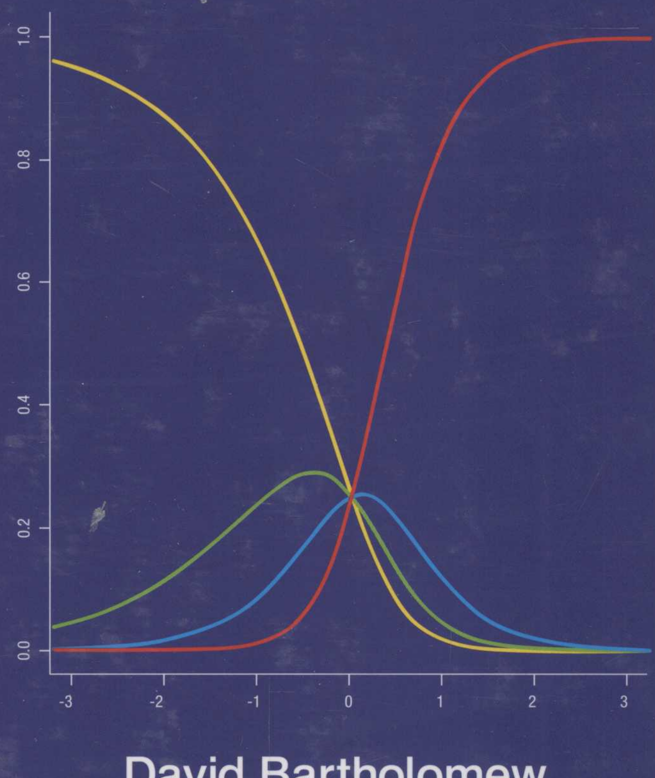
WILEY

Latent Variable Models and Factor Analysis

A Unified Approach

3rd Edition



David Bartholomew

Martin Knott • Irini Moustaki

WILEY SERIES IN PROBABILITY AND STATISTICS

Latent Variable Models and Factor Analysis

A Unified Approach

3rd Edition

David Bartholomew · Martin Knott · Irini Moustaki
London School of Economics and Political Science, UK

WILEY
A John Wiley & Sons, Ltd., Publication

This edition first published 2011

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.

Library of Congress Cataloging-in-Publication Data

Bartholomew, David J.
Latent variable models and factor analysis : a unified approach. – 3rd ed. / David Bartholomew, Martin Knott, Irini Moustaki.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-97192-5 (cloth)
1. Latent variables. 2. Latent structure analysis. 3. Factor analysis. I. Knott, M. (Martin) II. Moustaki, Irini. III. Title.
QA278.6.B37 2011
519.5′35–dc22

2011007711

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

Print ISBN: 978-0-470-97192-5
ePDF ISBN: 978-1-119-97059-0
oBook ISBN: 978-1-119-97058-3
ePub ISBN: 978-1-119-97370-6
Mobi ISBN: 978-1-119-97371-3

Set in 10/12pt Times by Aptara Inc., New Delhi, India.
Printed in Malaysia by Ho Printing (M) Sdn Bhd

WILEY SERIES IN PROBABILITY AND STATISTICS

Preface

It is more than 20 years since the first edition of this book appeared in 1987, and its subject, like statistics as a whole, has changed radically in that period. By far the greatest impact has been made by advances in computing. In 1987 adequate implementation of most latent variable methods, even the well-established factor analysis, was guided more by computational feasibility than by theoretical optimality. What was true of factor analysis was even more true of the assortment of other latent variable techniques, which were then seen as unconnected and very specific to different applications. The development of new models was seriously inhibited by the insuperable computational problems which they would have posed. This new edition aims to take full account of these changes.

The Griffin series of monographs, then edited by Alan Stuart, was designed to consolidate the literature of promising new developments into short books. Knowing that one of us (DJB) was attempting to develop and unify latent variable modelling from a statistical point of view, he proposed what appeared in 1987 as Volume 40 in the Griffin series. Ten years later the series had been absorbed into the Kendall Library of Statistics monographs designed to complement the evergreen volumes of Kendall and Stuart's *Advanced Theory of Statistics*. *Latent Variable Models and Factor Analysis* took its place as Volume 7 in that series in 1999. This second edition was somewhat different in character from its predecessor, and a second author (MK) brought his particular expertise into the project. After a further decade that book was in urgent need of revision, and this could only be done adequately by recruiting a third author (IM) who is actively involved at the frontiers of contemporary research. Throughout its long history the principal aim has remained unchanged and it is worth quoting at some length from the Preface of the second edition:

> the prime object of the book remains the same – that is, to provide a unified and coherent treatment of the field from a statistical perspective. This is achieved by setting up a sufficiently general framework to enable us to derive the commonly used models, and many more as special cases. The starting point is that all variables, manifest and latent, continuous or categorical, are treated as random variables. The subsequent analysis is then done wholly within the realm of the probability calculus and the theory of statistical inference.

The subtitle, added in this edition, merely serves to emphasise, rather than modify its original purpose.

Chapter 1 covers the same ground as before, but the order of the material has been changed. The aim of the revision is to provide a more natural progression of ideas from the most elementary to the more advanced.

Chapters 2 and 3, as before, are the heart of the book. Chapter 2 provides an overall treatment of the basic model together with an account of general questions of inference relating to it. It introduces what we call the general linear latent variable model (GLLVM) from which almost all of the models considered later in the book are derived as special cases. An important new feature is an introductory account of Markov chain Monte Carlo (MCMC) methods for parameter estimation. These are a good example of the computer-intensive methods which the growth in the power of computers has made possible. In principle, these methods are now capable of handling any of the models in this book and a general introduction is given in this chapter, leaving more detailed treatment until later.

In Chapter 3 the general model is specialised to the normal linear factor model. This includes traditional factor analysis, which is probably the most thoroughly studied and widely applied latent variable model. Little directly relevant research has appeared since the second edition, but our treatment has been revised and this chapter will serve as a source for the basic theory, much of which is now embodied in computer software.

Latent trait models are widely used, especially in educational testing, but they have a far wider field of application, as the examples in Chapter 4 show. The chapter begins with two versions of the model and then discusses the statistical methods available for their implementation. Although the traditional estimation methods, based on likelihood, are efficient and are present in the standard software, we have also taken the opportunity to demonstrate the MCMC method in some detail in a situation where it can easily be compared with established methods. There is no intention here to suggest that its use is limited to such relatively simple examples. On the contrary, this example is designed to illustrate the potential of the MCMC method in a broader context.

Chapters 5 and 7 extend the ideas into newer areas, particularly where ordered categorical variables are involved. A number of the models appeared for the first time in earlier editions. This work has been consolidated here and, now that computing is no longer a barrier, they should find wider application. Latent class models are often seen as among the simpler latent variable models, and in the first edition they appeared much earlier in the book. Here they appear in Chapter 6 where it can be seen more easily, perhaps, how they fit in to the broader scheme.

Chapter 8, on relationships between latent variables, has been supplemented by an account of methods of estimation and goodness-of-fit in the LISREL model, but otherwise is unchanged, apart from the transfer to Chapter 9 of some material noted below.

Chapter 9 is entirely new except for the inclusion of a little material from the old Chapter 8 which now fits more naturally in its new setting. It draws attention to a number of methods, especially principal components analysis, which serve much the same purpose as latent variable models but come from a different statistical tradition.

The examples are an important part of the text. They are intended not only to illustrate the mechanics of putting the theory into practice but they also bring to light many subtleties which are not immediately evident from the formal derivations. This is especially important in latent variable modelling where questions of interpretation need to be explored in numerical terms for their full implications to be appreciated. Many of the original examples have been retained because, although the data on which they are based are now necessarily older, it is the point that the examples make which is important. Where we felt that these could not be bettered, they have been retained. But, in some cases, we have replaced original examples and added new ones where we felt that an improvement could be made. However, all the original examples have been recalculated using the newer software described in the Appendix.

There was a website linked to the second edition which has been discontinued. There are two reasons for this. First, we have provided an appendix to this book which gives details of the more comprehensive software that is currently available: the new appendix has removed the need for the individual programs provided on the original website. Secondly, it is now much easier to find numerical examples on which the methods can be tried out. One convenient source is in Bartholomew *et al.* (2008) and its associated website, where there are extensive data sets and some of the methods are described in a form more suitable for users.

Acknowledgements

Alan Stuart died in 1998, but his encouragement and support in getting the first edition off the ground, when latent variable models were often viewed by statisticians with suspicion, if not hostility, still leave the statistical community in his debt.

Much of the earlier editions remains, as does our debt to those who contributed to them: Lilian de Menezes, Panagiota Tzamourani, Stephen Wood, Teresa Albanese and Brian Shea, all once at the London School of Economics. Fiona Steele read a draft of the new Chapter 9 and her comments have materially helped the exposition.

The anonymous advice garnered by our publisher, John Wiley, for this edition was invaluable both in encouraging us to proceed and in defining the changes and additions we have made.

We extensively used the IRTPRO software for producing output for the factor analysis model for categorical variables. The authors of the software, Li Cai, Stephen du Toit and David Thissen, have kindly provided us with a free version of the software, and Li Cai in particular helped us resolve any software-related questions. We would also like to thank Jay Magidson and Jeroen Vermunt for their help with Latent Gold and Albert Maydeu-Olivares for sharing with us the UK data on Eysenck's Personality Questionnaire–Revised.

The material relating to Sir Godfrey Thomson's work in Chapter 9 was covered in much greater detail in a research project at the University of Edinburgh in which one of us (DJB) was a principal investigator. References to relevant publications arising from the project are included here. This project was financed as part of research supported by the Economic and Social Research Council, grant no. RES-000-23-1246.

David J. Bartholomew
Martin Knott
Irini Moustaki
London School of Economics and Political Science
January 2011

Contents

1

Basic ideas and examples

1.1 The statistical problem

Latent variable models provide an important tool for the analysis of multivariate data. They offer a conceptual framework within which many disparate methods can be unified and a base from which new methods can be developed. A statistical model specifies the joint distribution of a set of random variables and it becomes a latent variable model when some of these variables – the latent variables – are unobservable. In a formal sense, therefore, there is nothing special about a latent variable model. The usual apparatus of model-based inference applies, in principle, to all models regardless of their type. The interesting questions concern why latent variables should be introduced into a model in the first place and how their presence contributes to scientific investigation.

One reason, common to many techniques of multivariate analysis, is to reduce dimensionality. If, in some sense, the information contained in the interrelationships of many variables can be conveyed, to a good approximation, in a much smaller set, our ability to 'see' the structure in the data will be much improved. This is the idea which lies behind much of factor analysis and the newer applications of linear structural models. Large-scale statistical enquiries, such as social surveys, generate much more information than can be easily absorbed without drastic summarisation. For example, the questionnaire used in a sample survey may have 50 or 100 questions and replies may be received from 1000 respondents. Elementary statistical methods help to summarise the data by looking at the frequency distributions of responses to individual questions or pairs of questions and by providing summary measures such as percentages and correlation coefficients. However, with so many variables it may still be difficult to see any pattern in their interrelationships. The fact that our ability to visualise relationships is limited to two or three dimensions places us under strong pressure to reduce the dimensionality of the data in a manner which preserves as much of the structure as possible. The reasonableness of such a course is often

Latent Variable Models and Factor Analysis: A Unified Approach, Third Edition.
David Bartholomew, Martin Knott and Irini Moustaki.

evident from the fact that many questions overlap in the sense that they seem to be getting at the same thing. For example, one's views about the desirability of private health care and of tax levels for high earners might both be regarded as a reflection of a basic political position. Indeed, many enquiries are designed to probe such basic attitudes from a variety of angles. The question is then one of how to condense the many variables with which we start into a much smaller number of indices with as little loss of information as possible. Latent variable models provide one way of doing this.

A second reason is that latent quantities figure prominently in many fields to which statistical methods are applied. This is especially true of the social sciences. A cursory inspection of the literature of social research or of public discussion in newspapers or on television will show that much of it centres on entities which are handled as if they were measurable quantities but for which no measuring instrument exists. Business confidence, for example, is spoken of as though it were a real variable, changes in which affect share prices or the value of the currency. Yet business confidence is an ill-defined concept which may be regarded as a convenient shorthand for a whole complex of beliefs and attitudes. The same is true of quality of life, conservatism, and general intelligence. It is virtually impossible to theorise about social phenomena without invoking such hypothetical variables. If such reasoning is to be expressed in the language of mathematics and thus made rigorous, some way must be found of representing such 'quantities' by numbers. The statistician's problem is to establish a theoretical framework within which this can be done. In practice one chooses a variety of indicators which can be measured, such as answers to a set of yes/no questions, and then attempts to extract what is common to them.

In both approaches we arrive at the point where a number of variables have to be summarised. The theoretical approach differs from the pragmatic in that in the former a pre-existing theory directs the search and provides some means of judging the plausibility of any measures which result. We have already spoken of these measures as indices or hypothetical variables. The usual terminology is *latent variables* or *factors*. The term *factor* is so vague as to be almost meaningless, but it is so firmly entrenched in this context that it would be fruitless to try to dislodge it now. We prefer to speak of latent variables since this accurately conveys the idea of something underlying what is observed. However, there is an important distinction to be drawn. In some applications, especially in economics, a latent variable may be real in the sense that it could, in principle at least, be measured. For example, personal wealth is a reasonably well-defined concept which could be expressed in monetary terms, but in practice we may not be able or willing to measure it. Nevertheless we may wish to include it as an explanatory variable in economic models and therefore there is a need to construct some proxy for it from more accessible variables. There will be room for argument about how best to do this, but wide agreement on the existence of the latent variable. In most social applications the latent variables do not have this status. Business confidence is not something which exists in the sense that personal wealth does. It is a summarising concept which comes prior to the indicators of it which we measure. Much of the philosophical debate which takes place on latent variable models centres on *reification*; that is, on speaking as though such things as quality of life and business confidence were real entities in the sense that length and weight are. However, the usefulness and validity of the methods to be described in this book do not depend primarily on whether one adopts a realist or an instrumentalist view of latent variables. Whether one regards the latent variables as existing in some real world or merely as a means of thinking economically about complex relationships, it is possible to use the methods for prediction or establishing relationships *as if* the theory were dealing with real entities. In fact, as we shall see, some methods, which appear to be purely empirical, lead their users to behave as if they had adopted a latent variable model. We shall return to the question of interpreting latent variables at the end of Chapter 9. In the meantime we note that an interesting discussion of the meaning of a latent variable can be found in Sobel (1994).

1.2 The basic idea

We begin with a very simple example which will be familiar to anyone who has met the notion of spurious correlation in elementary statistics. It concerns the interpretation of a 2×2 contingency table. Suppose that we are presented with Table 1.1. Leaving aside questions of statistical significance, the table exhibits an association between the two variables. If A was being a heavy smoker and B was having lung cancer someone might object that the association was spurious and that it was attributable to some third factor C with which A and B were both associated – such as living in an urban environment. If we go on to look at the association between A and B in the presence and absence of C we might obtain data as set out in Table 1.2. The original association has now vanished and we therefore conclude that the underlying variable C was wholly responsible for it. Although the correlation between the manifest variables might be described as spurious, it is here seen as pointing to an underlying latent variable whose influence we wish to determine.

Even in the absence of any suggestion about C it would still be pertinent to ask whether the original table could be decomposed into two tables exhibiting independence. If so, we might then look at the members of each subgroup to see if they had anything in common, such as most of one group living in an urban environment. The idea can be extended to a p-way table and again we can enquire whether it can be decomposed into sub-tables in which the variables are independent. If this were possible there would be grounds for supposing that there was some latent categorisation which fully explained the original association. The discovery of such a

Table 1.1 A familiar example.

	A	$\bar{A}$	Total
B	350	200	550
$\bar{B}$	150	300	450
	500	500	1000

Table 1.2 Effect of a hidden factor.

	C			$\bar{C}$		
	A	$\bar{A}$	Total	A	$\bar{A}$	Total
B	320	80	400	30	120	150
$\bar{B}$	80	20	100	70	280	350
	400	100	500	100	400	500

decomposition would amount to having found a latent categorical variable for which conditional independence held. The validity of the search does not require the assumption that the goal will be reached. In a similar fashion we can see how two categorical variables might be rendered independent by conditioning on a third continuous latent variable. We now illustrate these rather abstract ideas by showing how they arise with two of the best-known latent variable models.

1.3 Two examples

1.3.1 Binary manifest variables and a single binary latent variable

We now take the argument one step further by introducing a probability model for binary data. In order to do this we shall need to anticipate some of the notation required for the more general treatment given below. Thus suppose there are p binary variables, rather than two as in the last example. Let these be denoted by $x_1, x_2, \ldots, x_p$ with $x_i = 0$ or 1 for all i. Let us consider whether the mutual association of these variables could be accounted for by a single binary variable y. In other words, is it possible to divide the population into two parts so that the xs are mutually independent in each group? It is convenient to label the two hypothetical groups 1 and 0 (as with the xs, any other labelling would serve equally well). The prior distribution of y will be denoted $h(y)$, and this may be written

$$h(1) = P\{y = 1\} = \eta \quad \text{and} \quad h(0) = 1 - h(1). \tag{1.1}$$

The conditional distribution of x_i given y will be that of a Bernoulli random variable written

$$P\{x_i \mid y\} = \pi_{iy}^{x_i}(1 - \pi_{iy})^{1-x_i} \quad (x_i, y = 0, 1), \tag{1.2}$$

where π_{iy} is the probability that $x_i = 1$ when the latent class is y. Notice that in this simple case the form of the distributions h and $P\{x_i \mid y\}$ is not in question; it is only their parameters, η, $\{\pi_{i0}\}$ and $\{\pi_{i1}\}$ which are unspecified by the model.