A review of *Latent Variable Modeling with R*

DL Oberski

Dept of Methodology & Statistics, Tilburg University, The Netherlands

Author Note

Thanks are due to Minjeong Jeon and Jesper Tijmstra for their comments on earlier versions of this review. This work was supported by the Netherlands Organization for Scientific Research (NWO) [Veni grant number 451-14-017].

A review of *Latent Variable Modeling with R*

Latent variable modeling refers to a class of models that includes factor analysis, structural equation modeling (SEM), growth curve modeling, item response theory (IRT), and latent class analysis (LCA) – all staple techniques in educational and psychological research. Increasingly these techniques are treated in a common "latent variable" framework. However, statistics textbooks that take this integrated approach, such as Skrondal and Rabe-Hesketh (2004), may be somewhat inaccessible for applied researchers and graduate students in these fields and do not contain examples to use as a starting point. A textbook that provides a bird's eye view of these techniques and shows how to apply them in R (R Core Team, 2014) is therefore a very welcome addition to the literature.

Latent Variable Modeling with R (Finch & French, 2015) aims to be such a contribution. According to the publisher's website[1], the book is

> "intended for use in graduate or advanced undergraduate courses in latent variable modeling, factor analysis, structural equation modeling, item response theory, measurement, or multivariate statistics taught in psychology, education, human development, and social and health sciences, researchers in these fields also appreciate this book's practical approach. The book provides sufficient conceptual background information to serve as a standalone text. Familiarity with basic statistical concepts is assumed but basic knowledge of R is not."

The excellent *Latent variable modeling using R: A step-by-step guide* (Beaujean, 2014), also published by Routledge, has a similar remit, but limits itself to SEM with continuous and categorical variables, omitting latent class (mixture) models. The Finch and French (2015) book has a slightly wider scope by including mixture models. It also discusses IRT models using the language and statistics common in that field, whereas Beaujean discusses these as categorical data SEM's. Finch & French's IRT perspective will be an advantage for some researchers–for instance those in the field of educational assessment.

[1] `https://www.routledge.com/products/9780415832458`

The book is organized by model, covering exploratory (chapter 2) and confirmatory (chapter 3) factor analysis, structural equation modeling (chapters 5 and 6), growth curve modeling (chapter 7), mixture modeling (chapter 8), and IRT (chapters 9 and 10). The first chapter introduces a few R commands, whereas the last chapter aims to demonstrate how to simulate data for Monte Carlo and power studies. In each chapter, the model is explained conceptually, an applied example using data from psychology and education is analyzed using various R packages, and the output is discussed in some detail in a didactic manner. Data, as well as Microsoft Word files containing R code and output are provided on the publisher's website[2].

Specific content of the book

The first chapter shows how data can be read into R and how to remove missings. The reader is not shown explicitly how to read in the SPSS data files provided on the website, although the appropriate function is mentioned.

Chapter 2 discusses EFA, using `factanal` and `fa` from the **psych** library (Revelle, 2014). The model is introduced using linear algebra. Methods to determine the likely number of factors, including parallel analysis, are demonstrated.

Chapter 3 is about CFA and uses **lavaan** (Rosseel, 2012). Again linear algebra is used to introduce the model, and various fit measures are discussed. An example is then analyzed. The authors have a preference for the DWLS estimation method rather than the more standard robust ML; in my view this choice is somewhat ideosyncratic but defensible.

Chapter 4 introduces structural equation modeling. Path models with latent variables are fit and model comparison is demonstrated didactically.

Chapter 5 discusses multiple group SEM and invariance testing. Contrary to most of the literature on invariance testing (e.g. Davidov, Schmidt, & Billiet, 2011; Steenkamp & Baumgartner, 1998), the authors use the term "invariance" to mean "metric invariance", and "fully invariant" in this book includes all latent variable means and (co)variances. Generally the

[2] `https://www.routledge.com/products/9780415832458`

discussion of the model comparisons and their interpretation is clear.

Chapter 6 discusses models with feedback loops ("nonrecursive models" in SEM parlance) using **systemfit** (Henningsen & Hamann, 2007), interactions in SEM, and SEM trees using **OpenMx** (Boker et al., 2015) and **semtree** (Brandmaier & Prindle, 2014; Brandmaier, von Oertzen, McArdle, & Lindenberger, 2013). This inclusion of more advanced and relatively modern techniques is laudable and has been missing from the literature. Nonrecursive SEM's are exclusively discussed as instrumental variable models, which is one approach, but not the most commonly used one in my perception. Again, though, this choice is certainly defensible.

Chapter 7 on growth curve modeling is concise and clear. Model estimation, and interpretation are demonstrated, as well as average growth curve plotting using **ggplot2** (Wickham, 2009). It would have been nice to see how individual growth curves can be plotted as well.

Chapter 8 discusses latent class models using **poLCA** (Linzer & Lewis, 2011) and mixture regression using **flexmix** (Gruen, Leisch, & Sarkar, 2013). The applied example is clearly presented and easy to follow. The authors recommend cross-validation in their summary on p. 175 but do not explain in the text how this can be done in R.

Chapters 9 and 10 discuss parametric and nonparametric IRT using the R packages **psychometric** (Fletcher, 2010), **ltm** (Rizopoulos, 2006), **mirt** (Chalmers, 2012), **difR** (Magis, Beland, & Raiche, 2015), **mokken** (van der Ark, 2007, 2012), and **KernSmoothIRT** (Mazza, Punzo, & McGuire, 2014). Chapter 9 discusses how to fit Rasch, 1PL, 2PL, 3PL, (G)PCM, and GRM models, and how to obtain IIC and IIF plots. Model fit evaluation in the form of omnibus tests and item fit is also discussed. Chapter 10 discusses methods to assess unidimensionality, DIF, mokken scaling, and kernel smoothing IRT.

The final chapter is a nice idea: it discusses how to perform Monte Carlo studies for the models in the other chapters. This may help readers perform simulation studies and power analysis. Simulations are exemplified for a SEM, IRT, and LCA model. The simulation code for SEM models uses self-programmed R code rather than existing packages for SEM simulation.

The simulation code for **ltm** and **poLCA** uses those packages' built-in functions.

Cautionary notes

The set up of the book is sensible and it will certainly contain useful information for those interested in latent variable modeling in R. Unfortunately, however, this first edition of the book is mired by a number of issues. Since some readers will benefit from guidance regarding these issues, some of them are listed in this section.

Throughout. The R packages used in the book are, for the most part, not cited. Readers who wish to use these packages can find information on proper citation under `https://stat.ethz.ch/R-manual/R-patched/library/utils/html/citation.html`.

Chapter 1. There is a syntax error on p. 3 (quotes are missing).

Chapter 2. There is some confusion as to whether Σ is the correlation or the covariance matrix on p. 12. Another syntax error can be found on p. 19 (variable names cannot start with numbers in R). In addition, the output formatting is jumbled so that the reader cannot tell which numbers belong to which columns (one example is on p. 34).

Chapter 3. On p. 53 models with different numbers of factors are compared using a chi-square difference test, which is not the correct reference distribution[3] (e.g. Andrews, 2001). Equation (3.4) is missing a bracket.

Chapter 4. The model is introduced in Equation (4.1) as $\eta = B\gamma + \zeta$ rather than the usual $\eta = B\eta + \Gamma\xi + \zeta$ (e.g. Bollen, 1989). Thus, it omits any effects between endogenous latent variables (η's), although such effects are actually used in later example models. I also found the use of γ for the latent exogenous variable confusing since this usually denotes a regression coefficient. Page 75 incorrectly asserts that "mathematically a covariance that is bidirectional behaves in the same fashion as a direct path that is unidirectional", which is not true in general: it is possible to generate non-equivalent models by replacing a bidirectional path with a

[3] The problem is that models with differing numbers of factors are nested by fixing some parameters (e.g. latent variable variances) to their boundary values. This means that a regularity condition for the usual chi-square distribution of the likelihood ratio does not hold.

undirectional one in SEM[4]. I also found it potentially confusing for students that the term "essentially identical [model fit]" is used to indicate both that two models are equivalent (have *identical* fit) and that one model does not fit much worse than another (have *similar* fit).

Chatper 5. A notational inconsistency is introduced: x instead of y is now used as observed indicator, and later on in the same chapter x changed to mean a covariate (Equation 5.5 on p. 101). On p. 85, the notation makes it seem that the random variable ϵ is confused with its variance (it is the variance that is restricted to be equal, not the variable itself). On p. 93 a model is tested with equal intercepts, residual variances, and latent variable (co)variances but free loadings, which is not a standard approach (e.g. Davidov et al., 2011; Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2000).

Chapter 6. Equations (6.5) and (6.7) have a notational inconsistency (B is assigned a new meaning twice). The definition of a nonrecursive model as a model in which "some of the relationships between the latent variables are bidirectional" (glossary on p. 314) is not that employed by other textbooks on SEM (e.g. Bollen, 1989), since this definition excludes cyclical models without bidirectional relationships such as $x \longrightarrow y \longrightarrow z$, which are generally also considered nonrecursive.

Chapter 8. The model is introduced in Equation (8.1) which has incorrect subscripts (these should not be $1, 2, ..., j$ but the values of the indicator variables). Equation (8.2) is missing the conditioning on the covariate z and uses the idiosyncratic notation $(\pi_t^Y | z_k)$ to mean $P(Y = t|Z = z)$ or $\pi_t^{Y|Z}{}_z$ in "categorical data" notation (Agresti, 2002). The regression mixture model in Equation (8.3) is incorrect[5]. Note that contrary to the book's notation, most texts use X for the latent variable and Y for the observed dependent variable.

[4]For example, $y_1 \rightarrow y_2 \rightarrow y_3$ is not equivalent to $y_1 \rightarrow y_2 \leftrightarrow y_3$ though both have 1 df.

[5]It says

$$\pi^{X_1 X_j .., Y}_{1,\ j...,\ t} = \sum_t^T \pi_t^T (m_i = \beta_0 + \beta_1 f_{1i} + \epsilon_i),$$

which is not a regression but a model for the joint distribution of multivariate X and Y, and omits the conditioning on Z, the distribution of the dependent variable, and the dependency on the mixture.

Chapter 9. Eq. (9.1) has an error (j should be J), as does Equation (9.7) (Pearson's chi-square but omits the square). There are several notational inconsistencies in the Equations; for example, p sometimes is the number of parameters and sometimes a probability, and subscripts are omitted or introduced.

Chapter 10. Formatting issues make some output difficult to read (e.g. p. 241).

Chapter 11. Unfortunately, the authors appear to have been unaware of the existence of the **simsem** package (Pornprasertmanit, Miller, & Schoemann, 2015), which would have obviated the need for the highly complex three-page simulation for a simple SEM on pp. 281–283. Moreover, this code contains a serious error: the `phi` variable is overwritten so that there is no error in the equation, contrary to the authors' stated intention. This can also be seen in the results on page 289, which show `phi~~phi` is negative, whereas it should be close to 1. To me this demonstrates that the longer code and more complex code is, the more room there is for human error. Packages such as Pornprasertmanit et al. (2015)'s excellent **simsem** are therefore very useful, not just to save us time, but also to prevent such errors.

Conclusion

Latent variable modeling comprises an important set of techniques for a wide range of fields, including educational and behavioral statistics. A clear textbook demonstrating how such models can be fit in open source software is therefore a great idea. The authors have done a good job of selecting the methods to be discussed, and in providing some easy-to-follow explanations of R output. The book may be most appropriate for more experienced researchers who already know the models behind the techniques and are merely seeking to learn how to apply them in R. Graduate students and others seeking to learn about these techniques should beware of the issues with the current edition of the book listed in the previous section.

References

Agresti, A. (2002). *Categorical data analysis, 2nd ed*. New York: Wiley-Interscience.

Andrews, D. W. (2001). Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica*, 683–734.

Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*. New York: Routledge.

Boker, S. M., Neale, M. C., Maes, H. H., Wilde, M. J., Spiegel, M., Brick, T. R., ... Kirkpatrick, R. M. (2015). **OpenMx** user guide release 2.2.6 [Computer software manual]. Retrieved from `http://openmx.psyc.virginia.edu/documentation`

Bollen, K. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.

Brandmaier, A. M., & Prindle, J. J. (2014). **semtree**: Recursive partitioning for structural equation models [Computer software manual]. Retrieved from `http://www.brandmaier.de/semtree` (`R` package version 0.9.7.7)

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological methods*, *18*(1), 71.

Chalmers, R. P. (2012, 5). **mirt**: A multidimensional item response theory package for the `R` environment. *Journal of Statistical Software*, *48*(6), 1–29. Retrieved from `http://www.jstatsoft.org/v48/i06`

Davidov, E., Schmidt, P., & Billiet, J. (2011). *Cross-cultural analysis: Methods and applications*. New York: Routledge.

Finch, W. H., & French, B. F. (2015). *Latent variable modeling with R*. New York: Routledge.

Fletcher, T. D. (2010). **psychometric**: Applied psychometric theory [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=psychometric` (`R` package version 2.2)

Gruen, B., Leisch, F., & Sarkar, D. (2013). **flexmix**: Flexible mixture modeling [Computer software manual]. Retrieved from

http://CRAN.R-project.org/package=flexmix (R package version 2.3-13)

Henningsen, A., & Hamann, J. D. (2007). **systemfit**: A package for estimating systems of simultaneous equations in R. *Journal of Statistical Software*, *23*(4), 1–40. Retrieved from http://www.jstatsoft.org/v23/i04/

Linzer, D. A., & Lewis, J. B. (2011). **poLCA**: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, *42*(10), 1–29. Retrieved from http://www.jstatsoft.org/v42/i10/

Magis, D., Beland, S., & Raiche, G. (2015). **difR**: Collection of methods to detect dichotomous differential item functioning (DIF) [Computer software manual]. (R package version 4.6)

Mazza, A., Punzo, A., & McGuire, B. (2014, 6). **KernSmoothIRT**: An R package for kernel smoothing in item response theory. *Journal of Statistical Software*, *58*(6). Retrieved from http://www.jstatsoft.org/v58/i06

Pornprasertmanit, S., Miller, P., & Schoemann, A. (2015). **simsem**: SIMulated structural equation modeling [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=simsem (R package version 0.5-11)

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/ (ISBN 3-900051-07-0)

Revelle, W. (2014). **psych**: Procedures for psychological, psychometric, and personality research [Computer software manual]. Evanston, Illinois. Retrieved from http://CRAN.R-project.org/package=psych (R package version 1.4.8)

Rizopoulos, D. (2006). **ltm**: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. Retrieved from http://www.jstatsoft.org/v17/i05/

Rosseel, Y. (2012). **lavaan**: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and

implications. *Human Resource Management Review*, *18*(4), 210–222.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling : multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Steenkamp, J., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–107.

Vandenberg, R., & Lance, C. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70.

van der Ark, L. A. (2007, 2). Mokken scale analysis in `R`. *Journal of Statistical Software*, *20*(11), 1–19. Retrieved from `http://www.jstatsoft.org/v20/i11`

van der Ark, L. A. (2012, 5). New developments in Mokken scale analysis in `R`. *Journal of Statistical Software*, *48*(5), 1–27. Retrieved from `http://www.jstatsoft.org/v48/i05`

Wickham, H. (2009). ***ggplot2****: elegant graphics for data analysis*. New York: Springer.