Springer Proceedings in Mathematics & Statistics

L. Andries van der Ark
Marie Wiberg
Steven A. Culpepper
Jeffrey A. Douglas
Wen-Chung Wang *Editors*

Quantitative Psychology

The 81st Annual Meeting of the Psychometric Society, Asheville, North Carolina, 2016

Springer

Springer Proceedings in Mathematics & Statistics

Volume 196

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at http://www.springer.com/series/10533

L. Andries van der Ark • Marie Wiberg
Steven A. Culpepper • Jeffrey A. Douglas
Wen-Chung Wang
Editors

Quantitative Psychology

The 81st Annual Meeting of the Psychometric Society, Asheville, North Carolina, 2016

Springer

Editors
L. Andries van der Ark
Research Institute for Child
Development and Education
University of Amsterdam
Amsterdam, The Netherlands

Steven A. Culpepper
Department of Statistics
University of Illinois at Urbana-Champaign
Champaign, IL, USA

Wen-Chung Wang
Department of Psychology
The Educational University of Hong Kong
Hong Kong, China

Marie Wiberg
Department of Statistics, USBE
Umeå University, Umeå, Sweden

Jeffrey A. Douglas
Department of Statistics
University of Illinois at Urbana-Champaign
Champaign, IL, USA

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-319-56293-3 ISBN 978-3-319-56294-0 (eBook)
DOI 10.1007/978-3-319-56294-0

Library of Congress Control Number: 2017940525

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This volume represents presentations given at the 81st annual meeting of the Psychometric Society in Asheville, North Carolina, during July 11–15, 2016. The meeting, organized by the University of North Carolina at Greensboro, was one of the largest Psychometric Society meetings in the United States, both in terms of participants and number of presentations. It attracted 415 participants, with 204 papers being presented, along with 95 poster presentations, 3 pre-conference workshops, 3 keynote presentations, 6 invited presentations, 2 career-award presentations, a debate, 2 dissertation-award winners, 9 symposia, a trivial-pursuit lunch, and *Psychometrika'*s 80th anniversary celebration.

Since the 77th meeting in Lincoln, Nebraska, Springer publishes the proceedings volume from the annual meeting of the Psychometric Society so as to allow presenters to quickly make their ideas available to the wider research community while still undergoing a thorough review process. The first four volumes of the meetings in Lincoln, Arnhem, Madison, and Beijing were received successfully, and we expect a successful reception of these proceedings too.

We asked authors to use their presentation at the meeting as the basis of their chapters, possibly extended with new ideas or additional information. The result is a selection of 36 state-of-the-art chapters addressing a diverse set of topics, including item response theory, equating, classical test theory, factor analysis, structural equation modeling, dual scaling, multidimensional scaling, power analysis, cognitive diagnostic models, and multilevel models.

Amsterdam — L. Andries van der Ark
Umeå — Marie Wiberg
Urbana-Champaign, IL — Steven A. Culpepper
Urbana-Champaign, IL — Jeffrey A. Douglas
Hong Kong — Wen-Chung Wang

Contents

New Results on an Improved Parallel EM Algorithm for Estimating Generalized Latent Variable Models

Matthias von Davier

Abstract The second generation of a parallel algorithm for generalized latent variable models, including MIRT models and extensions, on the basis of the general diagnostic model (GDM) is presented. This new development further improves the performance of the parallel-E parallel-M algorithm presented in an earlier report by means of additional computational improvements that produce even larger gains in performance. The additional gain achieved by this second-generation parallel algorithm reaches factor 20 for several of the examples reported with a sixfold gain based on the first generation. The estimation of a multidimensional IRT model for large-scale data may show a larger reduction in runtime compared to a multiple-group model which has a structure that is more conducive to parallel processing of the E-step. Multiple population models can be arranged such that the parallelism directly exploits the ability to estimate multiple latent variable distributions separately in independent threads of the algorithm.

Keywords Parallel EM-algorithm • MIRT • Diagnostic modeling • Estimation • Latent variable modeling

1 Introduction

This chapter reports on the second generation of a parallel algorithm for generalized latent variable models on the basis of the general diagnostic model (von Davier 2005, 2008, 2014). This new development further improves the performance of the parallel-E parallel-M algorithm presented in an earlier report (von Davier 2016) by

The original version of this chapter was revised. An erratum to this chapter can be found at https://doi.org/10.1007/978-3-319-56294-0_37

This work was partially completed while the author was at the Educational Testing Service.

M. von Davier (✉)
National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA, 19104-3102, USA
e-mail: mvondavier@nbme.org

L.A. van der Ark et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 196, DOI 10.1007/978-3-319-56294-0_1

means of additional computational improvements that produce even larger gains in performance. The additional gain achieved by this second-generation parallel algorithm reaches factor 20 for several of the examples were reported with a sixfold gain based on the first generation. The estimation of a multidimensional IRT model for large-scale data may show a larger reduction in runtime compared to a multiple-group model which has a structure that is more conducive to parallel processing of the E-step. Multiple population models can be arranged such that the parallelism directly exploits the ability to estimate multiple latent variable distributions separately in independent threads of the algorithm.

This development allows estimation of advanced psychometric models for very large datasets in a matter of seconds or minutes, rather than hours. Unlike methods that rely on simplifications of the likelihood equations that are only available for a specific set of constrained problems such as bifactor models, the approach presented here is applicable to all types of multidimensional latent variable models, including multidimensional models, multigroup and mixture models, as well as growth curve and growth mixture models.

Parallel processing is now available in a number of compilers and hence found its way into software packages such as LatentGold, Mplus, and FlexMirt. While these packages allow users to utilize one or multiple cores, their documentation is somewhat limited. In the present report, the approach to parallelism is detailed at the level of algorithmic description, and the types of gains are exemplified based on a range of hardware platforms that are typically available as workstations or servers. Moreover, the software presented here is available for research purposes on all major operating systems, in particular, on Linux, Microsoft Windows, and Apple OS X platforms.

2 A General Latent Variable Model

The general latent variable model used in this evaluation of an improved algorithm for parallel processing is based on the general diagnostic model (GDM) (von Davier 2005). This family of models contains a large class of well-known psychometric approaches as special cases, including IRT, MIRT, latent class models, HYBRID models, and mixture models (von Davier 2008), as well as models for longitudinal data (von Davier et al. 2011) and several diagnostic models (von Davier 2014, 2016).

The probability of a correct item response $X = 1$ by a respondent from a population $C = c$ and with skill attribute pattern $a = (a_1, \ldots, a_k)$ on item i can be written as

$$P(X = 1|i, a, c) = \frac{\exp\left(\beta_{ic} + \sum_{k=1}^{K} \gamma_{ick} h(q_{ik}, a_k)\right)}{1 + \exp\left(\beta_{ic} + \sum_{k=1}^{K} \gamma_{ick} h(q_{ik}, a_k)\right)} \tag{1}$$

This is the general model introduced by von Davier (2005). The q_{ik} are indicator variables for $i = 1, \ldots, I$ and $k = 1, \ldots, K$ and are provided as an input. These

Q-matrix entries q_{ik} describe which of the skill attributes is required for which item. Note that Eq. (1) also contains a population indicator c, which makes it suitable both for multiple-group and mixture distribution models (von Davier and Rost 1995; von Davier and Yamamoto 2004; von Davier 2005, 2008; von Davier and Rost 2016). While the general model given in Eq. (1) served as the basis for the formal specification of the log-linear cognitive diagnostic model (L-CDM) (Henson, Templin and Willse 2009) and other developments for binary skill attributes and data, von Davier (2005, 2008) utilized the general form to derive the linear or partial-credit GDM:

$$P(X = x|i, a, c) = \frac{\exp\left(\beta_{\text{ixc}} + \sum_{k=1}^{K} x\gamma_{\text{ick}} h\left(q_{\text{ik}}, a_k\right)\right)}{1 + \sum_{i}^{m}\left(\beta_{\text{iyc}} + \sum_{k=1}^{K} y\gamma_{\text{ick}} h\left(q_{\text{ik}}, a_k\right)\right)} \tag{2}$$

Note that this leads to a model that contains located latent class models, multiple classification latent class models, IRT models, and multidimensional IRT models, as well as a compensatory version of the reparameterized unified model, as special cases (von Davier 2005). In addition, the linear GDM as well as the general family is suitable for binary, polytomous ordinal, and mixed-format item response data.

One common application of generalized latent variable models is the use for confirmatory analysis. In this case, a Q-matrix provides the required loading pattern, and constraints on the skill attribute space provide the structure of the model. One example is what is commonly called a multi-trait-multi-method model, in which each observed indicator variable is cross classified with respect to two different sets of latent variables. In the examples analyzed for this report, a seven-dimensional model of this type is analyzed, which contains two loadings for each item, one for a set of four latent variables (subdomains) and one for a set of three variables (processes). Figure 1 provides an illustration of this model used as example.

3 Method

The EM algorithm (Dempster et al. 1977) is one of the most frequently used approaches for estimating latent variable models (e.g., McLachlan and Krishnan 1997). The name of the algorithm stems from the alternating, iterative repetition of two steps, the E (expectation) step and the M (maximization) step. The estimation of generalized latent variable models using the EM algorithm requires the estimation of expected values for all required sufficient statistics of the structural parameters of the measurement model as well as the estimation of latent variable distributions in one or more populations. In the M-step, the expected values serve as sufficient statistics for the maximization of parameters. Parallel implementations of the EM algorithm have been used in image processing, in particular in Gaussian mixture modeling for some time (Cui et al. 2014; Cui et al. 2010; Das et al. 2007; Lopez de Teruel et al. 1999). In contrast to Gaussian mixtures, certain latent variable models require computationally more costly calculations in the M-step as well.

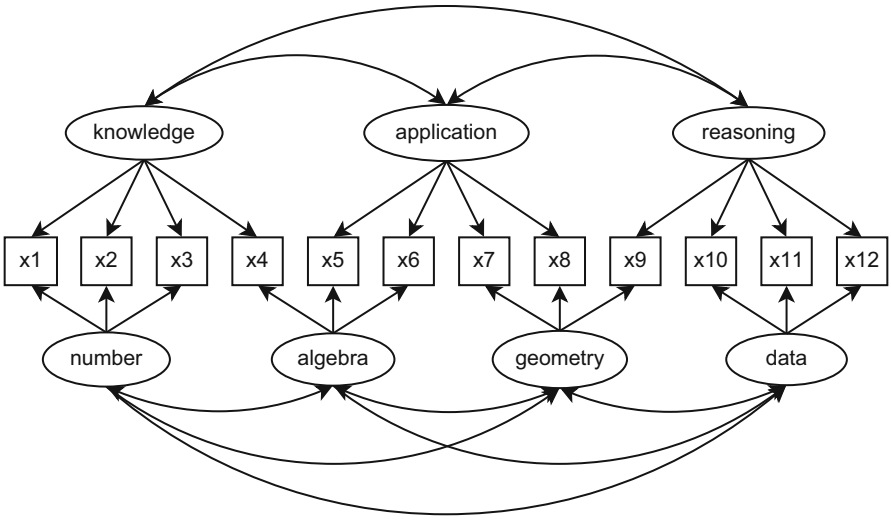

Fig. 1 Confirmatory multidimensional IRT model with seven dimensions. Three of the variables describe processing skills, and four variables describe subdomains of mathematics. While the figure uses on only 12 items, the real data example contains 214 items in a balanced incomplete block design administered to approx. 8000 students

While parallelizing the E-step is straightforward in terms of distribution of the work, the aggregation of the partial results obtained in distributed ways separately by each core is again a potentially costly calculation or aggregation process. A new algorithm was developed based on the first generation parallel-E parallel-M algorithm described in von Davier (2016). This new algorithm can be described as a parallel-E parallel-M algorithm with tiling-based aggregation of results. This new approach is based on three different phases of parallel execution of the necessary calculations:

1. Parallel E-step: Distributed calculation of expected counts for sufficient statistics
2. Tiling: Rearranging distributed latent variable space and parallel aggregation
3. Parallel M-Step: ML-estimation of parameters based on aggregated counts

Gains are largest for the parallelism introduced in part (1) that concerns the E-step by conducting estimation of expected counts separately in subsamples distributed across cores. The smallest gains are obtained by the conversion of the M-step to parallel execution in part (3). The aggregation step that follows the E-step in part (2) provides somewhat more advantages than the parallel M-step, either in the form of a multiple-group approach where aggregation can be completely avoided, or in the form of tiling, where the latent variable space aggregation is rearranged so that it can take place in parallel as well. Shared memory allocation of all latent variable distributions and rearranging the direction of aggregation are crucial in the process. More details about the different approaches utilized in version 1 can be

Table 1 Summary of example analyses used in the comparisons

Case	Scales	Model	Groups	Items	Sample	QPT	Total	Ncat
A	1	2PL/GPCM	312	293	1,614,281	21	21	2–4
B	1	2PL/GPCM	283	133	1,803,599	21	21	2–4
C	7	MTMM	1	214	7377	3	2187	3
D	2	MIRT	1	175	5763	15	225	3
E	2	MIRT	1	175	5763	31	961	3
F	NA	LCA	54	54	246,112	1	1	6
G	5	MIRT	1	150	2026	5	3125	2

The examples cover a wide range of latent variable models from IRT to MIRT, confirmatory models, and latent class models. The items are mixed format, and their number varies from 54 to 293; the number of respondents varies from 2026 to 1.8 million

found in von Davier (2016). The tiling process resulting in an improved ability to utilize parallelism in aggregation is detailed in Gan et al. (2009).

4 Data

Table 1 shows an overview of the test cases. All test cases reported here are based on sequential and parallel versions, except two additional ones, that were only run on the fastest hardware platform, and only in parallel mode, since sequential mode or running these on a laptop would take unacceptably long periods of time. The test cases are from typical applications of generalized latent variable models, ranging from IRT, to classification of respondents by means of a latent class analysis, to MIRT applications with 2, 5, and 7 dimensions, and finally multiple population IRT for linking in large-scale international assessments with approximately 300 populations and 2,000,000 test takers distributed across these populations.

5 Results

Table 2 shows the results for a (somewhat older) Dual-CPU 12-Core Intel Xeon workstation running at 3.46 GHZ per core. These are given for the sequential algorithm, running only on a single core, as well as for parallel-E parallel-M version 1 and the improved version 2 of the PE-PM algorithm, running on all available cores.

Table 3 shows the results for a 4-CPU AMD Opteron (Piledriver architecture) server with 64/32 cores, running at 2.6 GHZ per core. This architecture offers 64 integer arithmetic cores, with each CPU offering 16 integer units that share eight floating point units. In this sense, we see a performance that is more reflective of 32 FPUs, but with some added capacity for caching and pre-fetching and integer processing. The measures are given for the sequential algorithm, running only on a

Table 2 Results of the comparison of parallel-E parallel-M versions 1 and 2 on a 12-Core Xeon workstation as well as the sequentially executed algorithm

		Parallel V1		Parallel V2		
Case	Iterations	Likelihood	Sec.	Sec.	Speedup	Sec.
A	126	−14,547,731.24	1356	168	807%	153
B	112	−14,639,728.20	1127	117	963%	96
C	165	−125,200.51	2465	314	785%	343
D	76	−14,468,510.90	44	11	400%	11
E	86	−14,468,485.63	1155	163	708%	145
F	1028	−1,234,570.30	7444	1039	716%	964
G	277	−130,786.39	2499	949	263%	726

Table 3 Results of the comparison of parallel-E parallel-M versions 1 and 2 on a 64/32-Core AMD Piledriver server as well as the sequentially executed algorithm

		Parallel V1		Parallel V2		
Case	Iterations	Likelihood	Sec.	Sec.	Speedup	Sec.
A	126	−14,547,731.24	2074	256	810%	114
B	112	−14,639,728.20	1553	185	839%	90
C	165	−125,200.51	6131	889	689%	300
D	76	−14,468,510.90	116	21	552%	5
E	86	−14,468,485.63	1945	150	1296%	116
F	1028	−1,234,570.30	6427	377	1704%	227
G	277	−130,786.39	6771	287	2359%	563

single core, as well as for parallel-E parallel-M version 1 and the improved version 2 of the PE-PM algorithm, running on all available cores.

These results show that a gain in the order of 800% for a 12 core workstation and in the order of 2000% for a 32/64 core 4-CPU server is well within reach. The examples provided here show also that for most cases, the version 2 of the parallel algorithm that uses tiling reduction performs for most cases at a much higher level than version 1. Unlike algorithms that either utilize reduction of dimensionality (Gibbons and Hedeker 1992; Rijmen et al. 2014; Cai 2010, 2013), the algorithm presented here is a general-purpose solution for speeding up calculations and can be applied to any latent variable model available through this family of models (von Davier and Rost 2016) to speed up estimation substantially.

6 Discussion

Massive gains in processing speed can be realized by using the parallel-E parallel-M algorithm with tile reduction (PEPM-TR) for estimating generalized latent variable models. The present paper shows that gains in the order of 2000% in processing

speed are not uncommon. That is, according to Amdahl's (1967) law, the percent parallel processing with 32 cores is at a level of

$$P = \left(1 - \frac{1}{G}\right)\left(\frac{C}{C-1}\right)$$

see von Davier (2016). For $G = 20$ and $C = 64$, we obtain a value of $P = 0.965$ or a level of parallism of 96.5% for this algorithm. For gains around 800% obtained with the 12 core hardware, we obtain a very similar estimate of a level of 95.5% parallelism. This is a gain that allows using all available data in almost any psychometric analysis. Recent analyses of the combined database of the first five PISA data collections were conducted with almost two million students in more than 300 populations (approximately 60 countries or country/language groups participating on average across 5 cycles) and up to 300 items. The analysis with an IRT model of this very large dataset takes about 2–3 min on the workstation and about 90 s on the server hardware tested here. Multidimensional models for this type of massive databases are easily within reach and can be estimated in less than an hour. This enables a much more rigorous quality control and allows analysts to rerun and to obtain results based on more stringent convergence criteria, resulting in more accurate estimates.

References

G.M. Amdahl, Validity of the single processor approach to achieving large-scale computing capabilities. AFIPS Conf. Proc. **30**, 483–485. doi:10.1145/1465482.1465560

L. Cai, Metropolis–Hastings Robbins–Monro algorithm for confirmatory item factor analysis. J. Educ. Behav. Stat. **35**, 307–335 (2010)

L. Cai, *flexMIRT: A Numerical Engine for Flexible Multilevel Multidimensional Item Analysis and Test Scoring (Version 2.0) [Computer software]* (Vector Psychometric Group, Chapel Hill, NC, 2013)

H. Cui, A. Tumanov, J. Wei, L. Xu, W. Dai, J. Haber-Kucharsky, E. Xing, in *Proceedings of the ACM Symposium on Cloud Computing*. Exploiting iterativeness for parallel ML computations (ACM, New York, NY, 2014), pp. 1–14

H. Cui, X. Wei, M. Dai, Parallel implementation of expectation-maximization for fast convergence (2010). Retrieved from http://users.ece.cmu.edu/~hengganc/archive/report/final.pdf

A.S. Das, M. Datar, A. Garg, S. Rajaram, in *Proceedings of the 16th International Conference on World Wide Web, WWW 07*. Google news personalization: scalable online collaborative filtering (ACM, New York, NY, 2007), pp. 271–280. doi:10.1145/1242572.1242610

A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood estimation from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B **39**, 1–38 (1977)

G. Gan, X. Wang, J. Manzano, G.R. Gao, in *Evolving OpenMP in an Age of Extreme Parallelism: 5th International Workshop on OpenMP, IWOMP 2009 Dresden, Germany, June 3–5, 2009*, ed. by M. S. Muller, B. R. de Supinski, B. M. Chapman. Tile reduction: the first step towards tile aware parallelization in OpenMP (Springer, Berlin, Heidelberg, 2009). doi:10.1007/978-3-642-02303-3_12

R.D. Gibbons, D. Hedeker, Full-information item bi-factor analysis. Psychometrika, **57**, 423–436 (1992)

R. Henson, J. Templin, J. Willse, Defining a family of cognitive diagnosis models using log linear models with latent variables. Psychometrika, **74**, 191–210 (2009)

P.E. Lopez de Teruel, J.M. Garcia, M. Acacio, O. Canovas, P-EDR: an algorithm for parallel implementation of Parzen density estimation from uncertain observations, in *Proceedings of 13th International Parallel Processing symposium and 10th Symposium on Parallel and Distributed Processing (IPPS/SPDP)* (1999). http://ditec.um.es/~jmgarcia/papers/P-EDR.pdf

G. McLachlan, T. Krishnan, The EM Algorithm and Its Extensions (Wiley, New York, 1997)

F. Rijmen, M. Jeon, S. Rabe-Hesketh, M. von Davier, A third order item response theory model for modeling the effects of domains and subdomains in large-scale educational assessment surveys. J. Educ. Behav. Stat. **38**, 32–60 (2014)

M. von Davier, *A General Diagnostic Model Applied to Language Testing Data (Research Report No. RR-05-16)* (Educational Testing Service, Princeton, NJ, 2005). doi:10.1002/j.2333-8504.2005.tb01993.x

M. von Davier, A general diagnostic model applied to language testing data. Br. J. Math. Stat. Psychol. **61**, 287–307 (2008)

M. von Davier, *The Log-Linear Cognitive Diagnostic Model (LCDM) as a Special Case of the General Diagnostic Model (GDM) (Research Report No. RR-14-40)* (Educational Testing Service, Princeton, NJ, 2014). doi:10.1002/ets2.12043

M. von Davier, *High-Performance Psychometrics: The Parallel-E Parallel-M Algorithm for Generalized Latent Variable Models.* ETS Research Report ETS-RR-16-34 (2016)

M. von Davier, J. Rost, Polytomous mixed rasch models, in *Rasch Models: Foundations, Recent Developments and Applications,* ed. by G.H. Fischer, I.W. Molenaar (Springer, New York, 1995), pp. 371–379

M. von Davier, J. Rost, in *Handbook of Item Response Theory*, 2nd edn., ed. by W. van der Linden. Logistic mixture-distribution response models, vol 1 (CRC Press, Boca Raton, FL, 2016), pp. 393–406

M. von Davier, K. Yamamoto, Partially observed mixtures of IRT models: an extension of the generalized partial credit model. Appl. Psychol. Meas. **28**(6), 389–406 (2004)

M. von Davier, X. Xu, C.H. Carstensen, Measuring growth in a longitudinal large scale assessment with a general latent variable model. Psychometrika **76**, 318 (2011). doi:10.1007/s11336-011-9202-z

Properties of Second-Order Exponential Models as Multidimensional Response Models

Carolyn J. Anderson and Hsiu-Ting Yu

Abstract Second-order exponential (SOE) models have been proposed as item response models (e.g., Anderson et al., J. Educ. Behav. Stat. 35:422–452, 2010; Anderson, J. Classif. 30:276–303, 2013. doi: 10.1007/s00357-00357-013-9131-x; Hessen, Psychometrika 77:693–709, 2012. doi:10.1007/s11336-012-9277-1 Holland, Psychometrika 55:5–18, 1990); however, the philosophical and theoretical underpinnings of the SOE models differ from those of standard item response theory models. Although presented as reexpressions of item response theory models (Holland, Psychometrika 55:5–18, 1990), which are reflective models, the SOE models are formative measurement models. We extend Anderson and Yu (Psychometrika 72:5–23, 2007) who studied unidimensional models for dichotomous items to multidimensional models for dichotomous and polytomous items. The properties of the models for multiple latent variables are studied theoretically and empirically. Even though there are mathematical differences between the second-order exponential models and multidimensional item response theory (MIRT) models, the SOE models behave very much like standard MIRT models and in some cases better than MIRT models.

Keywords Dutch Identity • Log-multiplicative association models • Formative models • Reflective models • Composite indicators • Skew normal • Bi-variate exponential

1 Introduction

Philosophical, theoretical, and empirical differences between second-order exponential (SOE) models and multidimensional item response theory (MIRT) models exist; however, these differences that have not been fully discussed nor

C.J. Anderson (✉)
University of Illinois at Urbana-Champaign, Champaign, IL, USA
e-mail: cja@illinois.edu; http://faculty.education.illinois.edu/cja/homepage

H.-T. Yu
National Chengchi University, Taipei City, Taiwan
e-mail: hsiutingyu@gmail.com

L.A. van der Ark et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 196, DOI 10.1007/978-3-319-56294-0_2

widely recognized in the literature on SOE models are derived based on the Dutch Identity (Holland 1990; Hessen 2012). Equivalent to SOE models, log-multiplicative association (LMA) models were derived as latent variable models from statistical graphical models (Anderson and Vermunt 2000), as well as from item response models using rest scores in lieu of the latent variables (Anderson and Yu 2007; Anderson et al. 2010). Anderson and Yu (2007) studied unidimensional LMA models for dichotomous data. The LMA models are formative measurement models, and they are item response models in their own right. A better understanding of the properties of LMA models as item response models leads to implications regarding the use and performance of LMA models for analyzing response data. The LMA models have a number of advantages, including maximum likelihood estimation does not require an assumption for the marginal distribution of the latent variables and the models can be fit directly to response patterns using Newton-Raphson. The goal of this paper is to extend Anderson and Yu (2007) to study the properties of multidimensional LMA (or equivalently SOE) models for dichotomous and polytomous items.

Holland (1990) proposed and used the Dutch Identity to derive SOE models for data based on underlying uni- and multidimensional IRT models for dichotomous items. The SOE models are equivalent to LMA models, which are special cases of a log-linear model with two-way interactions. Hessen (2012) extended the Dutch Identity to polytomous items and derived an LMA model; however, he focused on models analogous to the partial credit model (i.e., models in the Rasch family), even though his extension of the Dutch Identity is more general. For models in the Rasch family, category scores are set to fixed values (e.g., consecutive integers). Hessen (2012) mentioned that the category scores could be treated as parameters and estimated. We treat category scores as parameters that are estimated. We extend and generalize the results in Anderson and Yu (2007) and Hessen (2012) to the case of multidimensional models for dichotomous and polytomous items. We highlight the philosophical, theoretical, and empirical differences between LMA and MIRT models.

In the first section of this paper, we discuss the philosophical and theoretical differences between standard MIRT and LMA models. In the following two sections, two properties of LMA are theoretically and empirically studied: the downward collapsibility of LMA models and the effect of different marginal distributions of the latent variables on the models' performance. We conclude with a discussion the potential uses of LMA models in measurement contexts.

2 Reflective and Formative Models

The differences between reflective and formative latent variable models have been discussed by Markus and Borsboom (2013), Bollen and Bauldry (2011), and others. Our intent here is to show the philosophical differences between LMA and MIRT models and how they lead to different mathematical models.

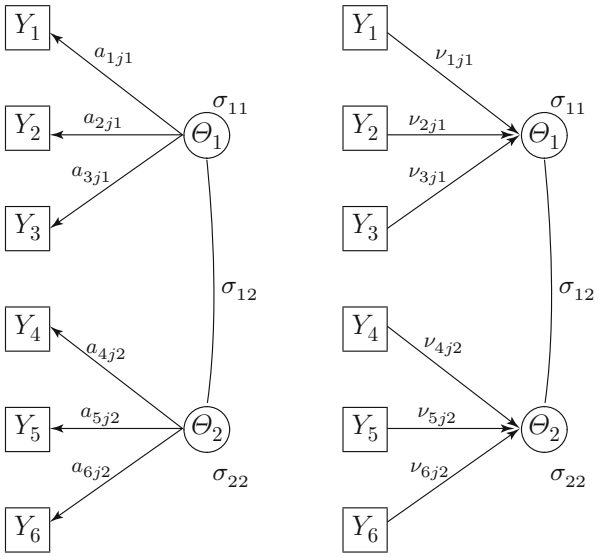

Fig. 1 Graphs corresponding to reflective (*left*) and formative (*right*) models for six items and two latent continuous variables

A reflective model posits that latent variables are prior to behavior, and the latent variables are conceived of as existing whether they are measured or not. A reflective model is illustrated by the graph on the left in Fig. 1. The values on the latent variables lead to observed responses; therefore, behavior indicates or reflects a person's value on the unobserved quantity. A change in the value of a latent variable causes a change in the response behavior. The items are *effect indicators* (Bollen and Bauldry 2011).

To algebraically take into account the directional nature of the relationship between $\boldsymbol{\theta}$ and $\boldsymbol{y}$, models are developed by writing the joint distribution of $\boldsymbol{\theta}$ and $\boldsymbol{y}$ as $f(\boldsymbol{y}, \boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})$. For a MIRT model, the marginal distribution of the latent variables $f(\boldsymbol{\theta})$ is typically assumed to be multivariate normal, and the distribution for the responses conditional on the latent variables $f(\boldsymbol{y}|\boldsymbol{\theta})$ is a product of multinomial logistic regression models. The model for responses to items is found by numerically integrating over the latent variables; that is, the probability of response pattern $\boldsymbol{y}$ is

$$P(\boldsymbol{y}) = \int_{\theta_1} \cdots \int_{\theta_M} \prod_{i=1}^{I} \frac{\exp[\beta_{ij} + \sum_m \alpha_{ijm}\theta_m]}{\sum_h \exp[\beta_{ih} + \sum_m \alpha_{ihm}\theta_m]} f(\boldsymbol{\theta}) d(\boldsymbol{\theta}), \tag{1}$$

where β_{ij} is a location parameter for response option j of item i, and α_{ijm} is the slope parameter for response option j of item i on latent variable θ_m.

In a formative model, the direction of the relationship between $\boldsymbol{\theta}$ and $\boldsymbol{y}$ is reversed relative to the reflective model. A graph representing a formative model

is illustrated on the right in Fig. 1. Items define and give meaning to latent variables. The items are *composite indicators* because $\boldsymbol{\theta}$ are composites of the values of the items (Bollen and Bauldry 2011). The joint distribution of $\boldsymbol{y}$ and $\boldsymbol{\theta}$ is found by first specifying the distribution for $f(\boldsymbol{y})$ and then the distribution for $f(\boldsymbol{\theta}|\boldsymbol{y})$; that is, $f(\boldsymbol{y}, \boldsymbol{\theta}) = f(\boldsymbol{\theta}|\boldsymbol{y})f(\boldsymbol{y})$. Assuming that $f(\boldsymbol{y})$ is multinomial and $f(\boldsymbol{\theta}|\boldsymbol{y})$ is a homogeneous conditional, Gaussian distribution leads to an LMA model for the probabilities of observed response patterns $\boldsymbol{y}$ (Anderson and Vermunt 2000; Anderson et al. 2010). The model for data is

$$P(\boldsymbol{y}) = \exp\left[\lambda + \sum_{i=1}^{I}\lambda_{ij} + \sum_{i}\sum_{k\geq i}\sum_{m}\sum_{m'\neq m}\sigma_{mm'}\nu_{ijm}\nu_{kjm'}\right], \tag{2}$$

where λ ensures probabilities sum to 1 over response patterns, λ_{ij} is the marginal effect term for response option j to item i, ν_{ijm} is the category scale value for response j to item i on latent variable m, and $\sigma_{mm'}$ is a within response pattern variance or covariance of the latent variable(s). The λ_{ij}s and ν_{ijm}s in (2) are analogous to the β_{ij}s and α_{ijm}s, respectively, in (1). Based on the LMA model, the conditional means of the latent variables given $\boldsymbol{y}$ equal:

$$E(\theta_m|\boldsymbol{y}) = \sum_{m'=1}^{M}\sigma_{mm'}\left(\sum_{i=1}^{I}\nu_{ijm'}\right). \tag{3}$$

Models (1) and (2) are very general models. In this paper, we study the case where each item is directly related to one and only one latent variable, that is, $\nu_{ijm} \neq 0$ and $\alpha_{ijm} \neq 0$ for one and only one m. We expect that the results we find will be the same for more complex models, but we leave this for future study.

The MIRT model given in (1) is not only philosophically different but mathematically different from the LMA model given in (2).

3 Downward Collapsibility of LMA Models

If an item is dropped from data generated from a MIRT model, the data excluding the item still follow a MIRT model and theoretically yield the same estimates of item parameters for the remaining items. If an item is dropped from (or added to) an LMA model, the resulting model is a different model with different parameter estimates. We theoretically and empirically study the effect on LMA model parameter estimates when dropping an item from data (i.e., collapse data over an item). In the first section, we consider the case when data are generated from an LMA model (not collapsible), and in the second section, we consider the case when data are generated from a MIRT model (downward collapsible).

3.1 LMA-Generated Data

Suppose that item 1 is directly related to θ_1 and it is dropped from the data. Let $\boldsymbol{y}_{-1}$ indicate the data excluding item 1. Rather than (3), the conditional means for θ_1 and θ_m are

$$E(\theta_1|\boldsymbol{y}_{-1}) = \sigma_{11}\sum_{i\neq 1}\nu_{ij1} + \sum_{m>1}\left(\sigma_{1m}\sum_{k}\nu_{kjm}\right) + \sigma_{11}\sum_{j}\nu_{1j1}P(Y_1 = j|\boldsymbol{y}_{-1})$$

and

$$E(\theta_m|\boldsymbol{y}_{-1}) = \sigma_{1m}\sum_{i\neq 1}\nu_{ij1} + \sum_{m'>1}\left(\sigma_{mm'}\sum_{k}\nu_{kjm'}\right) + \sigma_{1m}\sum_{j}\nu_{1j1}P(Y_1 = j|\boldsymbol{y}_{-1}),$$

respectively. The last term in each of these equations for the conditional means is unobserved and equals the expected biases of the means due to dropping item 1.

Dropping an item that is directly related to θ_1 changes the conditional variances of θ_1 and any θ_m directly related to θ_1 (i.e., $\sigma_{1m} \neq 0$). In particular, the conditional variances after collapsing over item 1 are

$$\text{var}(\theta_1|\boldsymbol{y}_{-1}) = \sigma_{11} + \sigma_{11}^2\left(\sum_{j}\nu_{ij1}^2P(Y_1 = j|\boldsymbol{y}_{-1}) - \left(\sum_{j}\nu_{ij1}P(Y_1 = j|\boldsymbol{y}_{-1})\right)^2\right),$$

and

$$\text{var}(\theta_m|\boldsymbol{y}_{-1}) = \sigma_{mm} + \sigma_{1m}^2\left(\sum_{j}\nu_{ij1}^2P(Y_1 = j|\boldsymbol{y}_{-1}) - \left(\sum_{j}\nu_{ij1}P(Y_1 = j|\boldsymbol{y}_{-1})\right)^2\right).$$

The conditional variances will increase for larger values of σ_{11} and σ_{1m}. The change of $\text{var}(\theta_m|\boldsymbol{y}_{-1})$ is smaller than that for $\text{var}(\theta_1|\boldsymbol{y}_{-1})$ because $\sigma_{1m}^2 \leq \sigma_{11}^2$. Regardless of the value of σ_{11} and σ_{1m}, the conditional means and variances are affected the most when an item with the largest values of ν_{ij1} is dropped, and they are least affected when the item with the smallest values of ν_{ij1} is dropped.

Our interest is in the theoretical behavior of the LMA models; therefore, $P(\boldsymbol{y})$s were computed from an LMA (six items, three response options per item), so the LMA model fits the data perfectly. The size of the scale values for an item was measured by $\sum_j \nu_{ijm}^2$. Two additional data sets were created by collapsing over the item with the smallest value and the largest value of $\sum_j \nu_{ijm}^2$. The item with the weakest relationship to a θ_m should have the smallest effect on the results, and collapsing over the item with the strongest relationship to a θ_m should have the largest effect.

Throughout this paper, maximum likelihood estimation was used to estimate parameters of LMA and MIRT models. The LMA models were fit to data using SAS$^{\delta}$ PROC NLP (version 9.4, SAS Institute Inc. 2015). The MIRT models were fit to data using *flexMIRT* (Houts and Cai 2013) assuming bivariate (multivariate) normality.[1]

In terms of goodness of fit, the likelihood ratio goodness-of-fit statistic (G^2) is used as an index but is not compared to a χ^2 distribution because there is no sampling variability. As a second index, we used the dissimilarity index:

$$D = \sum_{\boldsymbol{y}} \frac{|P(\boldsymbol{y}) - \hat{P}(\boldsymbol{y})|}{2},$$

where the sum is over all response patterns, $P(\boldsymbol{y})$ is the probability of response pattern $\boldsymbol{y}$, and $\hat{P}(\boldsymbol{y})$ is the fitted value of the probability of response pattern $\boldsymbol{y}$ from a model. The index D is interpretable as the proportion of data that would have to be moved from one response pattern to another for the model to fit perfectly (Agresti 2013).

Any misfit of the LMA model fit to the six items is due to numerical inaccuracy in the data generation and/or model estimation. The LMA model fits the probabilities of response patterns for the six items nearly perfectly. When collapsing over the weak item, the parameter estimates and goodness-of-fit statistics of the LMA model were nearly identical to those when the model was fit to all six items. Specifically, collapsing over the weak item had a smaller impact on the goodness of fit than collapsing over the strong item (i.e., $G^2 = 0.0000$ versus $G^2 = 0.0002$ and $D = 0.0002$ versus $D = 0.0049$). All of the LMA models fit the probabilities better than all of the MIRT models.

When collapsing over the strong item, there were noticeable differences between the estimated parameters from the LMA model fit to those used to generate the data. The variance of θ_m increased the most when the item dropped is the strong item. Specifically, when the strong item is dropped, the variance of the latent variable to which it is connected goes from 0.87 to 1.66, but when the weak item is dropped, the variance of the latent variable to which it is connected goes from 0.77 to 0.88. As predicted, both $\hat{\sigma}_{11}$ and $\hat{\sigma}_{22}$ increased when collapsing over either the weak or strong item. The change in both variances occurs because when we collapse over an item related to, say θ_1, leads to less information to estimate the latent variable θ_2, which increases uncertainty (i.e., larger σ_{22}).

[1] Files containing code and data that reproduce all analyses can be downloaded from http://faculty.education.illinois.edu/cja/homepage.

3.2 MIRT-Generated Data

If θ_1 and θ_2 were discrete, then we could collapse over, for example, item 1 and expect $\hat{\nu}_{ijm}$ for $i \neq 1$ to remain the same. Since for the LMA models $\hat{\boldsymbol{\theta}}$ equals a weighted sum of category scores, $\hat{\boldsymbol{\theta}}$ is empirically discrete and LMA models might be collapsible. When data are generated from a model that implies collapsibility, whether LMA scale values are affected by dropping items is an open question. Since MIRT models imply collapsibility, probabilities were generated from a two-dimensional MIRT model with $\boldsymbol{\theta} \sim MVN(\boldsymbol{\mu} = (0, 0), \rho = 0.5)$ for eight items where items 1–4 were related to θ_1, and items 5–8 were related to θ_2. The generated probabilities were collapsed over one item at a time until there were only four items remaining. We alternated collapsing over an item related to θ_1 and one related to θ_2.

Since LMA models are formative measurement models, we are primarily interested in the $\hat{\nu}_{ijm}$s, which are used to compute estimates of the conditional means of the latent variables (i.e., $\hat{\boldsymbol{E}}(\theta_m|y)$). The scale values $\hat{\nu}_{ijm}$ were essentially unaffected by collapsing the data. When data were collapsed over an item, both of the $\hat{\sigma}_{mm}$s increased. When the first item was dropped, which was related to θ_2, the increase of $\hat{\sigma}_{22}$ was greater than that for $\hat{\sigma}_{11}$. When the second item was dropped, which was related to θ_1, the increase in $\hat{\sigma}_{11}$ was greater than that for $\hat{\sigma}_{22}$. This pattern continued until there are only four items remaining.

In sum, if data are generated from a MIRT model, which collapsibility, then the LMA model yields nearly the same $\hat{\nu}_{ijm}$s when items are dropped. Conversely, we can consider adding items. If the data come from a model that implies collapsibility and then when adding items (assuming that the added items are related to the underlying latent variable(s)), the $\hat{\nu}_{ijm}$s are not expected to change, and $\hat{\sigma}_{mm}$s are expected to be smaller.

4 Different Marginal Distributions

A property often given as an advantage of LMA models is that a marginal distribution of the latent variables is a mixture of normals, which can take on many different shapes. The goal of this section is to determine whether and when an LMA model may perform well in terms of goodness of fit and parameter recovery and compare LMA model performance with a corresponding MIRT model.

In this study, we generated probabilities for response patterns by numerically integrating out the latent variables from a MIRT model assuming one of four different underlying distributions. The multivariate normal (MVN) was chosen because this is the typical assumption made when fitting a MIRT model. The multivariate skew normal was chosen because the MVN is a special case of the skew normal, and there has been some interest in using the skew normal as an alternative to the normal distribution (Azevedo et al. 2011; Casabianca and Junker 2016; Lee 2002). Marshall-Olkin bivariate exponential distribution (Mardia et al. 1979) was

chosen because some variables in the data that we often analyze are very skewed. Lastly, a mixture of two multivariate normal distributions was chosen to mimic a situation where individuals have opposite attitudes or views. This also reflects a situation where there is an important group variable that has not been included in the model, and there is differential item functioning.

As the number of items increases to ∞, $\sigma_{mm} \longrightarrow 0$, an LMA model will yield the actual marginal distribution of the latent variables. The behavior of LMA models for short tests or subscales is less certain; therefore, we empirically examine the behavior of the models when fit to generated data for short tests. Probabilities of response patterns were generated using the MIRT model in (1) for $M = 2$ latent variables and $I = 4$ or 6 items with $J = 2$, 3, or 4 response options. For the multivariate normal distribution, we also fit models to data with $M = 3$ and $I = 6$ items.

Both the MIRT and the LMA models were fit to all of the data sets. Albeit naive, when the distribution generating the data is not normal, the MIRT models were fit to data assuming multivariate normality. Although not reported here, two additional models were fit as baseline models: the log-linear model of independence and the homogeneous (all two-way interaction) log-linear model. The probabilities of response patterns were multiplied by 1,000,000 to retain more decimal places and accuracy. Besides the dissimilarity index D, a second measure of goodness of fit is reported for the models: the percent of association accounted for by a model,

$$\text{Percent association} = \frac{G^2_{independence} - G^2_{model}}{G^2_{independence}} \times 100\%,$$

where the likelihood ratio statistic G^2 from the independence model is a measure of the amount of association in the data.

To examine parameter recovery, we used the correlation between the parameters used to generate the data and the estimated parameters from the LMA and MIRT models. Given our focus on LMA models, we are primarily interested in the estimation of the ν_{ijm} parameters. The marginal effect terms λ_{ij} generally are viewed as nuisance parameters from an LMA model framework, but the correlations for marginal terms are reported for the sake of completeness.

The results for different numbers of items and response options are all very similar; therefore, we only report the result for one case (i.e., six items, three response options, and two latent variables). Goodness-of-fit statistics are reported in Table 1, and correlations between estimated parameters and those used to generate the data are reported in Table 2.

When data were generated using the bivariate normal distribution ($\boldsymbol{\mu} = \mathbf{0}$, $\rho = 0.5$), the MIRT model should fit perfectly. Any misfit is due to numerical inaccuracy in generating the probabilities and/or estimating the model. The MIRT models essentially fit perfectly; however, the goodness-of-fit indices for the LMA models are just shy of perfect. When data were generated using a skew normal (i.e., $\boldsymbol{\mu} = \mathbf{0}$, $\rho = 0.75$, and shape parameters 2 and 3) or a bivariate exponential distribution (i.e.,

Table 1 Goodness-of-fit statistics for LMA and MIRT models fit to date generated from a MIRT model with different underlying distributions for $f(\boldsymbol{\theta})$

	Dissimilarity		Percent association	
Underlying distribution	LMA	MIRT	LMA	MIRT
Bivariate normal	0.0016	0.0002	99.99	100.00
Skew normal	0.0268	0.0268	97.14	97.16
Bivariate exponential	0.0127	0.0129	98.44	98.39
Mixture of normals	0.0346	0.0708	99.43	96.05

Table 2 Correlations between LMA and MIRT model parameter estimates and parameters used generated MIRT model probabilities for different $f(\boldsymbol{\theta})$s

	LMA	MIRT	LMA	MIRT
Underlying distribution	$r(\alpha_{ijm}, \hat{\nu}_{ijm})$	$r(\alpha_{ijm}, \hat{a}_{ijm})$	$r(\beta_{ij}, \hat{\lambda}_{ij})$	$r(\beta_{ijm}, \hat{b}_{ij})$
Bivariate normal	0.9980	1.0000	0.9839	1.0000
Skew normal	0.9950	0.9962	0.8361	0.7506
Bivariate exponential	0.9326	0.9077	0.8257	0.7894
Mixture of normals	0.9971	0.9430	0.9665	0.9428

$f(\boldsymbol{\theta}) = \exp(-1.0\theta_1 - 0.5\theta_2 - 0.2\max(\theta_1, \theta_2))\kappa$ where κ normalized the function), the LMA and MIRT models both provide good representations of the data, and there are no systematic differences in terms of which model fits the data better. When data were generated from the mixture of two normals (i.e., $\boldsymbol{\mu}_1 = (-2, -2)'$, $\boldsymbol{\mu}_2 = (2, 2)'$, $\rho = 0.5$, and mixing weight of 0.5), the LMA models clearly fit the data better than the MIRT models.

More differences between the models' performance were found in terms of parameter recovery. When data were from the bivariate normal, MIRT parameters are perfectly correlated with those used to generate the data; however, the LMA parameters were just short of perfect. For the skew normal, the correlations between the α_{ijm}s used to generate the data and the estimated ν_{ijm}s parameters from the LMA models were about the same as the corresponding correlations of parameters from the MIRT models; however, the correlations for the β_{ijm}s were much larger for the LMA model than the MIRT model. For the exponential and mixture of normal distributions, the correlations for the estimated ν_{ijm}s and λ_{ij}s from the LMA models were considerably larger than those for parameters from the MIRT models.

5 Discussion

The LMA models and standard MIRT models were shown to be philosophically and mathematically different models; however, they share some important properties. For short tests, the LMA models performed nearly as well as standard MIRT models when the underlying distribution of the latent variables is multivariate normal, and

the LMA and MIRT models empirically perform equally well when the underlying distribution is skew normal. With the skew normal, the goodness of fit is about the same for both the LMA and MIRT models; however, the estimation of the β_{ij}s parameters had lower correlations with the parameters used to generate the data than the LMA model parameters. The LMA models perform better than MIRT models in terms of goodness of model fit to data and parameter recovery when data arise from an LMA model and when $f(\boldsymbol{\theta})$ follows either a bivariate exponential distribution or a mixture of two normal distributions.

The LMA models are more flexible than discussed in this paper. The LMA models can include covariates for the latent variables, the marginal effect terms (i.e., the λ_{ij}), and the conditional variances and covariances of the latent variables (Anderson 2013). The models also permit various restrictions on parameters, including equality, ordinal, partially ordinal, linear transformations, and/or any desired transformation (Anderson 2013). The LMA models can also represent more complex latent variable structures than those studied in this paper, such as those where items "load" on multiple correlated or uncorrelated latent variables (e.g., bi-factor models). Since the assumptions and theory are the same, we expect the same results for more complex models such as those that we found for the simpler models reported in this paper.

Our focus was on short tests because these are cases where LMA and MIRT models may differ. Although we used common commercial software (SAS) to fit the LMA models to data, one bottleneck to more widespread applications of LMA models is a limitation to the size of the problem that can be handled. The size of the cross-classification of items (i.e., number of response patterns) increases exponentially when adding items and/or categories per item. When scores are input, the pseudo-likelihood method given in Anderson et al. (2007) works well and can be implemented in any program that fits conditional multinomial logistic models. Recently Paek (2016), Paek and Anderson (2017) proposed a solution to the more general problem where scores are estimated. In simulations, Paek (2016) showed that the algorithm yields nearly identical parameter estimates as MLE of LMA models for short tests and that the algorithm recovers parameters used to simulate the data in longer tests (i.e., 20 and 50 items). The more general algorithm also can be implemented in any software program that fits conditional multinomial logistic regression models.

We do not advocate that LMA models replace MIRT models because they are philosophically and theoretically different measurement models. The LMA models actually may be complimentary to applications of MIRT models. Suppose a researcher desires a reflective model but does not know what marginal distribution of the latent variable(s) should be used when fitting a MIRT model to data. The LMA models can be used to estimate the marginal distribution of the latent variables, which could confirm or suggest a distribution to be used when fitting the MIRT model to data.

The empirical studies in this paper imply that one cannot conclusively determine whether the model should be formative or reflective. Whether one performs better

than the other is an empirical question. The choice between using an LMA model or a MIRT model for a particular case depends on a researcher's conceptualization of the latent variable.

References

A. Agresti, *Categorical Data Analysis,* 3rd edn. (Wiley, New York, 2013)
C.J. Anderson, Multidimensional item response theory models with collateral information as Poisson regression models. J. Classif. **30**, 276–303 (2013). doi: 10.1007/s00357-00357-013-9131-x
C.J. Anderson, J.K. Vermunt, Log-multiplicative association models as latent variable models for nominal and/or ordinal data. Sociol. Methodol. **30**, 81–121 (2000)
C.J. Anderson, H.-T. Yu, Log-multiplicative association models as item response models. Psychometrika **72**, 5–23 (2007)
C.J. Anderson, Z. Li, J.K. Vermunt, Estimation of models in a Rasch family for polytomous items and multiple latent variables. J. Stat. Softw. **20** (2007). http://www.jstatsoft.org/v20/i06/v20i06.pdf
C.J. Anderson, J.V. Verkuilen, B.L. Peyton, Modeling polytomous item responses using simultaneously estimated multinomial logistic regression models. J. Educ. Behav. Stat. **35**, 422–452 (2010)
C.L.N. Azevedo, H. Bolfarine, D.F. Andrade, Bayesian inference for a skew-normal IRT model under the centred parameterization. Comput. Stat. Data Anal. **55**, 353–365 (2011)
K.A. Bollen, S. Bauldry, Three C's in measurement models: causal indicators, composite indicators, and covariates. Psychol. Methods **16**, 265–284 (2011)
J.M. Casabianca, B.W. Junker, Multivariate normal distribution, in *Handbook of Item Response Theory, Volume Two: Statistical Tools*, ed. by W.J. van der Linden (Talyor & Fransics/CRC Press, Boca Raton, 2016), pp. 35–46
D.J. Hessen, Fitting and testing conditional multinomial partial credit models. Psychometrika **77**, 693–709 (2012). doi:10.1007/s11336-012-9277-1
P.H. Holland, The Dutch identity: a new tool for the study of item response models. Psychometrika **55**, 5–18 (1990)
C.R. Houts, L. Cai, *flexMIRT: Flexible Multilevel Multidimensional Item Analysis and Test Scoring* (Vector Psychometric Group, LLC, Chapel Hill, 2013)
J. Lee, Multidimensional item response theory: an investigation of interaction effects between factors on item parameter recovery using Markov Chain Monte Carlo. Unpublished Doctoral Dissertation, Michigan State University (2002)
K.V. Mardia, J.M. Kent, J.M. Bibby, *Multivariate Analysis* (Academic, Orlando, 1979)
K.A. Markus, D. Borsboom, *Frontiers of Test Validity Theory: Measurement, Causation and Meaning* (Routlege, New York, 2013)
Y. Paek, Pseudo-likelihood estimation of multidimensional polytomous item response theory models. Unpublished Doctoral Dissertation, University of Illinois at Urbana-Champaign (2016)
Y. Paek, C.J. Anderson, Pseudo-likelihood estimation of multidimensional response models: polytomous and dichotomous items, in *The 81st Annual Meeting of the Psychometric Society, Asheville, NC*, ed. by L.A. van der Ark, S.A. Culpepper, J.A. Douglas, W.C. Wang, M. Wiberg (Springer, New York, 2017)
SAS Institute Inc., *Statistical Analysis System*, version 9.4 (SAS Institute, Cary, 2015)

Pseudo-likelihood Estimation of Multidimensional Response Models: Polytomous and Dichotomous Items

Youngshil Paek and Carolyn J. Anderson

Abstract Log-multiplicative association (LMA) models have been proposed as uni- and multidimensional item response models for dichotomous and/or polytomous items. A problem that prevents more widespread use of LMA models is that current estimation methods for moderate to large problems are computationally prohibitive. As a special case of a log-linear model, maximum likelihood estimation (MLE) of LMA models requires iteratively computing fitted values for all possible response patterns, the number of which increases exponentially as the number of items and/or response options per item increases. Anderson et al. (J. Stat. Softw. 20, 2007, doi:10.18637/jss.v020.i06) used pseudo-likelihood estimation for linear-by-linear models, which are special cases of LMA models, but in their proposal, the category scores are fixed to specific values. The solution presented here extends pseudo-likelihood estimation to more general LMA models where category scores are estimated. Our simulation studies show that parameter estimates from the new algorithm are nearly identical to parameter estimates from MLE, work for large numbers of items, are insensitive to starting values, and converge in a small number of iterations.

Keywords Log-multiplicative association models • Log linear-by-linear models • Second-order exponential models • Multidimensional item response theory • Formative measurement models

1 Introduction

Log-multiplicative association (LMA) models have been proposed as uni- and multidimensional item response models for dichotomous and/or polytomous items (Anderson et al. 2010; Holland 1990; Hessen 2012, and others). They are formative measurement models (Anderson and Yu 2017) that do not require an assumption for the marginal distribution of the latent variables. Although maximum likelihood

Y. Paek (✉) • C.J. Anderson
University of Illinois at Urbana-Champaign, Champaign, IL, USA
e-mail: ypaek2@illinois.edu; http://faculty.education.illinois.edu/cja/homepage

L.A. van der Ark et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 196, DOI 10.1007/978-3-319-56294-0_3

estimation can be accomplished for small numbers of items, the estimation of LMA models for moderate to large problems is computationally prohibitive because fitted values for all possible response patterns must be iteratively computed. The number of response patterns increases exponentially as the number of items and/or response options per item increases. Pseudo-likelihood estimation (PLE) was proposed by Anderson et al. (2010) for log linear-by-linear models which are special cases of LMA models where the category scores (e.g., slopes for the latent variables) are set to fixed values at input. We extend the pseudo-likelihood approach to general LMA models where category scores are treated as parameters and are estimated. This method works for large numbers of items and response options.

One of the most widely used programs for estimating LMA models is ℓ_{EM} (Vermunt 1997), which used quasi- or unidimensional Newton-Raphson. With ℓ_{EM} we were able to fit an LMA model to 12 binary items (i.e., $2^{12} = 4096$ response patterns). LMA models can also be fit using analytic derivatives and a Newton-Raphson algorithm as implemented in SAS® procedure NLP (SAS Institute Inc. 2015). Using SAS, the largest problem that we successfully fit had seven items each with five response categories (i.e., $5^7 = 78,125$ response patterns). Adding a single item increased the number of response patterns to $390,625$, and estimation became problematic. Ten items with five response categories per item (i.e., $9,765,625$ response patterns) are beyond the capability of current estimation methods.

Pseudo-likelihood estimation simplifies estimation of large complex models by maximizing the product of likelihoods of a set of conditional models based on the complex model. The method, first proposed by Besag (1974), has been used to solve estimation problems in a number of different settings (Huwang and Huwang 2002; Geys et al. 1999; Liang and Yu 2003; Johnson and Riezler 2002; Strauss and Ikeda 1990; Wasserman and Pattison 1990; Molenberghs and Verbeke 2005). The original uses of PLE to estimate parameters of Rasch models were limited to unidimensional models for pairs of binary items (Arnold and Strauss 1991; Zwinderman 1995). Smit (2000) extended the use of PLE to a set of dichotomous items and studied the quality of the estimates relative to other standard estimation methods. Pseudo-likelihood estimation (Anderson et al. 2007) of LMA models was developed to handle only the special case, when category scale values are assumed and set to fixed values. The estimation method and algorithm that we propose use pseudo-likelihood estimation but add a step for estimating the category scores.

PLE parameter estimates are asymptotically normal and consistent (Geys et al. 1999; Aerts et al. 2002), which is important for forming confidence intervals and hypothesis testing. Other advantages of PLE are that it is fast and stable, and implementation is straightforward.

The structure of the paper is as follows. In the first section, LMA models are presented in a form that is key to our algorithm. In the second section, we discuss pseudo-likelihood estimation and present our algorithm. In the subsequent sections, we present the results of simulation studies showing that the new step for estimation of category scores works (i.e., one latent variable) and simulation studies showing that the method works for multidimensional models. We conclude with a discussion and possible extensions of the algorithm.