Psychometrika

CONTENTS

VOLUME SIXTY-SIX **MARCH 2001** **NUMBER 1**

MANUSCRIPTS RECENTLY ACCEPTED FOR PUBLICATION

Listed in the Order They will be Published

Standard Errors of Fit Indices Using Residuals in Structural Equation Modeling
Haruhiko Ogasawara

Nonparametric Goodness-of-Fit Tests for the Rasch Model
Ivo Ponocny

PSYCHOMETRIKA—VOL. 66, NO. 1, 5–24
MARCH 2001

AN INTERACTIVE MULTIOBJECTIVE PROGRAMMING APPROACH TO COMBINATORIAL DATA ANALYSIS

MICHAEL J. BRUSCO

INFORMATION & MANAGEMENT SCIENCES
FLORIDA STATE UNIVERSITY

STEPHANIE STAHL

TALLAHASSEE, FLORIDA

Combinatorial optimization problems in the social and behavioral sciences are frequently associated with a variety of alternative objective criteria. Multiobjective programming is an operations research methodology that enables the quantitative analyst to investigate tradeoffs among relevant objective criteria. In this paper, we describe an interactive procedure for multiobjective asymmetric unidimensional seriation problems. This procedure uses a dynamic-programming algorithm to partially generate the efficient set of sequences for small to medium-sized problems, and a multioperation heuristic to estimate the efficient set for larger problems. The interactive multiobjective procedure is applied to an empirical data set from the psychometric literature. We conclude with a discussion of other potential areas of application in combinatorial data analysis.

Key words: combinatorial optimization, multiobjective programming, seriation.

1. Introduction

The solution of combinatorial optimization problems within the context of data analysis has been prevalent in quantitative psychology and related disciplines for many years (see Arabie and Hubert (1996) for a thorough review and extensive bibliography of research in this area). The formal underpinnings for these problems are frequently based on classic mathematical-programming paradigms. For example, Hubert and Schultz (1976) outlined applications in quantitative psychology related to the quadratic assignment problem (QAP), originally posed by Koopmans and Beckmann (1957). Similarly, Hubert and Baker (1978) illustrated data-analysis strategies associated with the traveling salesman problem, which is a special case of the QAP. Despite the similarity in the underlying structure of combinatorial problems in the social and behavioral sciences, the number of alternative objective criteria can be quite large. Quantitative psychologists and behavioral researchers—hereafter referred to as quantitative analysts—are often faced with difficult choices regarding the appropriate objective criterion and solution methodology for the particular problem at hand.

If the quantitative analyst is uncertain as to the selection of the appropriate criterion, then evaluating a variety of solution alternatives regarding their objective-function values across multiple criteria would be helpful. The consideration of multiple objective criteria can be useful for a couple of reasons. First, the quantitative analyst gains confidence in a solution if it performs well across two or more criteria. In other words, multiple criteria can provide evidence of the robustness of a solution. Second, the consideration of multiple objectives during the solution process can uncover intuitively appealing alternatives that might be overlooked if only a single criterion is considered. To realize these benefits, the quantitative analyst needs a sound methodological approach for combinatorial data analysis in a multiobjective environment. This paper presents one possible approach based on interactive multiobjective programming.

Stephanie Stahl is a freelance writer and editor. She can be reached via e-mail at s-stahl@worldnet.att.net.

Requests for reprints should be sent to Michael J. Brusco, IMS Department, Florida State University, Tallahassee, FL 32306-1110. E-Mail: mbrusco@cob.fsu.edu

0033-3123/2001-1/1999-0746-A $00.75/0

PSYCHOMETRIKA—VOL. 66, NO. 1, 25–43
MARCH 2001

RANDOM EFFECTS DIAGONAL METRIC MULTIDIMENSIONAL SCALING MODELS

Douglas B. Clarkson

DATA ANALYSIS PRODUCTS DIVISION, MATHSOFT, INC.

Richard Gonzalez

UNIVERSITY OF MICHIGAN

By assuming a distribution for the subject weights in a diagonal metric (INDSCAL) multidimensional scaling model, the subject weights become random effects. Including random effects in multidimensional scaling models offers several advantages over traditional diagonal metric models such as those fitted by the INDSCAL, ALSCAL, and other multidimensional scaling programs. Unlike traditional models, the number of parameters does not increase with the number of subjects, and, because the distribution of the subject weights is modeled, the construction of linear models of the subject weights and the testing of those models is immediate. Here we define a random effects diagonal metric multidimensional scaling model, give computational algorithms, describe our experiences with these algorithms, and provide an example illustrating the use of the model and algorithms.

Key words: multidimensional scaling, random coefficients.

1. Random Effects Diagonal Metric Multidimensional Scaling Models

Multidimensional scaling analysis estimates the coordinates of a set of n objects in a τ dimensional space given only a measure (observed with error) of the distances between the objects. The dimensionality of the space, τ, is specified by the model. In the models considered here, the distance measurements, called *dissimilarities*, are obtained from a sample of m subjects. The observed data are denoted by y_{lj}, where $l = 1, \ldots, m$ indexes subjects, and $j = 1, \ldots, N_t$ indexes distances between pairs of stimuli. The y_{lj} are usually not distances in the strict sense, but are thought to satisfy the various properties required of distance measures. Supposing that there are n stimuli, a maximum of $N = n(n+1)/2$ dissimilarities on each subject are possible, but only a subset of these need to be measured for estimation, i.e., $N_l \leq N$.

Here we are concerned with the diagonal metric (also called the individual differences or INDSCAL) model first proposed by Caroll and Chang (1970). Let x_{ik} denote the coordinates for $i = 1, \ldots, n$ stimuli in a $k = 1, \ldots, \tau$ dimensional space, and call the matrix $\mathbf{X} = (x_{ik})$ the *configuration matrix*. The diagonal metric model assumes that the expected distance between stimulus i and stimulus j for individual l is given as

$$d_{lij} = \left(\sum_{k=1}^{\tau} w_{lk} (x_{ik} - x_{jk})^2 \right)^{1/2}$$

In this model the *subject weights*, w_{lk}, account for variability between individuals—each individual gives different weight to each of the τ dimensions.

We would like to thank J. Douglas Carroll for early consultation of this research, and Robert I. Jennrich for commenting on an earlier draft of this paper and for help on the computational algorithms. James O. Ramsay and Forrest W. Young were instrumental in providing the example data. This work was supported in part by National Institute of Mental Health grant 1 R43 MH57559-01. We would also like to thank the anonymous referees for comments that helped to clarify our work.

Requests for reprints should be sent to Douglas Clarkson, Data Analysis Products Division, MathSoft, Inc., 1700 Westlake Ave. N., Suite 500, Seattle, WA, 98109-3044. E-Mail: clarkson@statsci.com

0033-3123/2001-1/1998-0637-A $00.75/0

PSYCHOMETRIKA—VOL. 66, NO. 1, 45–62
MARCH 2001

MIXED-EFFECTS ANALYSES OF RANK-ORDERED DATA

ULF BÖCKENHOLT

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

This paper presents a synthesis of Bock's (1972) nominal categories model and Luce's (1959) choice model for mixed-effects analyses of rank-ordered data. It is shown that the proposed ranking model is both parsimonious and flexible in accounting for preference heterogeneity as well as fixed and random effects of covariates. Relationships to other approaches, including Takane's (1987) ideal point discriminant model and Croon's (1989) latent-class version of Luce's ranking model, are also discussed. The application focuses on a ranking study of behavioral traits that parents find desirable in children.

Key words: latent trait models, latent class analysis, probabilistic choice models, nominal categories model, Luce's choice model.

1. Introduction

The appeal of collecting ranking data as opposed to rating data stems from the fact that individual differences in rating scale usage and possibly arbitrary definitions regarding the number of response categories and category labels can be avoided. However, despite these attractive features, the ranking method is used only infrequently in applied work. To a great extent, the lack of ranking applications can be explained by computational difficulties that arise from taking into account heterogeneous responses. This paper presents a parsimonious and yet flexible solution to this problem by combining Bock's (1972) nominal categories model and Luce's (1959) ranking model within a latent-class framework. The synthesis of these approaches leads to a mixed-effects logistic regression model for ranking data that considers both observed and omitted or latent covariates.

Thurstone (1931) initiated the analysis of heterogeneity in ranking responses by proposing that the judgments underlying the rankings follow a multivariate normal distribution. This framework facilitates a straightforward formulation of both linear models for the means and structural equations for the covariance matrix of the judgments (Böckenholt, 1992; Chan & Bentler, 1998; Maydeu-Olivares, 1999; Takane & de Leeuw, 1987; Yao & Böckenholt, 1999). As pointed out by Marden (1995), this approach is most useful when there is substantial agreement among the rankers about the underlying representation of the choice options. For example, Thurstone's model is appropriate when rankers agree on the values of a set of latent attributes but differ in weighing these attributes for their overall judgments. Despite its conceptual appeal, only few applications of Thurstone's ranking model are reported in the literature because parameter estimation requires high-dimensional numerical integration. Gamma or Gumbel distributions (Luce, 1959; Stern, 1990) provide computationally attractive, albeit more restrictive, alternatives to Thurstone's specification of a multivariate normal distribution.

The assumption of a preference structure that is common to all rankers is not always adequate and can be rather limiting. For instance, in elections it is usual to observe several groups of voters with distinct opinions about the political candidates. In this case, it is more appropriate to represent the heterogeneous population of respondents by a finite mixture of subpopulations

---

The manuscript for this article was submitted and accepted during my tenure as the Editor of *Psychometrika*. — Willem Heiser

This research was partially supported by NSF grant SBR-9730197. The author is grateful to Rung-Ching Tsai and three anonymous reviewers for their helpful comments on this research.

Requests for reprints should be sent to Ulf Böckenholt, Department of Psychology, University of Illinois, Champaign, IL 61820. E-Mail: ubockenh@s.psych.uiuc.edu

0033-3123/2001-1/1998-0682-A $00.75/0

PSYCHOMETRIKA—VOL. 66, NO. 1, 63–68
MARCH 2001

THE PROPER SEQUENCE FOR CORRECTING CORRELATION COEFFICIENTS FOR RANGE RESTRICTION AND UNRELIABILITY

JOSEPH M. STAUFFER

CENTER FOR WOMEN IN SMALL BUSINESS
SAINT MARY-OF-THE-WOODS COLLEGE

JORGE L. MENDOZA

UNIVERSITY OF OKLAHOMA

Corrections of correlations for range restriction (i.e., selection) and unreliability are common in psychometric work. The current rule of thumb for determining the order in which to apply these corrections looks to the nature of the reliability estimate (i.e., restricted or unrestricted). While intuitive, this rule of thumb is untenable when the correction includes the variable upon which selection is made, as is generally the case. Using classical test theory, we show that it is the nature of the range restriction, not the nature of the available reliability coefficient, that determines the sequence for applying corrections for range restriction and unreliability.

Key words: correlation, range restriction, reliability, selection.

Basic formulas for correcting correlation coefficients for unreliability and range restriction (also known as selection or curtailment; Thorndike, 1949) have been available for nearly a century (Pearson, 1903; Spearman, 1904) and, consequently, are well known and frequently applied. Many applications such as test validation, validity generalization, and psychometric meta-analysis attempt to use these formulas jointly (see, e.g., Bobko, 1983; Callender & Osburn, 1980; Hunter & Schmidt, 1990, 1994; Lee, Miller, & Graham, 1982; Mendoza & Mumford, 1987; Raju & Burke, 1983). However, because the correction formulas for range restriction and unreliability were derived separately, careful attention must be paid to the sequence with which these corrections are made. Under current practice, the sequence with which these psychometric corrections are applied is determined using this basic rule of thumb: *If the reliability coefficient is itself restricted, correct the correlation for unreliability before correcting for range restriction; otherwise, correct the correlation for range restriction and then for unreliability*. This note demonstrates that this rule of thumb, while intuitive, is tenable only when the correction does not include the variable upon which selection occurs. We further show that it is the nature of the range restriction, not the nature of the reliability, that determines the sequence with which these joint corrections should be made.

1. Currently Applied Joint Correction Formulas

We begin by defining $x = t + e$, where x is an observed variable and t and e are its latent components, as the variable upon which selection is made. The latent variable t is the true score component of x, and e represents random measurement error. Using uppercase letters to indicate unrestricted values, we define the unrestricted correlation between t and some variable, y, as the

---

We would like to thank Malcolm James Ree for his encouragement and helpful comments as well as those of the editors, associate editor, and reviewers.

Requests for reprints should be sent to Joseph M. Stauffer, c/o Division Management, University of Oklahoma, 307 W. Brooks, Adams Hall, Room 206, Norman OK 73019. E-Mail: joestauffer@hotmail.com

0033-3123/2001-1/1995-0410-A $00.75/0

PSYCHOMETRIKA—VOL. 66, NO. 1, 69–77
MARCH 2001

ON MAXIMIZING ITEM INFORMATION AND MATCHING DIFFICULTY WITH ABILITY

PETER BICKEL

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA AT BERKELEY

STEVEN BUYSKE

DEPARTMENT OF STATISTICS
RUTGERS UNIVERSITY

HUAHUA CHANG

NATIONAL BOARD OF MEDICAL EXAMINERS

ZHILIANG YING

DEPARTMENT OF STATISTICS
RUTGERS UNIVERSITY

An important assumption in IRT model-based adaptive testing is that matching difficulty levels of test items with an examinee's ability makes a test more efficient. According to Lord, "An examinee is measured most effectively when the test items are neither too difficult nor too easy for him". This assumption is examined and challenged through a class of one-parameter IRT models including those of Rasch and the normal ogive. It is found that for a specific model, the validity of the fundamental assumption is closely related to the so-called van Zwet tail ordering of symmetric distributions. In this connection, the cosine distribution serves as the borderline between those satisfying the assumption and those violating the assumption. Graphic and numerical illustrations are presented to demonstrate the theoretic results.

Key words: item response theory, item characteristic curve, computerized adaptive testing, Rasch model, normal ogive, one-parameter IRT model, item information, van Zwet tail ordering.

1. Introduction

Computerized adaptive testing (CAT) was initially conceived by Lord (1970, 1971) to tailor a test according to an examinee's ability. An important assumption in the theoretical foundation for item response theory (IRT) model-based CAT is that matching difficulty levels of test items with an examinee's ability makes a test more efficient. In the words of Lord (1970, p. 139), "An examinee is measured most effectively when the test items are neither too difficult nor too easy for him". Because in adding an item to a test, the improvement of accuracy is an increasing function of the item information, this assumption amounts to claiming that the item information is maximized when its difficulty level matches the examinee's ability. That claim, for binary items, is the subject of this paper.

Research by Peter Bickel supported by NSF Grant DMS 95-04955. Research by Steven Buyske supported by ETS Psychometric Fellowship. Research by Huahua Chang supported by ETS research allocation PJ 79427. Research by Zhiliang Ying supported by ETS Visiting Scholar Program, NSF Grant DMS 96-26750, and NSA Grant MDA 96-1-0034.

We would like to thank the Associate Editor and two referees for their helpful and constructive comments, which led to many improvements. We also thank Ming-Mei Wang and Eiji Muraki for relating our work to Samejima (1979), Fumiko Samejima for sending us her manuscript, and Charles Davis for helpful conversations.

Requests for reprints should be sent to Zhiliang Ying, Department of Statistics, Hill Center, Busch Campus, Rutgers University, Piscataway NJ 08854-8019. E-Mail: zying@stat.rutgers.edu

0033-3123/2001-1/1997-0554-A $00.75/0

PSYCHOMETRIKA—VOL. 66, NO. 1, 79–97
MARCH 2001

GENERAL ABILITY MEASUREMENT: AN APPLICATION OF MULTIDIMENSIONAL ITEM RESPONSE THEORY

DANIEL O. SEGALL

DEFENSE MANPOWER DATA CENTER

Two new methods for improving the measurement precision of a general test factor are proposed and evaluated. One new method provides a multidimensional item response theory estimate obtained from conventional administrations of multiple-choice test items that span general and nuisance dimensions. The other method chooses items adaptively to maximize the precision of the general ability score. Both methods display substantial increases in precision over alternative item selection and scoring procedures. Results suggest that the use of these new testing methods may significantly enhance the prediction of learning and performance in instances where standardized tests are currently used.

Key words: computerized adaptive testing, general cognitive ability, general mental ability, multidimensional adaptive testing, multidimensional item response theory, personnel selection, tests.

During most of the 20th century, the role of general cognitive ability (g) for predicting future learning and job-performance has been hotly disputed. Schmidt and Hunter (1998) summarize 85 years of validity research, stating that the most well known conclusion from this work is that "for hiring employees without previous experience in the job the most valid predictor of future performance and learning is general mental ability" (p. 262). Ree and Earles (1991a, 1992, 1994) arrive at a similar conclusion for the prediction of both training and job success. These conclusions emphasize the importance and usefulness of well constructed measures of general ability. It follows that more precise measures of g will lead to a number of desirable personnel selection outcomes, including increases in employee performance, and increased learning of job-related skills (Hunter, Schmidt, & Judiesch, 1990).

The precursor to modern ability theory (Spearman, 1904) states that the variation in error-free mental measurements is due to two types of factors. One type consists of a single factor g termed general cognitive ability which is common to all mental ability measurements. The second type of factors s are assumed to exist for each individual test. Given that s and g are uncorrelated, and the s's are uncorrelated with each other, Spearman's two-factor formulation suggests that a composite score formed from a number of tests will have more g than any of the individual components used to form the composite. Although Spearman's two-factor theory has been largely supplanted by hierarchical models of intelligence[1] (which allow for the existence of intermediate group factors), g is widely assumed by modern cognitive theorists to exist at the apex of all mental measurements. Accordingly the prevailing approach to general cognitive ability measurement is consistent with Spearman's original formulation. This approach attempts to average out much of the variance due to the unique demands made by each test or test-grouping by computing composite scores from a large number of highly diverse tests.

Since Spearman's early insight, progress towards the construction of a perfect measure of general cognitive ability has been disappointing. This lack of progress has been partly due to

The author wishes to thank the three anonymous reviewers for their useful comments on an earlier version of this manuscript. The views expressed are those of the author and not necessarily those of the Department of Defense, or the United States government.

Requests for reprints should be sent to Daniel O. Segall, Defense Manpower Data Center, DoD Center Monterey Bay 400 Gigling Road, Seaside, CA 93955-6771. E-Mail: publications@danielsegall.com

---

[1] See Lohman (1989) and Mislevy (1994) for other modern perspectives about cognitive psychological measurement and the nature of cognitive abilities.

0033-3123/2001-1/1999-0723-A $00.75/0

PSYCHOMETRIKA—VOL. 66, NO. 1, 99–107
MARCH 2001

ROTATION IN THE DYNAMIC FACTOR MODELING OF MULTIVARIATE STATIONARY TIME SERIES

PETER C.M. MOLENAAR

DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF AMSTERDAM

JOHN R. NESSELROADE

DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF VIRGINIA

A special rotation procedure is proposed for the exploratory dynamic factor model for stationary multivariate time series. The rotation procedure applies separately to each univariate component series of a q-variate latent factor series and transforms such a component, initially represented as white noise, into a univariate moving-average. This is accomplished by minimizing a so-called state-space criterion that penalizes deviations of the rotated solution from a generalized state-space model with only instantaneous factor loadings. Alternative criteria are discussed in the closing section. The results of an empirical application are presented in some detail.

Key words: dynamic factor model, identifiability, polynomial division, moving-average, rotation criteria.

Application of the common factor model to multivariate time series obtained from a single case (subject, system) has been established in the psychological literature for over half a century (e.g., P-technique factor analysis, Cattell, Cattell, & Rhymer, 1947). One also finds related applications in the social science literature (Engle & Watson, 1981; Geweke & Singleton, 1981). It has also long been known that adaptations of the traditional common factor model are required in order to exploit the riches of time series data, for example, to explain the lagged covariance structure of manifest variables (Anderson, 1963; Cattell, 1963; Holtzman, 1963). Molenaar (1985) introduced a dynamic factor model that can handle stationary time series and can be fitted by means of standard structural equation modeling (SEM) software (see also Nesselroade & Molenaar, 1999; Wood & Brown, 1994). Although these more advanced modeling techniques have been in existence now for well over a decade, their ability to accommodate some of the more subtle features of factor analysis, especially exploratory factor analysis, has been greatly limited. In this article we will address one of the classical exploratory factor analysis problems rotation by presenting a special form of rotation for exploratory dynamic factor analysis.

On the one hand, the rotation method we present is special in that it applies separately to each of the q univariate latent factor series in a dynamic q-factor model. In particular, the method applies to a dynamic 1-factor model and hence involves a kind of rotation for which there is no analog in traditional factor analysis. On the other hand, our rotation method accomplishes two results that are in keeping with those of rotation in traditional exploratory factor analysis. First, the rotation "simplifies" the factor loading pattern. Second, the accompanying implied transformation of the factor scores induces properties into the latter that render them compatible with the new loading pattern. To denote the fact that the transformation we develop and apply is, strictly speaking, not a rotation of axes in the usual factor analytic sense, we will consistently refer to it as a special rotation. A merit of our proposal is that the properties induced into the

---

This research was supported by the Institute for Developmental and Health Research Methodology, University of Virginia.

Requests for reprints should be sent to Peter C.M. Molenaar, Department of Psychology, University of Amsterdam, Roetersstraat 15, room 502, 1018 WB Amsterdam, THE NETHERLANDS. E-Mail: op_molenaar@macmail.psy.uva.nl

0033-3123/2001-1/1998-0627-A $00.75/0

PSYCHOMETRIKA—VOL. 66, NO. 1, 109–132
MARCH 2001

TESTING FOR LOCAL DEPENDENCY IN DICHOTOMOUS AND POLYTOMOUS ITEM RESPONSE MODELS

EDWARD HAK-SING IP

MARSHALL SCHOOL OF BUSINESS
UNIVERSITY OF SOUTHERN CALIFORNIA

Researchers studying item response models are often interested in examining the effects of local dependency on the validity of the resulting conclusion from statistical inference. This paper focuses on the detection of local dependency. We provide a framework for viewing local dependency within dichotomous and polytomous items that are clustered by design, and present a testing procedure that allows researchers to specifically identify individual item pairs that exhibit local dependency, while controlling for false positive rate. Simulation results from the study indicate that the proposed method is effective. In addition, a discussion of its relation to other existing methods is provided.

Key words: item response theory, item cluster, Mantel-Haenzsel test, multiple inference, NAEP, DIMTEST, Yen's Q_3.

1. Introduction

For more than a decade, the item response theory (IRT) has provided a framework under which dichotomous and polytomous responses to items can be modeled under a specific set of assumptions. The popularity of IRT models can be witnessed by the large amount of IRT-related literature in psychometrics and its widespread use in large-scale measurement programs. Unfortunately, as pointed out by Harwell, Stone, Hsu and Kirisci (1996), "the enthusiasm for IRT has been tempered by the realization that the validity with which these methods can be applied to realistic data sets (e.g., small numbers of items and examinees, multidimensional data) is often poorly documented".

This paper concerns a common design that occurs in realistic datasets for which the validity of IRT -based method may not be justified. In this design, a test typically consists of one or more item clusters. Jannarone (1992a) describes an example in which adjacent items are semantically linked through common key words. The first item introduces a word and a test for word comprehension. The second item introduces a new word and requires the word and the one previously tested in item 1 to be correctly comprehended in order to pass this item. The third item similarly tests whether both the new word from the current item and the one from item 2 are understood, and so on. The entire test consists of one or more serially correlated item-chains.

Another example of the presence of clusters in a test occurs when two or more items are related to a specific task, such as the reading of a literary or scientific passage. The presence of reading passages in tests is common in psychological tests and large-scale assessment programs such as the Law School Admission Test (LSAT), the National Assessment of Educational Progress (NAEP) Reading assessment, and the Advanced Placement (AP) test of English Literature.

The local independence (LI) assumption in IRT models, which states that item responses given ability are independent, seems hardly defensible in the kind of test design we have just

---

The research was supported under the National Assessment of Educational Progress (Grant No. R902B990007) administered by the National Center of Education Statistics, U.S. Department of Education. This work was started when the author was at the Division of Statistics and Psychometrics at the Educational Testing Service. I thank Juliet Shaffer for her comments on the multiple testing procedure. I also thank three anonymous referees and the Associate Editor for suggestions that greatly improved the presentation of the manuscript.

Request for reprints should be directed to the author at the Marshall School of Business, Information and Operations Management Department, University of Southern California, Los Angeles, CA 90089-1421. E-Mail: eip@bus.usc.edu

0033-3123/2001-1/1996-0493-A $00.75/0

PSYCHOMETRIKA—VOL. 66, NO. 1, 133–152
MARCH 2001

A BAYESIAN ANALYSIS OF FINITE MIXTURES IN THE LISREL MODEL

HONG-TU ZHU AND SIK-YUM LEE

DEPARTMENT OF STATISTICS
THE CHINESE UNIVERSITY OF HONG KONG

In this paper, we propose a Bayesian framework for estimating finite mixtures of the LISREL model. The basic idea in our analysis is to augment the observed data of the manifest variables with the latent variables and the allocation variables. The Gibbs sampler is implemented to obtain the Bayesian solution. Other associated statistical inferences, such as the direct estimation of the latent variables, establishment of a goodness-of-fit assessment for a posited model, Bayesian classification, residual and outlier analyses, are discussed. The methodology is illustrated with a simulation study and a real example.

Key words: Bayesian analysis, finite mixtures, LISREL models, Gibbs sampler, conditional distributions, goodness-of-fit assessment, Bayesian classification, residual and outlier analyses.

1. Introduction

In general, a finite mixture model arises with a population which is a mixture of M components with associated probability densities $\{f_m, m = 1, \ldots, M\}$ and mixing proportions $\{\pi_m, m = 1, \ldots, M\}$. Such a situation is very common in many areas of applied statistics such as statistical pattern recognition, classification and clustering; see the survey paper by Redner and Walker (1984), and the excellent book by Titterington, Smith and Makov (1985), among others. In the literature, a variety of statistical methods have been proposed to analyze finite mixture models. Examples include the method of moments (Day, 1969; Lindsay, 1989; Lindsay & Basak, 1993), Bayesian and quasi-Bayes methods (Crawford, DeGroot, Kadane, & Small, 1992; Diebolt & Robert, 1994; Richardson & Green, 1997; Robert, 1996; Smith & Makov, 1978), the discriminant analysis approach (Do & McLachlan, 1984; Ganesalingam & McLachlan, 1981), and the maximum likelihood (Day, 1969; Hasselblad, 1966; Hathaway, 1985).

On the other hand, models for establishing substantive theory in behavioral and social sciences usually involve causal effects and correlations among manifest variables and the latent variables that cannot be measured by one single operationalization. Structural equation modeling (SEM) is an important method in finding the appropriate model and estimating the causal effects and the correlations, see Bentler (1983), Browne (1984), Jöreskog (1978), and Yuan and Bentler (1997), among others. Now, SEM represents a widely used multivariate method in behavioral, health and social sciences; see for examples, Bollen and Long (1993), Byrne (1994), Hoyle (1995) and the references therein. The most important factor accounting for the popularity of SEM is clearly due to the availability of the efficient computer softwares such as EQS (Bentler, 1992) and LISREL VIII (Jöreskog & Sörbom, 1996). Hence, a major trend of research in the field has been devoted to enlarge the scope of applicability of the LISREL model to non-standard situations; see for example, Lee and Poon (1992), Muthén (1989) and van Buuren (1997), among others.

Recently, a few important contributions on the maximum likelihood (ML) analysis of finite mixtures in structural equation models have been established. Jedidi, Jagpal and DeSarbo (1997a)

This research was supported by a Hong Kong UGC Earmarked grant CUHK 4026/97H. The authors are indebted to the Editor, the Associate Editor, and three anonymous reviewers for constructive comments in improving the paper, and also to ICPSR and the relevant funding agency for allowing the use of the data. The assistance of Michael K.H. Leung and Esther L.S. Tam is gratefully acknowledged.

Requests for reprints should be sent to S.Y. Lee, Department of Statistics, The Chinese University of Hong Kong, Shatin, N.T. HONG KONG. E-Mail: sylee@sparc2.sta.cuhk.edu.hk

0033-3123/2001-1/1998-0694-A $00.75/0

PSYCHOMETRIKA—VOL. 66, NO. 1, 153–155
MARCH 2001
REVIEWS

REVIEW

Susan E. Embretson and Scott L. Hershberger (Eds.). *The new rules of measurement: What every psychologist and educator should know*. Mahwah, NJ: Lawrence Erlbaum, 263 pp., 1999, ISBN 0-8058-2860-5.

The objective of the editors of this volume (Embretson and Hershberger) is to supplement "contemporary textbooks on measurement and testing by explicating some neglected topics" (preface, p. VIII), which must be known to every psychologist and educator. The editors consider Item Response Theory (IRT) as the most neglected topic of test development, and the emphasis of the book is on IRT. The book has eleven chapters: Six on the measurement of cognitive abilities, four on personality measurement, and a closing chapter.

Embretson (chapter 1) introduces the chapters on the measurement of cognitive abilities. She gives a brief introduction to IRT, which includes a very nice, intuitive, explanation of the maximum likelihood method for estimating persons' latent trait values. Moreover, she discusses four general IRT rules, which every test developer should know. From this chapter, it immediately becomes clear that the volume is mainly focused on a special type of IRT; that is, models having a continuous latent variable (latent trait) to explain discrete observed item responses. Thorndike (chapter 2) discusses early work on ability measurement by, among others, his grandfather and father. Father and son made the fourth revision of the Stanford-Binet (SB-IV) test. The SB-IV differs from the previous editions of the test in two respects. First, the 15 subtests are organized by content instead of difficulty level and, second, IRT (the Rasch model) was used to construct the subtests. Finally, Thorndike discusses some directions of psychological measurement, which will probably continue in the future. One trend is in the direction of computer-assisted test administration, which has many advantages above conventional test administration procedures. Another trend is in the direction of the use of ability scales, which have a latent-trait referenced interpretation instead of the conventional norm-referenced interpretation. Latent-trait referenced interpretations of test scores come up again in Woodcock's contribution (chapter 5). Daniel (chapter 3) demonstrates the use of IRT methods on the Differential Ability Scales (DAS) and the Kaufman Adolescent and Adult Intelligence Test (KAIT). Wright's contribution (chapter 4) reads like an advertisement text on the Rasch model. This chapter, however, does not refer to most of the recent developments, which were mainly made by European researchers (e.g., Andersen, Fischer, Glas, Kelderman, Molenaar, Rost, Verhelst). Woodcock (chapter 5) describes methods for reporting test scores under the Rasch model. His contribution highlights the usefulness of IRT for reporting test scores to test practitioners. Marcoulides' contribution (chapter 6) deviates from the other chapters because it is on Generalizability (G) Theory. He gives an introduction to G-theory for one- and multi-facet designs. G-theory is a generalization of Classical Test Theory (CTT), but Marcoulides relates G-theory to IRT. He gives some nice examples, but the text fails to make clear the theoretical relation between G-theory and IRT. Hershberger (chapter 7) introduces the chapters on personality measurement. He mentions three important themes: (1) the construct validity of personality tests, (2) the dimensional versus taxonomic conceptualization of personality characteristics, and (3) individual differences in the relevance of trait labels ("Some people are simply more traited than others", p. 154). Exner (chapter 8) reports on properties of the Comprehensive System of coding Rorschach protocols. He reports favorable results on intercoder agreement and test-retest reliability of some of the Rorschach codes. Moreover, he reports between-group differences in the frequency distributions of some of the codes. The author does not discuss the application of IRT to Rorschach data. A long time ago, Fischer and Spada (1973) used the Rasch model to describe Rorschach responses, but their work is rather unknown

0033-3123/2001-1/2000-0086-R $00.75/0

Psychometrika

Psychometrika, the official journal of the Psychometric Society, is devoted to the development of psychology as a quantitative rational science. Issued four times a year, on March 15, June 15, September 15, and December 15.

MARCH 2002, VOLUME 67, NUMBER 1

(ISSN 0033-3123) is published by the Psychometric Society, 326 Cox Statistics Building, Research Triangle Institute, P.O. Box 12194, Research Triangle Park, NC 27709-2194 U.S.A. and at additional mailing offices. Editorial office: *Psychometrika*, University of North Carolina at Greensboro, P.O. Box 26171, Greensboro, NC 27402-6171 U.S.A (e-mail: Pmetrika@s.psych.uiuc.edu). Managing Editor: P.O. Box 7104, San Jose, CA 95150-7104 U.S.A. (e-mail: Pmetrika@nullset.org).

Subscription Price: The regular subscription rate is $125.00 per volume. All annual subscriptions start with the March issue and cover the calendar year. All bills and orders are payable in advance. Checks should be made payable to the *Psychometric Society*. All checks, notices of change of address, and business communications (including notification of the nonreceipt of journals) should be sent to the Treasurer. Membership in the Psychometric Society includes a subscription to *Psychometrika*. Regular members pay annual dues of $60.00, except that family members pay annual dues of $75.00. The annual dues for a student in a recognized college or university is $25.00.

Application for membership and student membership in the Psychometric Society, together with a check for dues for the calendar year in which application is made, should be sent to the Treasurer.

Back Issues: Back issues for the previous 6 volumes only are available from the Treasurer at the address below. Persons claiming replacement for issues lost in the mail must do so with the Treasurer within one year. Microfilm copies of all volumes are available from Johnson Associates, Inc., P.O. Box 1017, Greenwich, CT, 06830, and Bell & Howell Information and Learning, P.O. Box 1346, Ann Arbor, MI 48106-1346 (http://wwwlib.umi.com/sim/). *Psychometrika* is abstracted in Statistical Theory and Method Abstracts (STMA).

LORI D. MCLEOD, Treasurer
Psychometric Society
326 Cox Statistics Building
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709-2194 USA

(Continued on the back inside cover page)

Psychometrika contains articles on the development of quantitative models for psychological phenomena, and on quantitative methodology in the social and behavioral sciences, including new mathematical and statistical techniques for the evaluation of psychological data, and the application of such techniques. Empirical studies will be considered only if they involve new or particularly interesting uses of quantitative techniques.

Submission of a paper to this journal will be taken to imply that it represents original work not previously published, and that it is not being considered elsewhere for publication. Authors submitting a manuscript do so on the understanding that if it is accepted for publication, copyright in the aticle, including the right to reproduce the article in all forms and media, shall be assigned exclusively to the Publisher. The Publisher will not refuse any reasonable request by the author for permission to reproduce any of his or her contributions to the journal.

The editor removes, as far as possible, the author's identity from manuscripts before sending them for review.

Paper submissions must include five copies of a manuscript. (Authors wishing to make an electronic submission should refer to the submission information posted on the Journal's web site.) Prospective authors are referred to the "Instructions to *Psychometrika* Authors" contained in the September 1994 issue. Reprints are available from the Editor upon request. A manuscript which fails to comply with these requirements will be returned to the author for revision. Authors will receive 100 reprints without covers, free of charge.

Manuscripts for publication in *Psychometrika* should be sent to:

Ulf Böckenholt, Editor
Psychometrika
Univ. of N.C. at Greensboro
210 Curry Building
P.O. Box 26171
Greensboro, NC 27402

Books for review in *Psychometrika* should be sent to:

Laura M. Koehly, Review Editor
Psychometrika
Department of Psychology
Texas A&M University
College Station, TX 77843-4235

Electronic Addresses:

Ulf Böckenholt, Editor: Pmetrika@uncg.edu
Cynthia H. Null, Managing Editor: Pmetrika@nullset.org
Tim Null, Technical Editor: Pmetrika@nullset.org
Laura M. Koehly, Review Editor: lmk@psyc.tamu.edu
24 Hour Fax/Voice Mail: 1 (888) 467-4101
24 Hour Fax (Paris): +331 5301 2721
Journal web site: http://www.psychometrika.org
Psychometric Society web site: http://www.psychometricsociety.org

The Psychometric Society

PRINTED BY CADMUS PROFESSIONAL COMMUNICATION—BYRD PRESS DIVISION
COMPOSITION BY INTEGRE TECHNICAL PUBLISHING CO., INC.

VOLUME 67
NUMBER 1
MARCH 2002

ISSN 0033-3123

Psychometrika

A JOURNAL OF QUANTITATIVE PSYCHOLOGY

THE PSYCHOMETRIC SOCIETY

MARCH 2002 PSYCHOMETRIKA VOLUME 67 NO. 1 1–184