Statistical Factor Analysis and Related Methods

Theory and Applications

ALEXANDER BASILEVSKY

Department of Mathematics & Statistics
The University of Winnipeg
Winnipeg, Manitoba
Canada

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Brisbane • Toronto • Singapore

This Page Intentionally Left Blank

Statistical Factor Analysis and Related Methods

This Page Intentionally Left Blank

Statistical Factor Analysis and Related Methods

Theory and Applications

ALEXANDER BASILEVSKY

Department of Mathematics & Statistics
The University of Winnipeg
Winnipeg, Manitoba
Canada

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Brisbane • Toronto • Singapore

A NOTE TO THE READER
This book has been electronically reproduced from digital information stored at John Wiley & Sons, Inc. We are pleased that the use of this new technology will enable us to keep works of enduring scholarly value in print as long as there is a reasonable demand for them. The content of this book is identical to previous printings.

This text is printed on acid-free paper.

Library of Congress Cataloging in Publication Data:

Basilevsky, Alexander.

Statistical factor analysis and related methods : theory and applications / by Alexander Basilevsky.

p. cm. -- Wiley series in probability and mathematical statistics. Probability and mathematical statistics)

"A Wiley-Interscience publication."

Includes bibliographical references and index.

ISBN 0-471-57082-6 (acid-free)

1. Factor analysis. I. Title. II. Series.

QA278.5.B37 1993

519.5'354--dc20 93-2323

To my mother Olha
and
To the memory of my father Mykola

This Page Intentionally Left Blank

If one is satisfied, as he should be, with that which is to be probable, no difficulty arises in connection with those things that admit of more than one explanation in harmony with the evidence of the senses; but if one accepts one explanation and rejects another that is equally in agreement with the evidence it is clear that he is altogether rejecting science and taking refuge in myth.

– Epicurus (Letter to Pythocles, Fourth Century B.C.)

Physical concepts are free creations of the human mind, and are not, however it may seem, uniquely determined by the external world. In our endeavour to understand reality we are somewhat like a man trying to understand the mechanism of a closed watch. He sees the face and the moving hands, even hears its ticking, but he has no way of opening the case. If he is ingenious he may form some picture of a mechanism which could be responsible for all the things he observes, but he may never be quite sure his picture is the only one which could explain his observations. He will never be able to compare his picture with the real mechanism and he cannot even imagine the possibility of the meaning of such a comparison.

– A. Einstein, *The Evolution of Physics*, 1938

This Page Intentionally Left Blank

Preface

More so than other classes of statistical multivariate methods, factor analysis has suffered a somewhat curious fate in the statistical literature. In spite of its popularity among research workers in virtually every scientific endeavor (e.g., see Francis, 1974), it has received little corresponding attention among mathematical statisticians, and continues to engender debate concerning its validity and appropriateness. An equivalent fate also seems to be shared by the wider class of procedures known as latent variables models. Thus although high-speed electronic computers, together with efficient numerical methods, have solved most difficulties associated with fitting and estimation, doubt at times persists about what is perceived to be an apparent subjectiveness and arbitrariness of the methods (see Chatfield and Collins, 1980, p. 88). In the words of a recent reviewer, "They have not converted me to thinking factor analysis is worth the time necessary to understand it and carry it out." (Hills, 1977.)

Paradoxically, on the more applied end of the spectrum, faced with voluminous and complex data structures, empirical workers in the sciences have increasingly turned to data reduction procedures, exploratory methods, graphical techniques, pattern recognition and other related models which directly or indirectly make use of the concept of a latent variable (for examples see Brillinger and Preisler, 1983). In particular, both formal and informal exploratory statistical analyses have recently gained some prominence under such terms as "soft modeling" (Wold, 1980) and "projection pursuit" (Huber, 1985; Friedman and Tukey, 1974). These are tasks to which factor analytic techniques are well suited. Besides being able to reduce large sets of data to more manageable proportions, factor analysis has also evolved into a useful data-analytic tool and has become an invaluable aid to other statistical models such as cluster and discriminant analysis, least squares regression, time/frequency domain stochastic processes, discrete random variables, graphical data displays, and so forth although this is not always recognized in the literature (e.g. Cooper, 1983).

Greater attention to latent variables models on the part of statisticians is now perhaps overdue. This book is an attempt to fill the gap between the mathematical and statistical theory of factor analysis and its scientific practice, in the hope of providing workers with a wider scope of the models than what at times may be perceived in the more specialized literature (e.g. Steward, 1981; Zegura, 1978; Matalas and Reicher, 1967; Rohlf and Sokal, 1962).

The main objections to factor analysis as a bona fide statistical model have stemmed from two sources—historical and methodological. Historically, factor analysis has had a dual development beginning indirectly with the work of Pearson (1898, 1901, 1927), who used what later becomes known as principal components (Hotelling, 1933) to fit "regression" planes to multivariate data when both dependent and independent variables are subject to error. Also, Fisher used the so-called singular value decomposition in the context of ANOVA (Fisher and Mackenzie, 1923). This was the beginning of what may be termed the statistical tradition of factor analysis, although it is clearly implicit in Bravais' (1846) original development of the multivariate normal distribution, as well as the mathematical theory of characteristic (eigen) roots and characteristic (eigen) vectors of linear transformations. Soon after Hotelling's work Lawley (1940) introduced the maximum likelihood factor model. It was Spearman (1904, 1913), however, who first used the term "factor analysis" in the context of psychological testing for "general intelligence" and who is generally credited (mainly in psychology) for the origins of the model. Although Spearman's method of "tetrads" represented an adaptation of correlation analysis, it bore little resemblance to what became known as factor analysis in the scientific literature. Indeed, after his death Spearman was challenged as the originator of factor analysis by the psychologist Burt, who pointed out that Spearman had not used a proper factor model, as Pearson (1901) had done. Consequently, Burt was the originator of the psychological applications of the technique (Hearnshaw, 1979). It was not until later however that factor analysis found wide application in the engineering, medical, biological, and other natural sciences and was put on a more rigorous footing by Hotelling, Lawley, Anderson, Joreskog, and others. An early exposition was also given by Kendall (1950) and Kendall and Lawley (1956). Because of the computation involved, it was only with the advent of electronic computers that factor analysis became feasible in everyday applications.

Early uses of factor analysis in psychology and related areas relied heavily on linguistic labeling and subjective interpretation (perhaps Cattell, 1949 and Eysenck, 1951 are the best known examples) and this tended to create a distinct impression among statisticians that imposing a particular set of values and terminology was part and parcel of the models. Also, questionable psychological and eugenic attempts to use factor analysis to measure innate (i.e., genetically based) "intelligence," together with Burt's fraudulent publications concerning twins (e.g., see Gould, 1981) tended to further alienate scientists and statisticians from the model. Paradoxically, the rejection has engendered its own misunderstandings and confusion amongst statisticians (e.g., see Ehrenberg, 1962; Armstrong, 1967; Hills, 1977), which seems to have prompted some authors of popular texts on multivariate analysis to warn readers of the "... many drawbacks to factor analysis" (Chatfield and Collins, 1980, p. 88). Such misunderstandings have had a further second-order impact on practitioners (e.g., Mager, 1988, p. 312).

Methodological objections to factor analysis rest essentially on two criteria. First, since factors can be subjected to secondary transformations of the coordinate axes, it is difficult to decide which set of factors is appropriate. The number of such rotational transformations (orthogonal or oblique) is infinite, and any solution chosen is, mathematically speaking, arbitrary. Second, the variables that we identify with the factors are almost never observed directly. Indeed, in many situations they are, for all practical intents and purposes, unobservable. This raises a question concerning exactly what factors do estimate, and whether the accompanying identification process is inherently subjective and unscientific. Such objections are substantial and fundamental, and should be addressed by any text that deals with latent variables models. The first objection can be met in a relatively straightforward manner, owing to its somewhat narrow technical nature, by observing that no estimator is ever definitionally unique unless restricted in some suitable manner. This is because statistical modeling of the empirical world involves not only the selection of an appropriate mathematical procedure, with all its assumptions, but also consists of a careful evaluation of the physical-empirical conditions that have given rise to, or can be identified with, the particular operative mechanism under study. It is thus not only the responsibility of mathematical theory to provide us with a unique statistical estimator, but rather the arbitrary nature of mathematical assumptions enables the investigator to choose an appropriate model or estimation technique, the choice being determined largely by the actual conditions at hand. For example, the ordinary least squares regression estimator is one out of infinitely many regression estimators which is possible since it is derived from a set of specific assumptions, one being that the projection of the dependent variable/vector onto a sample subspace spanned by the independent (explanatory) variables is orthogonal. Of course, should orthogonality not be appropriate, statisticians have little compunction about altering the assumption and replacing ordinary least squares with a more general model. The choice is largely based on prevailing conditions and objectives, and far from denoting an ill-defined situation the existence of alternative estimation techniques contributes to the inherent flexibility and power of statistical/mathematical modeling.

An equivalent situation also exists in factor analysis, where coefficients may be estimated under several different assumptions, for example, by an oblique rather than an orthogonal model since an initial solution can always be rotated subsequently to an alternative basis should this be required. Although transformation of the axes is possible with any statistical model (the choice of a particular coordinate system is mathematically arbitrary), in factor analysis such transformations assume particular importance in some (but not all) empirical investigations. The transformations, however, are not an inherent feature of factor analysis or other latent variable(s) models, and need only be employed in fairly specific situations, for example, when attempting to identify clusters in the variable (sample) space. Here, the coordinate axes of an initial factor solution usually represent mathematically arbitrary frames of references which are chosen on grounds of convenience and east of computation, and which may have to be altered because of interpretational or substantive requirements. The task is much simplified, however, by the existence of well-defined statistical criteria which result in unique rotations, as well as by the availability of numerical algorithms for their implementation. Thus once a criterion function is selected and optimized, a unique set of estimated coefficients (coordinate axes) emerges. In this sense the rotation of factors conforms to general and accepted statistical practice. Therefore, contrary to claims such as those of Ehrenberg (1962) and Temple (1978), our position on the matter is that the rotation of factors is not intrinsically subjective in nature and, on the contrary, can result in a useful and meaningful analysis. This is not to say that the rotational problem represents the sole preoccupation of factor analysis. On the contrary, in some applications the factors do not have to be rotated or undergo direct empirical interpretation. Frequently they are only required to serve as instrumental variables, for example, to overcome estimation difficulties in least squares regression. Unlike the explanatory variables in a regression model, the factor scores are not observed directly and must also be estimated from the data. Again, well-defined estimators exist, the choice of which depends on the particular factor model used.

The second major objection encountered in the statistical literature concerns the interpretation of factors as actual variables, capable of being identified with real or concrete phenomenon. Since factors essentially represent linear functions of the observed variables (or their transformations), they are not generally observable directly, and are thus at times deemed to lack the same degree of concreteness or authenticity as variables measured in a direct fashion. Thus, although factors may be seen as serving a useful role in resolving this estimation difficulty or that measurement problem, they are at times viewed as nothing more than mathematical artifacts created by the model. The gist of the critique is not without foundation, since misapplication of the model is not uncommon. There is a difficulty, however, in accepting the argument that just because factors are not directly observable they are bereft of all "reality." Such a viewpoint seems to equate the concept of reality with that of direct observability (in principle or otherwise), a dubious and inoperative criterion at best, since many of our observations emanate from indirect sources. Likewise, whether factors correspond to real phenomena is essentially an empirical rather than a mathematical question, and depends in practice on the nature of the data, the skill of the practitioner, and the area of application. For example, it is important to bear in mind that correlation does not necessarily imply direct causation, or that when nonsensical variables are included in an analysis, particularly under inappropriate assumptions or conditions, very little is accomplished. On the other hand, in carefully directed applications involving the measurement of unobservable or difficult-to-observe variables—such as the true magnitude of an earthquake, extent and/or type of physical pain, political attitudes, empirical index numbers, general size and/or shape of a biological organism, the informational content of a signal or a two-dimensional image—the variables and the data are chosen to reflect specific aspects which are known or hypothesized to be of relevance. Here the retained factors will frequently have a ready and meaningful interpretation in terms of the original measurements, as estimators of some underlying latent trait(s).

Factor analysis can also be used in statistical areas, for example, in estimating time and growth functions, least squares regression models, Kalman filters, and Karhunen–Loève spectral models. Also, for optimal scoring of a contingency table, principal components can be employed to estimate the underlying continuity of a population. Such an analysis (which predates Hotelling's work on principal components—see Chapter 9) can reveal aspects of data which may not be immediately apparent. Of course, in a broader context the activity of measuring unobserved variables, estimating dimensionality of a model, or carrying out exploratory statistical analysis is fairly standard in statistical practice and is not restricted to factor models. Thus spectral analysis of stochastic processes employing the power (cross) spectrum can be regarded as nothing more than a fictitious but useful mathematical construct which reveals the underlying structure of correlated observations. Also, statisticians are frequently faced with the problem of estimating dimensionality of a model, such as the degree of a polynomial regression or the order of an ARMA process. Available data are generally used to provide estimates of missing observations whose original values cannot be observed. Interestingly, recent work using maximum likelihood estimation has confirmed the close relationship between the estimation of missing data and factor analysis, as indicated by the EM algorithm. Finally, the everyday activity of estimating infinite population parameters, such as means or variances, is surely nothing more than the attempt to measure that which is fundamentally hidden from us but which can be partially revealed by careful observation and appropriate theory. Tukey (1979) has provided a broad description of exploratory statistical research as

> ...an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as for those we believe might be there... its tools are secondary to its purposes.

This definition is well suited to factor and other latent variable models and is employed (implicitly or explicitly) in the text.

The time has thus perhaps come for a volume such as this, the purpose of which is to provide a unified treatment of both the theory and practice of factor analysis and latent variables models. The interest of the author in the subject stems from earlier work on latent variables models using historical and social time series, as well as attempts at improving certain least squares regression estimators. The book is also an outcome of postgraduate lectures delivered at the University of Kent (Canterbury) during the 1970s, together with more recent work. The volume is intended for senior undergraduate and postgraduate students with a good background in statistics and mathematics, as well as for research workers in the empirical sciences who may wish to acquaint themselves better with the theory of latent variables models. Although stress is placed on mathematical and statistical theory, this is generally reinforced by examples taken from the various areas of the natural and social sciences as well as engineering and medicine. A rigorous mathematical and statistical treatment seems to be particularly essential in an area such as factor analysis where misconception and misinterpretations still abound. Finally, a few words are in order concerning our usage of the term "factor analysis," which is to be understood in a broad content rather than the more restricted sense at times encountered in the literature. The reason for this usage is to accentuate the common structural features of certain models and to point out essential similarities between them. Although such similarities are not always obvious when dealing with empirical applications, they nevertheless become clear when considering mathematical-statistical properties of the models. Thus the ordinary principal components model, for example, emerges as a special case of the weighted (maximum likelihood) factor model although both models are at times considered to be totally distinct (e.g., see Zegura, 1978). The term "factor analysis" can thus be used to refer to a class of models that includes ordinary principal components, weighted principal components, maximum likelihood factor analysis, certain multidimensional scaling models, dual scaling, correspondence analysis, canonical correlation, and latent class/latent profile analysis. All these have a common feature in that latent root and latent vector decompositions of special matrices are used to locate informative subspaces and estimate underlying dimensions.

This book assumes on the part of the reader some background in calculus, linear algebra, and introductory statistics, although elements of the basics are provided in the first two chapters. These chapters also contain a review of some of the less accessible material on multivariate sampling, measurement and information theory, latent roots and latent vectors in both the real and complex domains, and the real and complex normal distribution. Chapters 3 and 4 describe the classical principal components model and sample-population inference; Chapter 5 treats several extensions and modifications of principal components such as Q and three-mode analysis, weighted principal components, principal components in the complex field, and so forth. Chapter 6 deals with maximum likelihood and weighted factor models together with factor identification, factor rotation, and the estimation of factor scores. Chapters 7–9 cover the use of factor models in conjunction with various types of data such as time series, spatial data, rank orders, nominal variables, directional data, and so forth. This is an area of multivariate theory which is frequently ignored in the statistical literature when dealing with latent variable estimation. Chapter 10 is devoted to applications of factor models to the estimation of functional forms and to least squares regression estimators when dealing with measurement error and/or multicollinearity.

I would like to thank by colleagues H. Howlader of the Department of Mathematics and Statistics, as well as S. Abizadeh, H. Hutton, W. Morgan, and A. Johnson of the Departments of Economics Chemistry, and Anthropology, respectively, for useful discussions and comments, as well as other colleagues at the University of Winnipeg who are too numerous to name. Last but not least I would like to thank Judi Hanson for the many years of patient typing of the various drafts of the manuscript, which was accomplished in the face of much adversity, as well as Glen Koroluk for help with the computations. Thanks are also owed to Rita Campbell and Weldon Hiebert for typing and graphical aid. Of course I alone am responsible for any errors or shortcomings, as well as for views expressed in the book.

Alexander Basilevsky

Winnigep, Manitoba
February 1994

Contents

This Page Intentionally Left Blank

Statistical Factor Analysis and Related Methods

CHAPTER 1

Preliminaries

1.1 INTRODUCTION

Since our early exposure to mathematical thinking we have come to accept the notion of a variable or a quantity that is permitted to vary during a particular context or discussion. In mathematical analysis the notion of a variable is important since it allows general statements to be made about a particular member of a set. Thus the essential nature of a variable consists in its being identifiable with any particular value of its domain, no matter how large that domain may be. In a more applied context, when mathematical equations or formulas are used to model real life phenomena, we must further distinguish between a deterministic variable and a probabilistic or random variable. The former features prominently in any classical description of reality where the universe is seen to evolve according to "exact" or deterministic laws that specify its past, present, and future. This is true, for example, of classical Newtonian mechanics as well as other traditional views which have molded much of our comtemporary thinking and scientific methodology.

Yet we know that in practice ideal conditions never prevail. The world of measurement and observation is never free of error or extraneous, nonessential influences and other purely random variation. Thus laboratory conditions, for example, can never be fully duplicated nor can survey observations ever be fully verified by other researchers. Of course we can always console ourselves with the view that randomness is due to our ignorance of reality and results from our inability to fully control, or comprehend, the environment. The scientific law itself, so the argument goes, does not depend on these nuisance parameters and is therefore fixed, at least in principle. This is the traditional view of the role of randomness in scientific enquiry, and it is still held among some scientific workers today.

Physically real sources of randomness however do appear to exist in the real world. For example, atomic particle emission, statistical thermodynamics, sun spot cycles, as well as genetics and biological evolution all exhibit random behavior over and above measurement error. Thus randomness does not seem to stem only from our ignorance of nature, but also constitutes an important characteristic of reality itself whenever natural or physical processes exhibit instability (see Prigonine and Stengers, 1984). In all cases where behavior is purely or partially random, outcomes of events can only be predicted with a probability measure rather than with perfect certainty. At times this is counterintuitive to our understanding of the real world since we have come to expect laws, expressed as mathematical equations, to describe our world in a perfectly stable and predictable fashion. The existence of randomness in the real world, or in our measurements (or both), implies a need for a science of measurement of discrete and continuous phenomena which can take randomness into account in an explicit fashion. Such a science is the theory of probability and statistics, which proceeds from a theoretical axiomatic basis to the analysis of scientific measurements and observations.

Consider a set of events or a "sample space" S and a subset A of S. The sample space may consist of either discrete elements or may contain subsets of the real line. To each subset A in S we can assign a real number $P(A)$, known as "the probability of the event A." More precisely, the probability of an event can be defined as follows.

Definition 1.1. A probability is a real-valued set function defined on the closed class of all subsets of the sample space S. The value of this function, associated with a subset A of S, is denoted by $P(A)$. The probability $P(A)$ satisfies the following axioms.*

(1) $P(S) = 1$

(2) $P(A) \geq 0$, all A in S

(3) For any r subsets of S we have $P(A_1 \cup A_2 \cup \cdots \cup A_r) = P(A_1) + P(A_2) + \cdots + P(A_r)$ for $A_i \cap A_j = \emptyset$ the empty set, $i \neq j$

From these axioms we can easily deduce that $P(\emptyset) = 0$ and $P(S) = 1$, so that the probability of an event always lies in the closed interval $0 \leq P(A) \leq 1$. Heuristically, a zero probability corresponds to a logically impossible event, whereas a unit probability implies logical certainty.

Definition 1.2. A real variable X is a real valued function whose domain is the sample space S, such that:

(1) The set $\{X \leq x\}$ is an event for any real number x

(2) $P(X = \pm\infty) = 0$

This definition implies a measurement process whereby a real number is assigned to every outcome of an "experiment." A random variable can

* Known as the Kolmogorov axioms.