LATENT STRUCTURE ANALYSIS AT FIFTY

Neil W. Henry, Virginia Commonwealth University
Neil W. Henry, Virginia Commonwealth University, Richmond, VA 23284-2014

Key Words: Latent class, Latent trait, Attitudes, Log-linear model

INTRODUCTION

During World War II multidisciplinary teams were employed by the War Department to study military personnel. Many of these social psychological studies were documented in the four volumes of *The American Soldier: Studies in Social Psychology in WW II,* (Stouffer, 1949-50). One of the major issues confronting these researchers was the validity and reliability of survey instruments and the methodology of survey analysis, and this led to interaction among psychologists, sociologists and statisticians who were concerned with the concept of attitude and the logic of attitude scale construction.

Volume 4 of *The American Soldier*, entitled *Measurement and Prediction,* appeared in 1950 and contained two chapters by Paul F. Lazarsfeld in which he formulated the theory and demonstrated the use of latent structure models. Lazarsfeld later contributed a chapter on latent structure analysis to the monumental work *Psychology: A Study of A Science*, published in 1959, and the developments of twenty-five years were collected and refined in the 1968 reference and textbook, *Latent Structure Analysis* (Lazarsfeld and Henry, 1968).

In the 30 years since then many statisticians, psychologists and sociologists have continued to study the models contained under the latent structure analysis umbrella. In this paper I will review some of the major contributions in an attempt to assess the progress that has been made in the use and understanding of these models.

A more ambitious project would be titled "Latent Variable Analysis at 100", since Charles Spearman's 1904 paper in the *American Journal of Psychology,* "'General Intelligence' Objectively Determined and Measured", is usually considered to be the starting point of the factor analysis movement. Whatever name is used for our subject (factors, latent variables, linear structural equations, etc.), there are a couple of critical aspects that are held in common by all contributors to the literature.

The first is that things are not always what they seem. The observations that we record are linked to, but do not exactly represent, the phenomena which we are seeking to measure and through measurement to understand. The second aspect is that the links between what Lazarsfeld called latent and manifest variables can be represented by formal mathematical models.

Aside: Lazarsfeld was, of course, Viennese, which probably accounts for his affection for the words "latent" and "manifest"! His colleague at Columbia, sociologist Robert K. Merton, adapted the terms to the study of the functions of social norms.

Manifest variables, or observable variables, are the basic measurements of an empirical study. The assumption that latent, or unobservable variables, exist, and have specific types of relationships to manifest variables, allows the empirical researcher to transcend his or her data and to speak as a theoretician, not merely as a statistician. Gigerenzer, et al (*The Empire of Chance: How probability changed science and everyday life", 1989)* point out that the theoretical concept of intelligence which Spearman proposed was, in fact, "a statistical construct defined by the method of factor analysis" (p. 244). L. L. Thurstone extended Spearman's research program to include the definition and scientific measurement of a variety of mental abilities (*The Vectors of the Mind*), and the social psychologists brought together to study the World War II soldiers adapted these methods to the study of attitudes.

Lazarsfeld's two chapters in *Measurement and Prediction*, were preceded by eight chapters by Louis Guttman that introduced "Guttman Scaling" to a wide audience. What distinguished their work from the psychometric studies of earlier decades was that for Guttman and Lazarsfeld the empirical data, the manifest variables, were items that consisted of the discrete categorical responses to questions rather than test scores. Lazarsfeld, in fact, concentrated for the most part on the simplest such items, those with only two possible responses, Yes/No or, generically, plus/minus, positive/negative.

AN EXAMPLE

Table 1 (Table 2 from Lazarsfeld, 1950b) shows a typical dataset for a latent structure analysis. Four dichotomous items provide an empirical breakdown of 1000 respondent-soldiers into 16 categories. The responses have been coded so that the + symbol indicates a positive feeling about the Army, - a negative feeling.

TABLE 1

Manifest Data of Four Items on Attitude toward the Army

In general how do you feel the Army is run?	Do you think when you are discharged you will [have] a favorable attitude toward the Army?	In general do you feel you yourself have gotten a square deal from the Army?	Do you feel that the Army is trying its best to look out for the welfare of enlisted men?	Count
+	+	+	+	75
+	+	+	-	69
+	+	-	+	55
+	-	+	+	42
-	+	+	+	3
+	+	-	-	96
+	-	+	-	60
+	-	-	+	45
-	+	+	-	16
-	+	-	+	8
-	-	+	+	10
+	-	-	-	199
-	+	-	-	52
-	-	+	-	25
-	-	-	+	16
-	-	-	-	229

TABLE 2

Computed Latent Structure for Attitude toward the Army

Latent Class Frequencies	Item 1 Probability +	Item 2 Probability +	Item 3 Probability +	Item 4 Probability +
424.3	.9240	.6276	.5704	.5125
575.7	.4324	.1871	.1008	.0635

A straightforward descriptive analysis of these data shows that negative responses are more numerous except on item 1; and that there is a positive association between each pair of items. A soldier who responds positively to any one item is more likely to respond positively to a second item. Lazarsfeld's analysis is based on the ***assumption*** that each soldier can be thought of as belonging to one of two latent classes. The probability of positive response to an item will generally be different in one class than in the other. Most importantly, he ***assumes*** for an individual respondent that the responses to items are statistically independent. That's the essence of a latent class model.

From an interpretive point of view, note the implications. The items are correlated not because there is some causal connection linking one to another, but because the population is heterogeneous. If only one class of soldiers had been interviewed no correlations would be observed. Lazarsfeld coined the term ***local independence*** to describe this condition.

In the two chapters in *Measurement and Prediction* Lazarsfeld showed how some latent structure models might be defined, and then showed how the parameters of those models could be estimated from the manifest data. The latent structure of this example is summarized in Table 2 (Table 6 of Lazarsfeld, 1950b).

The results tell us that the population is divided roughly 40%/60% between those who are generally favorable to the army and those who are generally negative. Almost everyone (92%) who belongs to the former class will answer positively to the question "In general how do you feel the Army is run", while almost everyone (94%) in the latter class will respond negatively to the fourth item, "The Army is trying its best to look out for the welfare of enlisted men."

It is possible to calculate, for each of the 16 manifest response patterns in Table 1, the probability that the responses came from a member of Class 1. Lazarsfeld proposed that this "posterior probability" be used as a numerical scale, a way of ordering the 16 response patterns, and ultimately as a characteristic of the respondent himself in subsequent analyses.

Of course this interpretation of the numbers in Table 2 is predicated on the assumption that the mathematical model accurately describes the behavior of soldiers answering questions about their feelings.

LATENT CLASSES OR LATENT TRAITS?

I have already noted that the use of discrete (categorical) manifest variables distinguished latent structure analysis (and Guttman scaling) from factor analysis. A related issue has to do with the nature of the latent structure, the nature of the latent variable (or variables). Intelligence was, for Spearman, a continuous numerically valued phenomenon, and so were almost all of the psychometrically defined concepts associated with factor analysis modelling. The goal of the attitude researchers in the 1940s was to measure attitudes, and in most cases attitudes were also conceptualized as numerical variables. Research characterized as "scale analysis" generally took this premise for granted, as the work of Guttman and Clyde Coombs (*A Theory of Data,* 1964) shows.

The example above takes a different position, by postulating the existence of a relatively small number of homogeneous subpopulations or latent classes. An individual is characterized by membership in a class, and by the response probabilities associated with such membership. Suppose that we assume that the model really is stochastic at the level of the individual, rather than descriptive of aggregate proportions. In other words, imagine that the responses of an individual are triggered by the equivalent of a series of tosses of some weighted coins. Although each response pattern has a probability of having been generated by a member of latent class 1, and these numbers could be used to order the response patterns, these probabilities are not intrinsic properties of the individual. They are simply expressions of the uncertainty the researcher feels about the proper assignment of the respondent to a latent class.

The term "latent trait" has long been used for latent structure models with continuous latent variables. Arguments over whether a particular concept is discrete or continuous, identifying a trait or a class membership, and whether it is even possible to empirically answer such a question, are part of the fifty year history of latent structure analysis.

Two of the most recent books on latent structure use the terms "class" and "trait" in their titles (Heinen, 1996; Langeheine and Rost, 1988). Heinen, and Jan DeLeeuw in his editor's introduction, point to the origin of both models as an explanation for why they are considered as different traditions in modelling categorical data. "Latent class analysis was developed mainly within the social and political sciences, whereas latent trait models have a clear psychometrical background." (Heinen, 1996, ix) Curiously, they have not perceived how the elaboration of both types of models exists simultaneously in Lazarsfeld's work in the 1940s, '50s and '60s.

American sociologists have been fascinated from the very beginning with the idea of social class. Whether social status is enabled by membership in a

class of persons or should be conceptualized as a continuous variable has been hotly, warmly and coolly debated throughout this century. Lazarsfeld's latent class analysis was taken up by the class proponents as a method that would allow the scientific assessment of a person's class membership. It is, however, important to remember that the survey research environment of the mid-century was more psychological than sociological, and that Lazarsfeld had not achieved the status within sociology that he later enjoyed.

"The logical and mathematical foundation of latent structure analysis", chapter 10 of *Measurement and Prediction,* begins by discussing a latent trait model. The first chart in the chapter explains what Lazarsfeld called a *traceline,* a function that indicates how the probability of positive response to an item changes as the "ethnocentrism" of American soldiers varies along a continuum. Only later does he bring up the special case where the latent distribution "is assumed to be concentrated at a different point on the continuum." (p. 376) The existence of a continuum is taken for granted, here and elsewhere throughout these two chapters, and I believe this caused a great deal of confusion in later years. At the time, however, it made the topic more understandable and acceptable to the social psychologists who were the primary audience of these research studies, the heirs if not in fact the students of Thurstone and Spearman.

Lazarsfeld's ambivalent position between the disciplines of sociology and psychology is shown by his next major publication on the subject. It is a chapter in the multiple-volume set sponsored by the American Psychological Association, *Psychology: A Study of a Science* (Koch, 1959). Once again the latent structure approach to the analysis of dichotomous manifest data is explained in terms of traceline functions and a latent continuum (possible multidimensional!). When it comes time to explain what local independence his, however, he falls back on an example of three latent classes. It is, after all, easier to demonstrate how a mixture of 2x2 frequency tables, each satisfying the property of statistical independence, add together to form a table in which there is an association, than to demonstrate the analog for continuous variables X and Y.

Linear, and more generally, polynomial, tracelines have the embarrassing property of eventually exceeding 1.0 and/or falling below zero. Well before 1959 others who were busy modelling discrete phenomena had chosen to get around this problem by choosing a functional form that was constrained to fall between zero and 1. D. J. Finney had used a normal ogive (cumulative distribution function) in his *probit analysis* model for an animal's response to dosages of a poison, as had Fred Lord and Ledyard Tucker in their elaboration of models for ability tests. Logit models, in which the log of the odds of a response instead of the probability of the response had a linear relationship to the latent continuum, had also been developed.

Lazarsfeld chose to study and analyze the properties of polynomial traceline models because of their intimate connection to latent class models: *two points determine a line.* If a traceline is a polynomial of degree m, then m+1 points on the curve will tell us what the polynomial is. A two-class model, he proved, was indistinguishable from a linear traceline model; a three-class model was indistinguishable from a quadratic traceline model, and so on, as long as the only data available are the manifest item responses. He could present a latent class analysis using words that made sense to an audience of "factor" psychologists by translating the class parameters into trait terminology, by "locating" the classes at discrete points along "the latent continuum"

THE ERA OF THE MATHEMATICAL STATISTICIAN

Mathematical statistics (like attitude research) came into its own as a distinct discipline in the U.S. only after the Second World War. It is instructive that in the 1950 and 1959 chapters the idea of evaluating the fit of a model to the manifest data is almost invisible. ("Professor Frederick Mosteller has suggested that the goodness of fit could be tested by using chi-square with 2^m degrees of freedom reduced by the number of parameters", Lazarsfeld 1950b: p.429). What is now called parameter estimation was formulated as a problem of decomposing a manifest distribution into its homogeneous components. The fact that with real data this program can never be exactly achieved was not ignored, but it was not dealt with systematically. Familiar (to statisticians) concepts like standard errors, unbiased estimation, and maximum likelihood do not appear in the 1950 or 1959 work.

Some of these issues were being addressed in journal articles by statisticians, notably by T.W. Anderson (1954, 1959). When the textbook *Latent Structure Analysis* was published in 1968, it used what was by then standard statistical terminology for the formulation, estimation and testing of stochastic models. It retained, however, Lazarsfeld's concern with the identification of model parameters (identifiability) and the primacy of the method of moments as the tool for deriving parameter estimates.

LOG-LINEAR MODELLING AND LATENT STRUCTURE ANALYSIS

Lazarsfeld considered that the statistical analysis of discrete categorical data had been ignored by the statisticians of his generation, and in a series of historical papers he drew attention to the 19th century statistician August Quetelet and the British statistician G. U. Yule. By 1975, however, a new approach to categorical data analysis had become current. Leo Goodman, in a series of papers that were reprinted in 1978, had made the terms "log-linear model" known among sociologists as well as among mathematical statisticians. In a long paper in 1974 in the *American Journal of Sociology* (Chapter 8 in Goodman, 1978), he showed how the latent class model could be interpreted as a log-linear model.

This article, and the publication in 1975 of *Discrete Multivariate Analysis* by Bishop, Fienberg and Holland, brought latent class models to the attention of mainstream statisticians. By thinking of the latent class variable as having "missing values" in a cross-classification, a wider range of interpretations and applications became available. Algorithms for parameter estimation in log-linear models could be adapted to deal with many of the classic latent structure models. Shelby Haberman's books (*Analysis of Qualitative Data,* 2 Volumes, 1978-79) and program **LAT** for estimation of latent class parameters were also important, though perhaps less influential in the short term.

REVISITING TRAIT/CLASS CONTROVERSY

An important issue in the literature has been the way in which special cases of general latent structure models are specified. "New" models have been proposed from time to time which turn out to be "old" models in which certain constraints have been placed on the parameters. Lazarsfeld's *latent distance model*, for instance, is equivalent to a latent class model in which some of the latent class probabilities are forced to be equal. Lazarsfeld had a great affection for this model, which explicitly combines characteristics of the trait and class approaches to latent variables. In these models the latent continuum is assumed to exist, but it has a discontinuous step function relationship with the probability of response to any particular item. In Lazarsfeld's writings the model is described as a generalization of Guttman's famous scale, with less restrictive but intuitively attractive assumptions about the items. The most interesting variations on this theme can be found in articles by Cliff Clogg and his students and colleagues (e.g. Clogg, 1988).

Rolf Langeheine provided a very nice review of the different approaches to constraining parameters in his chapter "New Developments in Latent Class Theory" in Langeheine and Rost (1988). Several other chapters in that collection also address the issue: John Bergan ("Latent Variable Techniques for Measuring Development") and Erling Andersen ("Comparison of Latent Structure Models") show how class and trait models can be used and interpreted in a complementary manner. While most latent trait models assume a one-dimensional latent variable, a multidimensional interpretation is often more consistent with a latent class approach.

Michael Sobel has criticized the traditional interpretation of the latent variable as the true cause of the manifest responses (Sobel, 1997), based on the assumption of local independence. This criticism, of course, applies to both the class and trait models.

WHERE ARE THE APPLICATIONS?

It had been my impression that latent structure analysis survived mainly in the writings of statisticians and methodologists who had been able to find some interesting special models to derive theorems about. A surprisingly large number of applications of latent class analysis have appeared since 1990, however. The diversity of sources is remarkable, but understandable, given the variety of disciplines that Lazarsfeld himself influenced directly: psychiatry, medicine, marketing, and public opinion are all represented in titles found by a search of the Social Science Citation Index database. The most recent issue of ***JASA*** features an application to criminology (Roeder *et al.*, 1999). I plan to comment on some of these applications in a later paper.

REFERENCES

Anderson, T. W. (1954) "On Estimation of Parameters in Latent Structure Analysis", *Psychometrika,* 19: 1-10.

Anderson, T. W. (1959) "Some Scaling Methods and Estimation Procedures in the Latent Class Model", in *Probability and Statistics,* U. Grenander (ed.). New York: John Wiley & Sons.

Bishop, Y.M.M., S. E. Fienberg, and P.W. Holland (1975) *Discrete Multivariate Analysis,* Cambridge MA: MIT Press.

Clogg, Clifford C. (1988) "Latent Class Models for Measuring", in Langeheine and Rost, 1988.

Coombs, Clyde H. (1964) *A Theory of Data,* New York: John Wiley & Sons

Gigerenzer, Gerd et *al.* (1989) *The Empire of Chance,* Cambridge: Cambridge University Press

Goodman, Leo A. (1978) *Analyzing Qualitative / Categorical Data: Log-linear Models and Latent Structure Analysis.* Cambridge, MA: Abt Books.

Haberman, Shelby J. (1978, 1979) *Analysis of Qualitative Data, Vol. 1&2,* New York: Academic Press.

Heinen, Tod (1996) *Latent Class and Discrete Latent Trait Models,* Thousand Oaks, CA: SAGE.

Langeheine, Rolf and Jurgen Rost (1988) *Latent Trait and Latent Class Models,* New York: Plenum Press.

Lazarsfeld, Paul F. (1950a) "The Logical and Mathematical Foundations of Latent Structure Analysis", Chapter 10 in Stouffer (1950).

Lazarsfeld, Paul F. (1950b) "Some Latent Structures", Chapter 11 in Stouffer (1950).

Lazarsfeld, Paul F. (1959) "Latent Structure Analysis", in *Psychology: A Study of a Science, Vol. 3,* S. Koch (ed.). New York: McGraw-Hill.

Lazarsfeld, Paul F. and Neil W. Henry (1968) *Latent Structure Analysis,* Boston: Houghton Mifflin.

Roeder, Kathryn, K. G. Lynch and D. S. Nagin (1999) "Modeling Uncertainty in Latent Class Membership: A Case Study in Criminology", *JASA, 94: 766-776.*

Sobel, Michael E. (1997) "Measurement, Causation and Local Independence", in *Latent Variable Modeling and Applications to Causality,* M. Berkane (ed.). New York: Springer-Verlag.

Spearman, Charles (1904) "'General Intelligence', Objectively Determined and Measured", *American Journal of Psychology,* 15: 201-293.

Stouffer, Samuel A., *et al. (1950) Measurement and Prediction,* Volume IV of *The American Soldier: Studies in Social Psychology in World War II..* Princeton University Press. Reprinted 1973 by Peter Smith, Gloucester MA