

COWLES COMMISSION DISCUSSION PAPER: STATISTICS No. 360

NOTE: Cowles Commission Discussion Papers are preliminary materials circulated privately to stimulate private discussion and are not ready for critical comment or appraisal in publications. References in publications to Discussion Papers (other than mere acknowledgment by a writer that he has had access to such unpublished material) should be cleared with the author to protect the tentative character of these papers.

Identification Problems in Latent Structure Analysis

by T. C. Koopmans

May 11, 1951

Ref: P. Lazarsfeld, "The Logical and Mathematical Foundation of Latent Structure Analysis," Chapter 10 of Measurement and Prediction, Vol. IV of Studies in Social Psychology in World War II, Princeton 1950.

The following note presupposes familiarity with Lazarsfeld's ideas such as can be obtained from a reading of this reference, and with the terminology and concept of identifiability as illustrated in C. C. New Series Paper No. 39. It is highly probable that the propositions stated here, and others going beyond this, have already been developed by Lazarsfeld and others. They have been put together here in the spirit of Reiersøl's article on identification problems in factor analysis (reprinted in New Series Paper 39), and also as an aid in a course I am teaching in statistical problems of model construction.

Consider the following simple latent structure model.

$P_{ij\dots k}$  = probability that a person selected at random from the population will answer "yes" to each of the attitude test questions numbered  $i, j, \dots, k$ .

$P_{ij\dots k}^g$  = same probability for a person selected at random from the  $g$ -th latent group in the population.

$\sqrt{g}$  = percentage of population in  $g$ -th latent group.

The quantities  $p_{ij\dots k}$  and  $p_{ij\dots k}^g$  are independent of the ordering of their subscripts.

The fundamental hypothesis of latent structure analysis is

$$(1) \quad p_{ij\dots k}^g = p_i^g \cdot p_j^g \cdot \dots \cdot p_k^g$$

From the identity

$$(2) \quad p_{ij\dots k} = \sum_{g=1}^{\lambda} \nu^g p_{ij\dots k}^g \quad (\lambda = \text{the number of latent groups})$$

follows by (1) the system of so-called accounting equations

$$(3) \quad p_{ij\dots k} = \sum_{g=1}^{\lambda} \nu^g p_i^g \cdot p_j^g \cdot \dots \cdot p_k^g \quad i, j, \dots, k \text{ all different each running from 1 to } m \text{ (the number of test items)}$$

The left hand members represent parameters of an observable multinomial distribution. The quantities  $\nu^g, p_i^g$  represent structural parameters of which the identifiability is to be investigated. Following Lazarsfeld we define the matrices

$$(4) \quad N = \begin{bmatrix} \nu^1 & 0 & \dots & 0 \\ 0 & \nu^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \nu^\lambda \end{bmatrix}, \quad \text{with trace } N = 1,$$

$$(5) \quad L = \begin{bmatrix} 1 & p_1^1 & p_2^1 & \dots & p_m^1 \\ 1 & p_1^2 & p_2^2 & \dots & p_m^2 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & p_1^\lambda & p_2^\lambda & \dots & p_m^\lambda \end{bmatrix},$$

$$(6) \quad M = \begin{bmatrix} 1 & p_1 & p_2 & \dots & p_m \\ p_1 & p_{11} & p_{12} & \dots & p_{1m} \\ p_2 & p_{21} & p_{22} & \dots & p_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ p_m & p_{m1} & p_{m2} & \dots & p_{mm} \end{bmatrix},$$

where the diagonal entries of M except the first are as yet undefined. The top row and left column will be referred to as the zero-th row and column, their elements denoted alternatively by

$$(7) \quad p_{00} = 1, \quad p_{i0} = p_{0i} = p_i, \quad i = 1, \dots, m.$$

With these definitions

$$(8) \quad M = L' N L$$

represents those of the equations (5) with one and two subscripts respectively, plus definitions for the yet undefined diagonal elements of M.

Let  $(\sigma)$  represent a set of  $m_{(\sigma)}$  integers selected from the set  $i=1,2,\dots,m$ . Let  $(\sigma)$  be called a signature, and let  $\bar{\phi}_{(\sigma)}$  be a column-omitting matrix such that

$$(9) \quad L \bar{\phi}_{(\sigma)} = L_{(\sigma)}$$

is obtained from L by deleting the columns with order numbers in  $(\sigma)$ .

Define with Lazarsfeld

$$(10) \quad N_{(\sigma)} = \begin{bmatrix} \nu^{1,1}_{k(\sigma)} & 0 & \dots & 0 \\ 0 & \nu^{2,2}_{k(\sigma)} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \nu^{\lambda,\lambda}_{k(\sigma)} \end{bmatrix},$$

$$(11) \quad \nu^{k(\sigma)} = \prod_{i \in (\sigma)} p_i^g, \quad (\prod \text{ product symbol}),$$

$$(12) \quad M_{(\sigma)} = \bar{\phi}'_{i(\sigma)} \begin{bmatrix} p_{ij\dots k} & p_{ij\dots k1} & p_{ij\dots k2} & \dots & p_{ij\dots km} \\ p_{ij\dots k1} & p_{ij\dots k11} & p_{ij\dots k12} & \dots & p_{ij\dots k1m} \\ \dots & \dots & \dots & \dots & \dots \\ p_{ij\dots km} & p_{ij\dots km1} & p_{ij\dots km2} & \dots & p_{ij\dots kmm} \end{bmatrix} \bar{\phi}_{(\sigma)},$$

where  $i,j,\dots,k$  represents the integers included in  $(\sigma)$ . The matrices  $\bar{\phi}'_{i(\sigma)}$ ,  $\bar{\phi}_{(\sigma)}$  have the effect of removing from  $M_{(\sigma)}$  all elements involving a repetition of subscripts, except the as yet undefined diagonal elements.

The equations

$$(13) \quad M_{(\sigma)} = L'_{(\sigma)} N_{(\sigma)} L_{(\sigma)}$$

for all possible signatures  $(\sigma)$  redundantly express all remaining equations (5), and define the remaining undefined diagonal elements for all  $M_{(\sigma)}$ . The nondiagonal elements of the matrices  $M_{(\sigma)}$  (including  $M$  where  $(\sigma)$  is empty) are in principle observable. The matrices  $L$  and  $N$  are structural parameter matrices whose identifiability is now to be studied. The elements of  $\Phi_{(\sigma)}$  are known constants (0 or 1).

Suppose there are two observationally equivalent structures  $(L, N)$  and  $(L^*, N^*)$  where the matrices  $L, N, L^*, N^*$  all have their maximum ranks, which equals  $\lambda$  in all cases.

$$(14) \quad \rho(L) = \rho(N) = \lambda, \rho(L^*) = \rho(N^*) = \lambda,$$

and where all elements of  $L$  and  $L^*$  are positive

$$(15) \quad p_1^g > 0, p_1^{g^*} > 0.$$

Then we have for all  $(\sigma)$

$$(16) \quad M_{(\sigma)} = L'_{(\sigma)} N_{(\sigma)} L_{(\sigma)} = L'^{*}_{(\sigma)} N^*_{(\sigma)} L^*_{(\sigma)} + D_{(\sigma)}, \quad (D_{(\sigma)})_{00} = 0,$$

where  $D$  is diagonal, since the diagonal elements of  $M_{(\sigma)}$  (except  $p_{(\sigma)00}$ ) do not represent parameters accessible to observation.

We have from the definition of  $N_{(\sigma)}$  and (14) and (15),

$$(17) \quad \rho(N_{(\sigma)}) = \lambda, \rho(N^*_{(\sigma)}) = \lambda.$$

Consider the subset  $S_\lambda$  of signatures  $(\sigma)$  such that

$$(18) \quad \rho(L_{(\sigma)}) = \lambda, \rho(L^*_{(\sigma)}) = \lambda.$$

By (14)  $S_\lambda$  contains the empty signature. We define

$$(19) \quad M^*_{(\sigma)} = L'^{*}_{(\sigma)} N^*_{(\sigma)} L^*_{(\sigma)} = M_{(\sigma)} - D_{(\sigma)}.$$

Then, from (16), (17), (18) and (19),

$$(20) \quad \rho(M_{(\sigma)}) = \lambda, \rho(M^*_{(\sigma)}) = \lambda \text{ if } (\sigma) \in S_\lambda.$$

Consider first the empty signature, for which

$$(21) \quad L'NL = L'^*N^*L^* + D, \quad d_{00} = 0, \quad D \text{ diagonal.}$$

We shall specify that, for every test question, numbered  $i_0$ , say, the set of columns of  $L$  (the set of tests plus the first column  $e$  consisting of elements 1 only) possesses three subsets I, II, III, of  $m_I, m_{II}, m_{III}$  columns respectively, such that each column of  $L$  is present in one subset, and no column is present in more than one subset except that  $e$  is allowed to occur in both I and II, and such that, for any structures  $L, L^*, \dots$  in the model,

$$(22) \quad \begin{aligned} (22a) \quad & i_0 \in \text{III}, \\ (22b) \quad & m_I = m_{II} = \lambda, \\ (22c) \quad & f(L_I) = f(L_{II}) = f(L_I^*) = f(L_{II}^*) = \lambda. \end{aligned}$$

This is a restriction on the set of structures (the model) admitted a priori for consideration.

Then we have from (21)

$$(23) \quad M_{I \ II} = L_I' N L_{II} = M_{I \ II}^* = L_I'^* N^* L_{II}^*$$

because the only diagonal element of (21) included in the submatrix (23) is the element  $(0, 0)$  for which  $d_{00}$  vanishes. Moreover, because of (14) and (22c) the matrix (23) is nonsingular.

Let  $\bar{I}, \bar{II}$  be the two sets of  $\lambda+1$  columns each, obtained by adjoining the column  $i_0$  to I and II, respectively. Then, by (20),  $M_{\bar{I} \ \bar{II}}$  and  $M_{\bar{I} \ \bar{II}}^*$  are both singular, and differ in no elements except possibly the diagonal elements  $m_{i_0 i_0}$  and  $m_{i_0 i_0}^*$ , of which the cofactors are the nonsingular matrix (23). It follows that  $m_{i_0 i_0} = m_{i_0 i_0}^*$  and, since a partitioning I, II, III satisfying (22) was assumed to exist for each  $i_0$ , that

$$(24) \quad L'NL = L'^* N^* L^*$$

Define diagonal matrices  $K$  and  $K^*$ , with nonnegative elements, through

$$(25) \quad N = K^2 = K'K, \quad N^* = K^{1*}K^*, \quad k_{11} > 0, \quad k_{11}^* > 0.$$

Then we have from (24)

$$(26) \quad TKL = K^*L^*, \quad TT' = I$$

(where  $T$  is an orthogonal matrix).

Conversely, if  $(L, N)$  is one structure satisfying the conditions (14a) and (15a), and  $T$  an orthogonal matrix such that (26) has a solution  $K^*, L^*$  satisfying (15b) and  $\text{tr}N^* = 1$ , then  $(L^*, N^*)$  is a structure satisfying (14b), which is indistinguishable from  $(L, N)$  on the basis of the observations  $M$ . With the help of a definition, we can now formulate a lemma that has been proved.

Definition: Two structures  $(L, N)$  and  $(L^*, N^*)$  are called equivalent up to second order joint occurrences if they imply equality of corresponding nondiagonal elements of  $M$  and  $M^*$ .

Lemma: Necessary and sufficient for the equivalence up to second order joint occurrences of two structures  $(L, N)$  and  $(L^*, N^*)$  satisfying (22) for each  $i_0$  is that there exists an orthogonal matrix  $T$  satisfying (26).

In order to obtain the set of structures  $(L^*, N^*)$  equivalent up to second order to a given structure  $(L, N)$  we must allow in (26) the set of all orthogonal matrices  $T$  which permit  $(L^*, N^*)$  to satisfy the obvious restrictions (satisfied also by  $L, N$ )

$$(27) \quad \text{tr } N^* = 1, \quad (L^*)_{i0} = 1, \quad i = 1, \dots, \lambda$$

$$(28) \quad (L^*)_{ij} = 0, \quad j=1, \dots, m, \quad i=1, \dots, \lambda.$$

No restriction on T arises at all from (27). For any T we obtain

$$(29) \quad k_{ii}^* = k_{ii}^* (L^*)_{i0} = (K^* L^*)_{i0} = (TKL)_{i0} = \sum_{h=1}^{\lambda} t_{ih} k_{hh} = \sum_{h=1}^{\lambda} t_{ih} k_{hh}.$$

Treating

$$(30) \quad k_h = k_{hh} \quad k_i^* = k_{ii}^*$$

for a moment as the elements of column vectors k, k\* we have from (29)

$$(31) \quad k^* = T k$$

and hence

$$(32) \quad \text{tr } N^* = k'^* k^* = k' T' T k = k' k = \text{tr } N = 1$$

The restriction (28) of course does limit the choice of matrices T that lead to equivalent structures.

For later use we will now give an answer to the following question: Given L and L\*, to what extent are T, K and K\* thereby determined. If  $\bar{T}$ ,  $\bar{K}$ ,  $\bar{K}^*$  also satisfy

(26) we have from (14)

$$(33) \quad \bar{K}^{-1} \bar{T} \bar{K} = K^{-1} T K,$$

or

$$(34) \quad \bar{T} = \bar{K}^* \bar{K}^{-1} T K \bar{K}^{-1} = D^{-1} T D, \quad \text{say } (D \text{ and } D^* \text{ diagonal}).$$

Since  $\bar{T}$  is by assumption also orthogonal,

$$(35) \quad \bar{T} \bar{T}' = D^{-1} T D D' T' D^{-1} = I$$

or, writing

$$(36) \quad D^2 = D D' = E, \quad D^{\bar{2}} = D'^* D^* = E^*,$$

we have

$$(37) \quad T E T' = E^*$$

or

$$(38) \quad T E = E^* T, \quad E \text{ and } E^* \text{ diagonal.}$$

By permutation of rows and columns, it is always possible to bring T in the form

$$(39) \quad T = \begin{bmatrix} T_1 & 0 & \dots & 0 \\ 0 & T_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & T_r \end{bmatrix}, \quad 1 \leq r \leq \lambda$$

with no  $T_q$  resolvable into smaller diagonal blocks by further permutation. We now write (38) as

$$(40) \quad t_{ij} e_j = e_i^* t_{ij}$$

It follows from the orthogonality and nonresolvability of each  $T_q$  that any  $T_q$  of order  $> 1$  has at least one nondiagonal nonzero element in each row and in each column, and that, for all  $i, j$  such that  $t_{ij}$  is an element of  $T_q$ ,

$$(41) \quad e_i^* = e_j = d_q^2, \text{ say.}$$

It follows further that

$$(42) \quad E = E^* = D^2, \quad TE = E^*T = \begin{bmatrix} d_1^2 T_1 & 0 & \dots & 0 \\ 0 & d_2^2 T_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & d_r^2 T_r \end{bmatrix}$$

Tracing back the definitions of  $E, E^*$  in terms of  $K, K^*, \bar{K}, \bar{K}^*$ , we obtain

$$(43) \quad D = K \bar{K}^{-1} = D^* = K^* \bar{K}^{*-1},$$

so that there exists a diagonal matrix  $D$  such that

$$(44) \quad K = D \bar{K}, \quad K^* = D \bar{K}^*$$

and of which all diagonal elements corresponding to the same  $T_q$  are equal in value to  $d_q$ . Moreover, from (34) and (39)

$$(45) \quad \bar{T} = D^* T D = T.$$

We conclude that (26) with given  $L$  and  $L^*$  determines  $T$  uniquely, and  $K, K^*$  up to the transformation (44). The choice of the positive numbers  $d_q$  is not completely free. It is restrained by the condition on  $\text{tr } N$ , which now takes the form

$$(46) \quad \text{tr } N = \text{tr } K^2 = \text{tr } D^2 \bar{K}^2 = \sum_q d_q^2 \text{tr } \bar{K}_q^2 = 1$$

where it is given that

$$(47) \quad \text{tr } \bar{N} = \text{tr } \bar{K}^2 = \sum_q \text{tr } \bar{K}_q^2 = 1.$$

As a result, we notice that, if  $r=1$ , hence  $T$  not resolvable, hence  $d_q = d$ , we must have  $d = 1$ , and  $K, K^*$  and therefore also  $N, N^*$  are fully determined by  $L$  and  $L^*$ .

At the other extreme, if  $r = \lambda$ , hence  $T$  itself diagonal, then (26) implies

$$(48) \quad L^* = FL, \quad F \text{ diagonal,}$$

and the fact that the first rows of  $L$  and  $L^*$  consist of elements 1 only requires

$$(49) \quad F = I$$

so that this case occurs if and only if

$$(50) \quad L^* = L.$$

Because of the sign restrictions (25) on  $k_{ii}$  and  $k_{ii}^*$  and the orthogonality of  $T$  we must now also have

$$(51) \quad T = I, \quad K = K^*, \quad N = N^*$$

and the two structures are identical.

We are now ready to derive a sufficient criterion for identifiability of a structure  $(L, N)$ . It has already become clear that in order to identify a structure  $(L, N)$ , use must be made of information regarding higher order joint occurrences.

Assume now that the set  $\bar{S}_\lambda$  of signatures satisfying both the requirement (18) -

given (14) - and the requirement that, for all structures  $L, L^*, \dots$ , in the model, and for all  $i_0 \in \{ \sigma \}$ ,

$$(52) \quad (22) \text{ with } L_{(\sigma)} \text{ taking the place of } L,$$

contains more than the empty signature. Then, in analogy to (24),

$$(53) \quad M_{(\sigma)}^* = L_{(\sigma)}'^* N_{(\sigma)}^* L_{(\sigma)}^* = L_{(\sigma)}' N_{(\sigma)} L_{(\sigma)} = M_{(\sigma)}.$$

We define diagonal matrices with nonnegative elements

$$(54) \quad P_{(i)} = \begin{bmatrix} p_1^1 & 0 & \dots & 0 \\ 0^1 & p^2 & \dots & 0 \\ \dots & \dots & & \\ 0 & 0 & \dots & p_1^\lambda \end{bmatrix}, \quad P_{(\sigma)} = \prod_{i \in \{ \sigma \}} P_{(i)} = Q_{(\sigma)}^2, \quad K_{(\sigma)} = Q_{(\sigma)} K,$$

and similarly for starred parameters. Then, from (10) and (25),

$$(55) \quad N_{(\sigma)} = K'_{(\sigma)} K_{(\sigma)}, \quad N^*_{(\sigma)} = K'^*_{(\sigma)} K^*_{(\sigma)}$$

and, in analogy with (26), there is a matrix  $T_{(\sigma)}$  such that

$$(56) \quad T_{(\sigma)} K_{(\sigma)} L_{(\sigma)} = K^*_{(\sigma)} L^*_{(\sigma)}$$

and

$$(57) \quad T'_{(\sigma)} T_{(\sigma)} = I.$$

Because of (9), (14), (18), (26) and (56),

$$(58) \quad K^{*-1} TK = K'^{-1} T_{(\sigma)} K_{(\sigma)},$$

an equation of the same type as (33), of which we have already found the implications in (44) and (45). Hence

$$(59) \quad T_{(\sigma)} = T$$

and

$$(60) \quad K_{(\sigma)} = D_{(\sigma)} K, \quad K^*_{(\sigma)} = D_{(\sigma)} K^*, \quad D_{(\sigma)} \text{ diagonal.}$$

Comparing (60) with (54c) we have

$$(61) \quad D_{(\sigma)} = Q_{(\sigma)} = Q^*_{(\sigma)} \text{ for all } (\sigma) \in S_{\lambda}.$$

Reference to the definition (54) of  $Q_{(\sigma)}$  leads to

**Theorem 1:** A sufficient condition for the identifiability of the  $i$ -th column of  $L$  in the model  $M$  defined by the specifications (4), (5),

(14), (15), (22) is that (52) holds for some signature  $(\sigma)$  containing  $i$ .

For the proof, we note that if (52) holds for a non-empty signature  $(\sigma)$ , it holds also for all subsignatures  $(\bar{\sigma})$  of  $(\sigma)$ , since the column in  $(\sigma)$  but no in  $(\bar{\sigma})$  can be added to group III. Hence, from (61) for  $(\sigma)$  and for  $(\bar{\sigma}) \equiv (\sigma) - (i)$ , we conclude from (54b) that

$$(62) \quad P_{(i)} = P^*_{(i)}.$$

**Theorem 2:** A sufficient condition for the identifiability of  $N$  in the model  $M$  is that the condition of Theorem 1 holds for all columns of a submatrix  $\bar{L}$  of  $L$  of rank  $\lambda$ .

The proof follows from the observation that the reasoning leading from (50) to (51) remains valid if (50) is replaced by

$$(63) \quad \bar{L}^* = \bar{L}, \quad f(\bar{L}) = \lambda$$

It is noted that the proof of identifiability depends only on signatures containing just one test number. Hence, if the model is not subject to doubt and the sample large, the frequencies  $p_{ijk}$  of triple occurrences are sufficient to determine a latent structure. Of course, occurrences of higher multiplicity can be used to obtain better estimates from finite samples or to test specifications of the model.

The foregoing analysis leaves unanswered the question of necessary and sufficient criteria for identifiability of a column of L, and of N. It is believed that the conditions of Theorems 1 and 2 are not necessary. We have followed the easy way of setting the rank conditions on L in such a way that (24) is obtained by the comparison of corresponding submatrices of M and M\* containing only one unknown diagonal element. The analysis also needs to be extended to the case of a latent continuum of attitudes instead of a finite number of homogeneous latent classes.

Note added on May 31 after a conversation with Lazarsfeld: A hectographed memorandum "The Complete Solution of the Three Class Latent Structure" by Lazarsfeld contains a method of determining L and N from given manifest joint occurrence probabilities  $p_i, p_{ij}, p_{ijk}$  of the first three orders, which will work whenever the conditions of our theorems are satisfied. This method is based on the matrix

$$(64) \quad R_i(u) \equiv \Phi_{(i)}' M \Phi_{(i)} u - M_{(i)} = L_{(i)}' N (uI - P_{(i)}) L_{(i)}$$

where u is a scalar, and (i) the signature consisting of the i-th test item only. The matrix  $R_i(u)$  is a product of matrices of maximal rank equal

to  $\lambda$ ) and hence itself of rank  $\lambda$ , unless

$$(65) \quad u = p_1^E \quad \text{for some } g, \quad 1 \leq g \leq \lambda.$$

Hence, if  $\Psi_{(\sigma)}$  and  $\Psi_{(\tau)}$  be column-selecting (instead of column-omitting) matrices that select the columns corresponding to signatures  $(\sigma)$  and  $(\tau)$ , there must exist two signatures  $(\sigma)$ ,  $(\tau)$ , containing  $\lambda$  items each, such that

$$(66) \quad f \left( \Psi_{(\sigma)}' R_1(u) \Psi_{(\tau)} \right) = \lambda - 1 \text{ if } u \text{ satisfies (65)} \\ = \lambda \text{ for all other values of } u.$$

The method described by Lazarsfeld depends on  $(\sigma)$  and  $(\tau)$  having no items in common (except possibly the "items" corresponding to the zero-th columns of L).

Hence  $(\sigma)$  and  $(\tau)$  play the roles of I and II in our discussion, and, as conjectured on p. 1 above, our discussion has only produced results already known to the latent structure analysts. I am nevertheless circulating it in the hope that the method of approach may stimulate further work that would lead to the establishment of necessary and sufficient conditions for identifiability of the elements of L and N.