SUMMARY OF RESEARCH ON A NOVEL, INDIRECT FACTOR ANALYTIC SOLUTION[1]

Harry R. Barker & Barbara M. Barker, University of Alabama

This paper summarizes briefly some of the trials and tribulations (and a few successes) of approximately five years of research. Research efforts were aimed at achieving an inexpensive, fast and accurate computer method of factoring large data matrices. Despite the enormous advancement in computer technology and software, there are formidable obstacles to factoring large data sets.

As an example, Barker, Fowler and Peterson (1971) succeeded in factoring a large data matrix consisting of 373 items of the short form MMPI (Minnesota Multiphasic Personality Inventory). This feat required the use of a special computer system (Matlan, 1968) developed by IBM. Approximately 50 hours of computer run time over a period of approximately two years was required. An enormous amount of computer storage space was required. (An IBM 360/MOD 50 computer with 512 K storage was used for the study.)

Subsequent efforts using the Matlan System to factor the 511 items of the original MMPI proved unsuccessful - defects in the revised Matlan System and/or changes in the computer configuration were apparently responsible for the failure. These computer efforts were made possible in part, by support from NSF funding and could not have been attempted under typical computer facilities. Typical computer installations are restricted to about 80-100 variables which can be factored at one time.

Various unconventional solutions have been proposed to enable factoring of large data matrices. Most of these methods have obvious defects. However, one method seems particularly attractive when evaluated from the standpoint of accuracy, and of computer storage space and computer run time required. Horst (1965) presents the mathematical proof for the method; development of the method is presented in Appendix A.

Translated into words, the method requires that a subset of variables be identified. These variables can be independent items or some combination of items in the larger matrix. Implementation of the method by Mees (1959) involved grouping MMPI items into arbitrary clusters from which totals are derived. These totals are intercorrelated and then each total is correlated with the remaining items. (Note that the larger matrix, intercorrelation of items is thus bypassed). The intercorrelation matrix of totals is factor analyzed and then by appropriate matrix manipulation the factor load of each item is estimated. Since the method arrives indirectly at an estimate of the factor load of items (variables), the method will hereafter be termed an indirect approach to the direct (conventional) solution.

The purpose of the paper is to summarize and evaluate five independent attempts to implement Horst's indirect method to estimate the factor structure of data matrices.

Method and Results

Study 1. Stallings (1973) refined an earlier computer program developed by Barker. She applied the indirect method to data matrices which varied in size; (36, 50, 75, 100, and 117 variables). The number of variables to be factored was kept well within computer range to enable direct checks on the accuracy of results. As anticipated, the computer time required for solution was drastically reduced by the indirect method (e.g., only three minutes of computer time was required for the 117 variable problem). Accuracy of the first factor was impressive whereas accuracy of the subsequent factors was disappointing. As the number of variables increased, the accuracy of the indirect method also increased.

Study 2. Sloan (1973) further refined the program for the indirect solution. She artificially generated two data matrices which differed in terms of degree of correlation between items (variables). One data set contained low intercorrelation (average = .3, range = .17 to .43); the other data set contained moderately high intercorrelation (average = .5; range = .34 to .67. Principal axes factor estimates from the indirect solution were rotated to maximal configuration with the direct factor solution and then the direct and indirect solution were compared. Accuracy of the indirect solution was excellent for the first factor but diminished for succeeding factors. Again, computer run time was minimal for the indirect method. Accuracy of solution was greater for the moderately intercorrelated data set than for the low-intercorrelated data set.

Study 3. Barker and Barker (1975) applied the indirect solution to the data orginally factored by Barker, Fowler and Peterson (1971). Interest centered on accuracy of the indirect method as a function of number of totals employed in the solution. The 373 items were divided successively into sets of approximately equal size to enable totaling of answers in sets. The matrix of totals employed in solution were varied from nine through twenty successively. Indirect factor solutions were rotated to maximal alignment with the original conventional solution (principal axes) and item cosines (which indicated closeness of item location in factor space) were examined.

As the number of totals increased, the accuracy of the indirect solution increased in estimating the direct solution. However, from a practical standpoint, varimax rotation of the indirect solution with even 20 totals was not useful in identifying the item factor clusters.

Study 4. Barker and Barker (1974) turned attention away from number of totals used to the selection of items which comprised each of the totals. Previously, items for totals were simply selected in succession. The number of totals was held constant at nine since nine factors were identified in the original conventional solution. However, the method of selecting items for the totals was varied. For the control condition, the items were selected in succession to form nine totals. For comparison, the nine sets of items previously identified by the conventional solution (Barker, Fowler, and Peterson, 1971) were used to form totals.

Accuracy of the indirect solution for which items were selected in succession was relatively poor. Varimax rotation of the estimated principal axes resulted in poor identification of the original factor item clusters.

Use of the original direct solution to assign items to respective totals resulted in dramatic improvement in accuracy of the indirect solution. Table 1 shows that identification of the item factor clusters was virtually perfect for the indirect solution, except for Factor II. Factor II was the only factor which in the original solution contained a number of both negative and positive loads. This finding was viewed, at the time, as a limitation of the indirect method. A further, fatal limitation, seemed to be that one needed a priori the correct solution before he could arrive at the correct solution using the indirect method. That is, one needed to know the factor identification of each item in order to place it in the appropriate total. A rationale for identifying items with totals was clearly needed.

Insert Table I About Here

Study 5. Barker and Barker (1975) used the F scale of the MMPI consisting of 64 items in an effort to arrive at a rationale for identifying items to form totals. The F scale was selected because of its relatively small size (64 items) and because of the extensive prior work by comrey (1958).

Several Procedures were developed for testing:

(1) Subsets of items were identified by a conventional factor method. These subsets were used as a basis for identifying items to comprise totals. The indirect solution was applied using the identified totals. Results of the indirect method were used as a basis for reassigning items to totals and the indirect method was again applied.

(2) The indirect solution was applied to totals, the items of which were sequentially assigned. Results of the solution were used as a basis for reassignment of items to totals and the indirect solution was reapplied. As many as three iterations were attempted.

All results were essentially disappointing. No approach resulted in improved accuracy of identification of item clusters for totals. Note: these results were reported to Southeastern Psychological Association in March 28, 1975. Subsequently, we discovered that certain unintended methodological steps were taken which did not permit the results to be properly observed. It was also reported that the formation of totals, by assigning items from actual factor clusters, did not result in a good solution. This research has been replicated, with appropriate change in methodology, and the accuracy of the indirect factor estimates appear quite good.

Break Throughs

Two decided breakthroughs have occurred during the course of these studies. The first occurred with the discovery that an almost perfect solution could be obtained with the indirect method if one knew beforehand the factor structure of the variables and could assign items to totals accordingly. However, even this solution proved inadequate in the case of a bipolar factor (as in Study 4).

The second breakthrough occurred a month ago when it was discovered that items which comprise a total should be aligned in the sense of directionality. This can be accomplished by a simple scoring scheme. The computer program for the indirect solution was revised to allow for directionality of items in forming totals and Study 1 was replicated. Table 1 indicates a near perfect solution was achieved.

Also, replication of Study 5 with directionality of items taken into account resulted in near perfect agreement with the direct conventional solution.

Discussion and Conclusions

Five years of research on Horst's indirect method of factor estimation appears to have uncovered gross and serious limitations of the method. It does not appear to have general applicability at this time.

Two critical matters are required for success of the method. Both matters require a prescience on the part of the investigator which is usually lacking. First, it is required that items to be used in totals be of homogeneous factor composition. This is almost paramount to requiring that the investigator know what the factor structure of the data set is before he does the indirect factor analysis!

Second, items which are slanted in opposite directions on a factor (Bipolar Factors) must display this fact by use of opposite sign as they enter totals.

These two prerequisites seem at first glance to be fatal to the method. However, on reflection, it seems that the indirect method may prove useful in either of two ways:

(1) As a test of two or more competing theories as to factor structure of a domain. Note that conventional scale designation on the MMPI theorizes a factor structure and the conventional scoring provides the required directionality for use of the method. The direct factor analysis of the MMPI by Barker, Fowler, and Peterson (1971) provides a competing theory with identification of items for totals together with directionality. Whichever theory results in the greatest correspondence between item factor clusters and item grouping for totals receives strongest support. This study is currently under way. Preliminary results strongly favor the factor theory of Barker, Fowler and Peterson.

(2) Preliminary factoring (or some suitable clustering method) may be used on a random sampling of variables from the data domain of interest in order to provide suggestion as to the appropriate grouping of variables for totals. Application of the indirect solution may be used in successive iterations until no further improvement in the fit between item total grouping and item factor cluster is attained. The next paper (Barker, Barker and Carlton) presents further research edidence on this approach.

Summary

Five previous studies, aimed at implementing Horst's indirect factor solution, were briefly detailed. Most research results served to delimit appropriate use of the method. At present, the method appears useful in evaluating the accuracy of competing factors theories of a data domain. Research continues in an effort to extend use of the method to exploratory research.

Footnote

[1]This research was supported by funds from the Research Grants Committee of the University of Alabama and Roche Laboratories, Division of Hoffmann-La Roche Inc.

References

Barker, H. R., and Barker, B. M. Accuracy of an indirect factor analytic solution as a function of the number of totals used. (Paper presented at Southeastern Psychological Association, Atlanta, Georgia, May 28, 1975).

Barker, B. M., and Barker, H. R. Search for a method of increasing the accuracy of a novel, indirect method of factor analysis. (Paper presented at Southeastern Psychological Association, Atlanta, Georgia, May 28, 1975).

Barker, H. R., Fowler, R. D., and Peterson, L. P. Factor analytic structure of the short form MMPI items in a VA Hospital population. Journal of Clinical Psychology, 1971, 27 (2), 228-233.

Comrey, A. L. A factor analysis of items on the F scale of the MMPI. Educational and Psychological Measurement, 1958, 18, 621-632.

Horst, P. Factor analysis of data matrices. New York: Holt, 1965.

Hunnicutt (Barker), B. M., and Barker, H. R. Application and evaluation of a novel and indirect factor method to a very large data matrix. (Paper presented at Southeastern Psychological Association, Hollywood, Florida, May 3, 1974).

IBM System/360 Matrix Language (Matlan

360 A-CM-05X). Program Description Manual, 1st Ed., 1968, IBM Technical Publications Department.

Mees, H. L. Preliminary steps in the construction of factor scales for the MMPI. Seattle: University of Washington Press, 1959.

Sloan, H. C. A simplified procedure for estimating the factor structure of large data matrices. Unpublished doctoral dissertation. University of Alabama, 1973.

Stallings, N. A. Evaluation of an indirect method of estimating the factor structure of large data matrices. Unpublished doctoral dissertation, University of Alabama, 1973.

Appendix A

COMPUTATIONAL PROCEDURE AND TERMINOLOGY

Proof for an indirect and relatively unconventional procedure for estimating factor loadings was given by Horst (1965).

Assume that we wish to factor analyze 700 items. A 700 x 700 matrix of correlations exceeds the capability of the computer. Therefore, we divided the items into arbitrary sets and total each set. If we divide the items into 10 sets, (each group consisting of 70 items), we will have 10 totals. Let

r_{II} = the matrix of intercorrelations between items.

A_I = the matrix of factor loadings derived from r_{II}.

A'_I = the transpose of the matrix of A_I.

r_{tt} = the matrix of intercorrelations between the 15 totals.

r_{It} = the non-symmetric vertical matrix of correlations between the 700 items and the 10 totals.

A_t = the matrix of factor loadings derived from r_{tt}.

A'_t = the transpose of the matrix of A_t.

Using these definitions, the following equations may be stated:

1. $r_{II} = A_I A'_I$
2. $r_{tt} = A_t A'_t$
3. $r_{It} = A_I A'_t$

The values of r_{tt}, r_{It}, and A_t may be readily computed. Then the factor matrix A_I may be estimated by the following:

4. $A_I = r_{It} \, (A'_t)^{-1}$

This formula can be used whenever the number of factors to be extracted equals the number of totals, thus producing a square matrix. If the number of totals does not equal the number of factors, the following formula is used:

5. $A_I = r_{It} A_t \, (A'_t A_t)^{-1}$

The derivation of A_I is quite feasible using either of these equations and circumvents the computation and factoring of r_{II}.

Table 1

Agreement Between Conventional and Three Indirect Factor Solutions

	Factors I	II	III	IV	V	VI	VII	VIII	IX	Other
Conventional	42	49	27	22	28	18	33	12	8	134
Indirect (1)	35	8	3	0	1	0	11	4	4	307
Indirect (2)	41	23	27	21	28	17	33	12	6	165
Indirect (3)	41	47	27	22	28	18	33	12	8	137

Note: Indirect (1) Sequential selection of items for totals.
Indirect (2) Selection of items for totals according to known factor structure.
Indirect (3) Same as Indirect (2) except allowance is made for directionality of items.