PSYCHOMETRIKA—VOL. 30, NO. 2
JUNE, 1965

A RATIONALE AND TEST FOR THE NUMBER OF FACTORS IN FACTOR ANALYSIS

JOHN L. HORN*

UNIVERSITY OF DENVER

It is suggested that if Guttman's latent-root-one lower bound estimate for the rank of a correlation matrix is accepted as a psychometric upper bound, following the proofs and arguments of Kaiser and Dickman, then the rank for a sample matrix should be estimated by subtracting out the component in the latent roots which can be attributed to sampling error, and least-squares "capitalization" on this error, in the calculation of the correlations and the roots. A procedure based on the generation of random variables is given for estimating the component which needs to be subtracted.

1. *The Rationale*

If m sets of very large samples of size N are drawn independently from a normally distributed population of random numbers and the resulting m "variables" are intercorrelated, it is to be expected that the m by m matrix of correlation coefficients, R, will approximate an identity matrix. (If the distribution is not normal, the expectations outlined here need to be modified.) Sampling theory would argue that the closeness of the approximation is a direct function of both N and m. Likewise, theory would predict that the average correlation is zero and that the variance of the correlation is inversely related to N. (In this development, although the m sets treated as variables are drawn independently in a univariate sense, the bivariate samples upon which the correlations are based are not independent. The standard error for the zero correlation is therefore not that usually employed, but rather a much more complex function (see Kendall and Stuart [5]).)

The latent roots for a matrix of correlations may be viewed as variances for variables that are derived from m intercorrelated variables. The first of these derived variables is the one "best" given by a weighted linear combination of the m original variables—"best" in the sense that all nonzero correlation is assumed to furnish a basis for estimation of the weights in the linear composite. The first component takes up as much of the total matrix variance, given in part by the intercorrelations, as is linearly possible. Successively determined roots have like interpretations among residual matrices. The roots that are first calculated on an R may thus be said to take advantage of chance fluctuations in a particular sample, since they

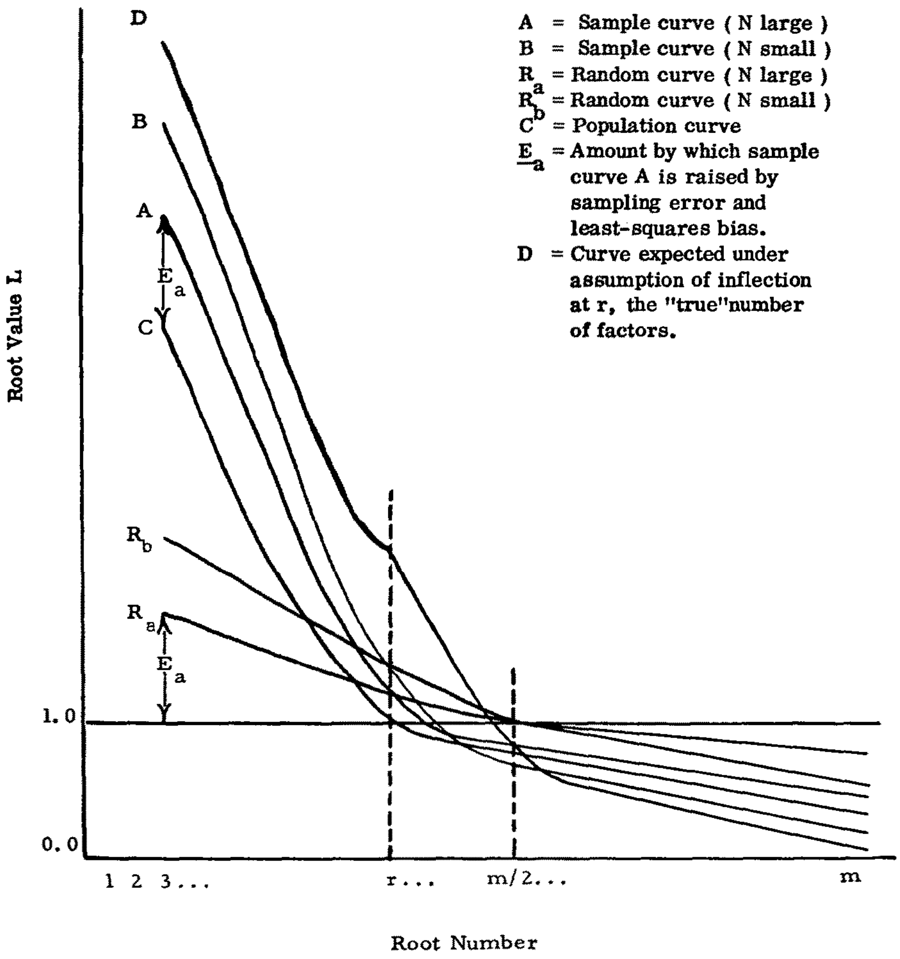
*I wish to acknowledge the valuable help given by J. Jaspers and L. G. Humphreys in the development of the ideas presented in this paper.

are inflated by whatever chance correlation may occur among the set of variables.

If R is determined on m random variables, it is expected to have m nonzero, positive roots, all nearly equal to 1.0. But insofar as the sample is less than infinite, the curve showing values of roots for successively calculated components is not a horizontal straight line at $L = 1.0$ (see Fig. 1) but a curve with slope indicating the extent to which sampling error and least-squares "bias" have combined to increase the value of the correlations and, thence, of the first $m/2$ roots. (In Fig. 1, curves R_a and R_b depict this expectation for large and small samples, respectively.) At the point $m/2$

FIGURE 1

LATENT ROOT CURVES EXPECTED FOR VARIOUS KINDS OF DATA



(on the average for many matrices) the curve is expected to cross the point $L = 1.0$ and thereafter, since the trace must equal m, the curve must decrease below 1.0, approaching but never reaching zero.

Guttman [3] has shown that the "weakest" of three lower bound estimates for the minimum rank of a Gramian R is given by the number of roots of R (with ones in the principal diagonal) which are greater than or equal to unity. This proof is based on an assumption that there is no error due to sampling in the population of subjects, that sampling takes place only in the universe of measures (tests): "We assume throughout that population parameters are used, and not sample statistics" (p. 151). Kaiser [4] and Dickman [2] have argued that this lower bound for the number of factors is also a psychometric upper bound. Kaiser shows that for a "principal component to have positive KR-20 internal consistency, it is necessary and sufficient that the associated eigenvalue be greater than one" ([4], p. 6). Dickman argues that it is not psychometrically reasonable to allow a factor, which is supposed to be a broad, "fundamental" dimension, to have less variance than the unity which is accorded a variable in the standard score space. The widely used "latent-root-one" criterion for when to stop factoring is based on these psychometric considerations.

If data are random, the latent-root-one criterion should (on the average) lead to the decision to estimate $m/2$ factors. But if the data are not random but are not infallible either, the variance of the component must be regarded as due in part to true correlation and in part to correlation resulting from sampling error and least-squares bias. Thus the plot showing the values of successive latent roots calculated in a sample (depicted in Fig. 1 by curves A and B for samples of different size) is expected to be above the curve for the population (curve C in Fig. 1). Therefore, if one accepts the argument that the lower bound, by Guttman's proof, or the most reasonable psychometric upper bound, by Kaiser's and Dickman's arguments, for the number of factors in the population is given at the point where curve C crosses $L = 1.0$, then it is apparent that this bound is given at some point where the latent root for a *sample* is larger than 1.0—how much larger depends upon the size of the sample. In particular, allowing the curve R_a to represent the extent to which sampling error and least-squares bias have increased the roots in the fallible sample A, the number of factors is given at the point where $A - E_a$ crosses $L = 1.0$. The curve $A - E_a$, which is C, could theoretically be obtained by subtracting $R_a - 1.0$ from A at each point along the curve. The abscissa of the point at which C crosses $L = 1.0$ is the same as the abscissa of the point of intersection of R_a and A. Hence, if R_a were known, the number of factors could be determined.

2. *A Test Based on Generated Random Variables*

As far as the writer knows, the statistical theory needed to handle the above problem at a purely analytical level does not exist. Anderson, in his survey of advanced multivariate statistical procedures ([1], p. 307 ff.), gives the joint and marginal distributions for the set of roots of the random vector having a given covariance matrix, but these developments do not lead directly to the equations needed to estimate the expected values of the ordered roots for the model of the rationale outlined above. Other texts (e.g., Rao [6]) and the journals likewise appear to lack an adequate solution for this problem. Several persons are now looking into this problem more fully. Meanwhile, since the values needed to construct the curve R_a cannot yet be obtained directly from formulas, the method below is suggested as a temporary expediency designed to furnish a large-sample solution which may serve until such time as the general statistical problem is solved.

Suppose that an investigator has obtained m measurements on N subjects. Call these measurements "real data." Now generate K matrices of random variables, each matrix of order m by N; intercorrelate the rows of these matrices; find the latent roots for the resulting R matrices; average (over K) the first root values, the second root values, etc. Insofar as K is reasonably large, these averages give the curve R_a in Fig. 1. For a given sample size and a given number of variables, the extent to which the real data latent roots are inflated by sampling error can now be estimated as the value E_a , the amount by which the "empirically obtained" R_a differs from 1.0. The number of factors can then be estimated as the abscissa of the point where R_a intersects A, i.e., by inference, the abscissa of the point where the population curve has the ordinate $L = 1.0$.

3. *An Example*

For purposes of illustration, the above test was tried out on an actual sample of "real data," using one matrix of random variables and comparing the results with those based on another rationale.

Sixty-five ability and nonability behavioral measures were obtained on a sample of 297 adults. These variables were converted to normal distribution form insofar as the ranges and distributions of raw scores permitted. In most cases there were more than 10 different scores in the raw score distributions and the transformed distributions were at least clearly symmetrical and bell-shaped, if not truly normal in form.

A 65 by 297 matrix of numbers from a normally distributed universe of random numbers was generated row by row using a 7090 program written for this purpose. The 65 distributions for rows were again clearly symmetrical and bell-shaped. The approximation to normality for these "variables" appeared to be slightly better than that for the real variables, but tests were not run to establish this as fact. The 65 by 65 matrices of product-moment correlations were determined for both sets of data. Unit values were inserted in the principal diagonals of these R matrices, and all latent roots were calculated. The roots are listed in Table 1 and portrayed graphically in Fig. 2.

In the real data, 16 of the roots are greater than 1.0. If the results giving the curve R_a are accepted at face value, however, and the rationale outlined above is used, only nine factors would be indicated. Interestingly, there is something of an inflection in the real data curve at this point. The ratio of the tenth root to the ninth is small. Some rules for when to stop factoring are, in effect, based on the assumption that there should be a rather

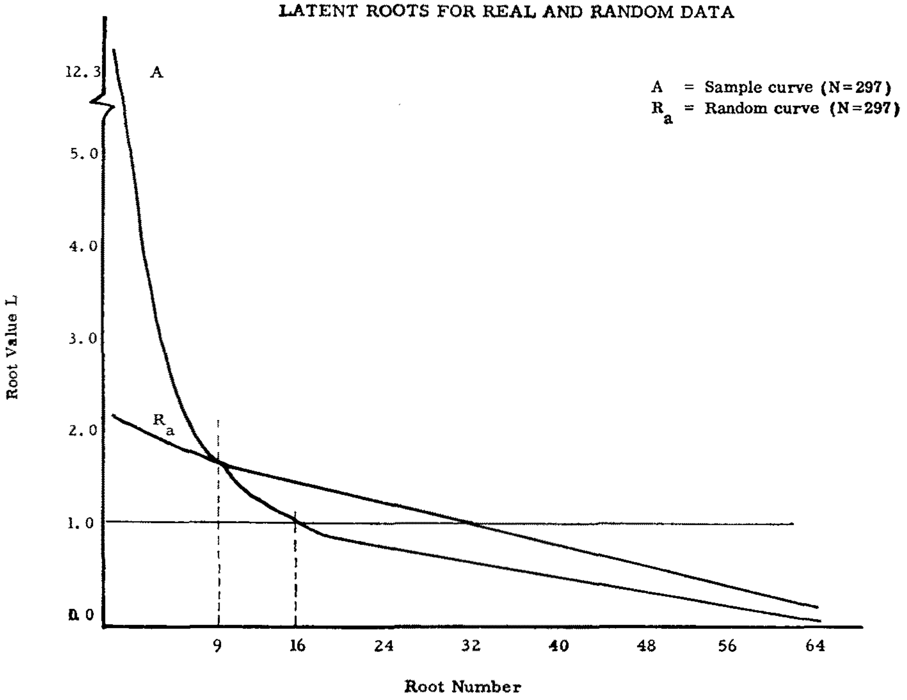
TABLE 1

Latent Roots for Random and Real Data

Root Number	Real Data Root	Random Data Root
1	12.29	2.10
2	4.86	1.95
3	3.75	1.91
4	2.83	1.83
5	2.14	1.76
6	1.84	1.72
7	1.78	1.67
8	1.69	1.64
9	1.62	1.61
10	1.39	1.56
11	1.33	1.54
12	1.25	1.52
13	1.21	1.50
14	1.15	1.42
15	1.09	1.40
16	1.03	1.38
17	.98	1.34
18	.96	1.31
19	.92	1.28
20	.88	1.25
21	.85	1.23
22	.80	1.18
23	.79	1.17
24	.77	1.11
25	.76	1.10
26	.73	1.07
27	.70	1.07
28	.68	1.03
29	.66	1.01
30	.64	1.00
31	.63	.99
32	.62	.92
33	.59	.91
34	.55	.90
35	.53	.87
36	.52	.85
37	.51	.83
38	.49	.81
39	.48	.80
40	.47	.79
41	.45	.75
42	.44	.75
43	.42	.73
44	.40	.71
45	.39	.67
46	.38	.66
47	.38	.65
48	.36	.63
49	.34	.62
50	.32	.60
51	.31	.59
52	.30	.58
53	.30	.56
54	.27	.54
55	.27	.51
56	.24	.50
57	.22	.49
58	.21	.47
59	.19	.46
60	.19	.42
61	.18	.40
62	.18	.40
63	.16	.36
64	.14	.32
65	.13	.31

FIGURE 2

LATENT ROOTS FOR REAL AND RANDOM DATA



sudden drop in the variance accounted for by a factor after the last *true* factor has been calculated. In the present case, application of this kind of rule leads to the estimation of the same number of factors as is suggested by the method developed in this paper. While this is an interesting outcome and suggests a hypothesis to be examined, the results here are not presented as a test of the hypothesis of congruency of the two approaches. Such a test would require variation over several samples of real data as well as variation over several samples of random variables.

It is to be hoped, of course, that the sampling theory required by the rationale given here will soon be developed to the point where the generation of samples of random variables will not be needed. Meanwhile, however, the procedures illustrated above can be rather easily adopted at any institution where fast computer facilities are available. The test based on random variables can be included in standard programs and used routinely.

REFERENCES

[1] Anderson, T. W. *An introduction to multivariate statistical analysis.* New York: Wiley, 1958.

[2] Dickman, K. W. Factorial validity of a rating instrument. Unpublished doctoral dissertation, Univ. Illinois, 1960.

[3] Guttman, L. Some necessary conditions for common-factor analysis. *Psychometrika*, 1954, **19**, 149–161.

[4] Kaiser, H. The application of electronic computers to factor analysis. (Paper read at a symposium on application of computers to psychological problems. Meeting of Amer. Psychol. Ass., 1959).

[5] Kendall, M. G. and Stuart, A. *The advanced theory of statistics* (Vols. I and II). London, Eng.: Griffin, 1958.

[6] Rao, C. R. *Advanced statistical methods in biometric research.* New York: Wiley, 1952.

Original manuscript received 1/16/64
Revised manuscript received 3/25/64