EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1981, 41

INTERPRETING THE FIRST EIGENVALUE OF A CORRELATION MATRIX

SALLY FRIEDMAN
The University of Michigan

HERBERT F. WEISBERG
Ohio State University

The first eigenvalue of a correlation matrix indicates the maximum amount of the variance of the variables which can be accounted for with a linear model by a single underlying factor. When all correlations are positive, this first eigenvalue is approximately a linear function of the average correlation among the variables. While that is not true when not all the correlations are positive, in the general case the first eigenvalue is approximately equal to a lower bound derived in the paper. That lower bound is based on the maximum average correlation over reversals of variables and over subsets of the variables. Regression tests show these linear approximations are very accurate. The first eigenvalue measures the primary cluster in the matrix, its number of variables and average correlation.

REGRESSION analysis employs R^2 to answer the familiar question of what proportion of the variance of a dependent variable can be explained by a set of independent variables. However, the distinction between independent and dependent variables is often not required, particularly in measurement studies where the researcher is interested in determining the extent to which several measures tap the same underlying syndrome. The appropriate statistic in this situation is the first eigenvalue (λ_1) of a principal component analysis. It indicates the maximum amount of the variance of the variables which can be accounted for with a linear model by a single underlying factor.

Since factor analysis is usually treated as a complicated black box,

there are no guidelines for intuitive understanding of its statistics, including the eigenvalues. However, all component analysis does is map the $n(n-1)/2$ correlations among n variables into n eigenvalues and their associated eigenvectors, so the eigenvalues must be functions of those underlying correlations. Even if the functions are complicated and nonlinear, they might still be approximated well by a linear rule.[1] Discovering how the eigenvalues relate to the correlations would increase our intuitive understanding of them and of component analysis more generally. That is our focus in this paper, beginning with the case of all positive correlations and moving on to the general case.

All Positive Correlations

Our linear relationship between the first eigenvalue and the underlying correlations is already known when all correlations are positive. Morrison (1967: 244–245) shows that when all correlations equal $r (r > 0)$,

$$\lambda_1 = 1 + (n-1)r. \tag{1}$$

What relationship holds when all correlations are not equal? Say that all the correlations are set equal but then some correlations are increased. For positive correlations, increasing some of the correlations would increase the amount of the variance which can be accounted for a single component. Of course, the central tendency of the correlations also increases as some correlations are increased. This suggests that the first eigenvalue may be a function of the central tendency of the correlations when the correlations are all positive. Following Morrison's result, we might expect that a linear relationship which would closely approximate the first eigenvalue would be[2]

$$\lambda_1 \approx 1 + (n-1)\bar{r} \tag{2}$$

where $\bar{r}$ is the mean correlation.

To examine the relationship between the first eigenvalue and the central tendency of the correlations, we have constructed a large num-

---

[1] For example, for three variables, the first eigenvalue, λ_1, is

$$\lambda_1 = 1 + 2\, rms\, cos(\tfrac{1}{3} \arccos(r_{12}r_{13}r_{23}/rms^3)),$$

where rms is the root mean square correlation: $\sqrt{(r_{12}^2 + r_{13}^2 + r_{23}^2)}/3$ a function which is not linear in the correlations, even if it might be well approximated by a linear function.

[2] This estimate would also be expected to apply when there are only a few, small, nonsystematic negative correlations.

ber of three and four variable correlation matrices[3] and computed their eigenvalues. Figure A plots the mean correlation for each three variable correlation matrix on the horizontal axis against the corresponding first eigenvalue on the vertical axis. The relationship between the first eigenvalue and the mean correlation is clearly near linear. When the first eigenvalue is regressed on the mean correlation, the best fitting linear regression equations are

$$3 \text{ variables: } \lambda_1' = 1.07 + 1.94\bar{r} \ (R^2 = .990;\ N = 220) \qquad (3)$$

$$4 \text{ variables: } \lambda_1' = 1.05 + 2.97\bar{r} \ (R^2 = .996;\ N = 300).$$

These regressions confirm the rule we suggest in equation (2).

As an example, consider the correlation matrix

1.0	.4	.3
.4	1.0	.3
.3	.3	1.0

[3] All three variable correlation matrices were generated with r values between 0.1 and 1.0 in steps of 0.1 (0.1, 0.1, 0.1; 0.1, 0.1, 0.2; and so on up to 1.0, 1.0, 1.0). A total of 220 correlation matrices resulted. Each point in Figure A represents the $\bar{r}$ and λ_1 obtained for one of these correlation matrices.

Ten sets of positive correlations among four variables were constructed with an emphasis on maintaining variance on the mean, maximum, and minimum correlations along with independence of those statistics. These sets were not chosen with respect to their clustering properties. The sets were as follows:

Set	*Correlations*						*Maximum*	*Minimum*	*Mean*
1:	.80	.75	.70	.65	.60	.10	.80	.10	.60
2:	.90	.40	.35	.30	.25	.20	.90	.20	.40
3:	.90	.84	.78	.72	.66	.60	.90	.60	.75
4:	.40	.34	.28	.22	.16	.10	.40	.10	.25
5:	.95	.92	.89	.86	.83	.80	.95	.80	.875
6:	.20	.17	.14	.11	.08	.05	.20	.05	.125
7:	.61	.60	.59	.41	.40	.39	.61	.39	.50
8:	.90	.80	.70	.30	.20	.10	.90	.10	.50
9:	.80	.76	.72	.68	.24	.20	.80	.20	.567
10:	.80	.76	.32	.28	.24	.20	.80	.20	.433

All 30 distinct permutations of values across variable pairs were yield a total of 300 correlation matrices for four variables.

Note that we have not restricted the correlation matrix to be positive semidefinite for several reasons. A matrix of Pearson r correlations need not be positive semidefinite if there is pairwise deletion of missing data. Also, the matrix is not necessarily positive semidefinite if Yule's Q or tetrachoric r are factor analyzed as an approximation to multidimensional Guttman scaling. As a result, we are investigating the eigenvalues of symmetric matrices rather than features specific to positive semidefinite correlation matrices.

Other central tendency measures—such as the median, the midrange, or the root mean square—could be used in this analysis instead of the mean. For three variables, λ_1 is actually an exact nonlinear function of the root mean square. Only results for the mean are reported here since it yields the highest R^2 values for a central tendency measure which is easy to compute. The median or midrange could be substituted because of their simpler calculation, but at greater loss of predictive accuracy.

Since the average correlation is .333 and the number of variables is three, equation (2) yields the estimate for λ_1 of $1 + (3 - 1)(.333) = 1.667$. The actual first eigenvalue is 1.669, so the estimate is excellent. Actually, the estimate deteriorates slightly as the variance of the correlations increases. Consider the correlation matrix

1.0	.9	.3
.9	1.0	.3
.3	.3	1.0

Since $\bar{r} = .5$ and $n = 3$, equation (2) yields the estimate $1 + 2(.5) = 2.000$. The actual first eigenvalue is 2.068, which is further away from the estimate than in the previous example but which is still quite close to the estimate.

While we have investigated only the three and four variable cases systematically, we would expect equation (2) to hold more generally. As an example, consider Verba and Nie's (1972: 388) matrix of the correlations among thirteen political activities in the United States:

1.0	.71	.60	.19	.19	.18	.13	.17	.18	.17	.09	.11	.12
.71	1.0	.64	.21	.19	.20	.14	.17	.18	.18	.10	.13	.14
.60	.64	1.0	.24	.26	.25	.18	.20	.23	.22	.14	.17	.18
.19	.21	.24	1.0	.47	.35	.27	.27	.22	.24	.23	.24	.22
.19	.19	.26	.47	1.0	.50	.46	.36	.27	.31	.24	.26	.22
.18	.20	.25	.35	.50	1.0	.45	.37	.30	.23	.24	.25	.21
.13	.14	.18	.27	.46	.45	1.0	.36	.20	.22	.22	.17	.19
.17	.17	.20	.27	.36	.37	.36	1.0	.24	.21	.12	.19	.20
.18	.18	.23	.22	.27	.30	.20	.24	1.0	.34	.28	.26	.27
.17	.18	.22	.24	.31	.23	.22	.21	.34	1.0	.38	.29	.23
.09	.10	.14	.23	.24	.24	.22	.12	.28	.38	1.0	.22	.19
.11	.13	.17	.24	.26	.25	.17	.19	.26	.29	.22	1.0	.23
.12	.14	.18	.22	.22	.21	.19	.20	.27	.23	.19	.23	1.0

Since $\bar{r} = .249$ and $n = 13$, equation (2) gives the estimate 3.990,[4] which is very close to the actual (Verba and Nie, 1972:62) 4.05.

The *proportion* of variance accounted for by the first component is λ_1/n. Substituting equation (2) for λ_1, the proportion is approximately equal to $(1/n) + (n - 1)\bar{r}/n$. As the number of variables becomes very large, $(1/n)$ approaches zero while $(n - 1)/n$ approaches one, so this proportion approaches $\bar{r}$. Thus, the proportion of variance accounted

[4] If the mean seems overly tedious to calculate, the median can be substituted. Here, for example, the median correlation is .22, which leads to an estamate of $1 + (13 - 1)(.22) = 3.64$. This would yield an underestimate of the true proportion of variance accounted for by the first component (λ_1/n) of only (4.05-3.64)/13 = 3.15%, so it is still a sufficiently accurate estimate.

for by the first principal component is seen to be basically a function of the central tendency of the underlying correlations.

Before concluding the all positive correlation case, it should be pointed out that Meyer (1975:68) has shown[5] that a lower bound for the first eigenvalue is

$$\lambda_1 \geq 1 + (n - 1)\bar{r}, \tag{4}$$

so equation (2) never overestimates the first eigenvalue. The regressions show that the first eigenvalue hugs the lower bound closely.[6]

The General Case

What happens when not all the correlations are positive? When the correlations are of mixed signs, it is less obvious that increasing some correlations would necessarily increase the first eigenvalue, even though that would increase the mean correlation. As a result, the mean correlation may not lead to as good an estimate of the first eigenvalue. This is shown vividly when we extend our regression experiments to the general case. Again we generated a large number of three and four variable correlation matrices,[7] extracted the eigenvalues, and regressed the first eigenvalue on the mean correlation. The regression equations are:

$$\begin{aligned} &3 \text{ variables: } 1.97 + .04\bar{r} \ (R^2 = .002;\ N = 1771) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (5) \\ &4 \text{ variables: } 2.21 + .41\bar{r} \ (R^2 = .128;\ N = 2400) \end{aligned}$$

[5] Meyer's proof is based on Rayleigh's principle for symmetric matrices (R) that

$$\lambda_1 \geq x'Rx/x'x \geq \lambda_n' \qquad x \neq 0.$$

When x is taken to be a column vector of 1's, equation (4) results.

[6] Morrison (1967) also gives an upper bound for λ_1, which is the largest row-sum of absolute values of the entries in the matrix:

$$\lambda_1 \leq \max_k \sum_i |r_{ik}| .$$

This upper bound is sometimes close to the lower bound, but often it is not, as in the correlation matrix from Verba and Nie where the upper bound is 4.73. In general, we find that the first eigenvalue is closer to its lower bound than to its upper bound, and it is very close to its lower bound as contrasted to its theoretical range from 1 to n.

[7] For three variables, 1771 correlation matrices were generated by using all three variable correlation matrices with r values between -1.0 and $+1.0$ in steps of 0.1.

For four variables, the signs of the correlations of the 300 matrices described in footnote 3 were permuted in eight ways: all correlations positive, only the first correlation negative, only the second negative, only the third negative, . . ., only the sixth negative, and all negative. Any other set of signs on the correlations can be obtained by reversing variables from one of these eight permutations. This procedure results in 2400 correlation matrices for the four variable test.

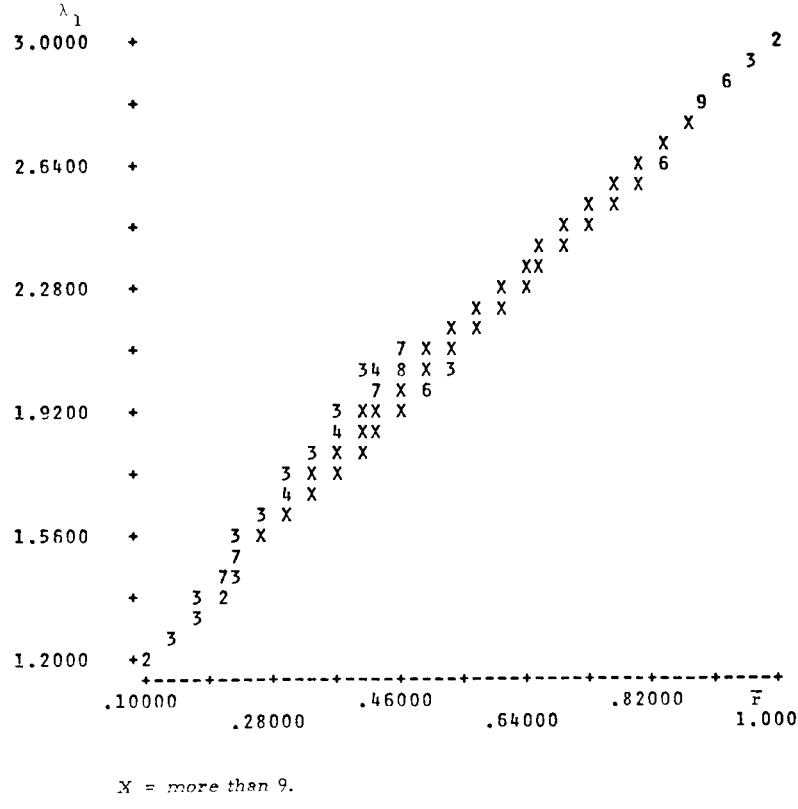


Figure 1.

Obviously the first eigenvalue is not always a linear function of the mean correlation. However, we can use Meyer's lower bound for λ_1 given in equation (4) to examine the general relationship between the first eigenvalue and the mean correlation.

First, say that all correlations are positive except that one variable has negative correlations with all other variables, as in the correlation matrix

$$\begin{matrix} 1.0 & .4 & -.3 \\ .4 & 1.0 & -.3 \\ -.3 & -.3 & 1.0 \end{matrix}$$

Meyer's result shows that $1 + (n - 1)\bar{r} = .867$ is a lower bound for the first eigenvalue. However, reversing the variable with the negative correlations does not affect the eigenvalues, even though it would in-

crease the $\bar{r}$. Therefore, a higher lower bound for the first eigenvalue would be obtained by using in Meyer's formula the average correlation based on the reversed variable. For the above example, reversing the third variable would give an $\bar{r}$ of .333 and a lower bound of 1.667 which is very close to the actual value of 1.669. In general, let $\bar{r}_{\max}$ represent the maximum value of $\bar{r}$ over all possible reversals of the variables. The lower bound for the first eigenvalue then becomes[8]

$$\lambda_1 \geq 1 + (n - 1)\bar{r}_{\max}. \tag{6}$$

Next, say that all correlations are positive except for one variable whose positive correlations with some variables exactly balance off its negative correlations with the other variables, as in the correlation matrix

1.0	.4	.3
.4	1.0	−.3
.3	−.3	1.0

Applying Meyer's result would yield a lower bound of 1 + 2(.133) = 1.267. However, if the variable with the inconsistent correlation signs is deleted, the lower bound of the first eigenvalue for the reduced matrix would be $1 + (n - 2)\bar{r}_{n-1}$, or here 1 + 1(.4) = 1.400. The inclusion principle (Franklin, 1968:149) implies that the first eigenvalue of a matrix is at least as large as the first eigenvalue of any submatrix,[9] so this must also be a lower bound for the first eigenvalue of the original matrix. In the present example, the higher (and, therefore, operative) lower bound is obtained with the submatrix, and the actual first eigenvalue of the original matrix is 1.400 which is identical to that lower bound. If we now let $\bar{r}_{\max}$ be the maximum average correlation among m of the variables for all subsets of size m and for all possible reversals of the variables, the new general lower bound[10] would be

$$\lambda_1 \geq 1 + \max_m ((m - 1)\bar{r}_{\max}). \tag{7}$$

[8] This result can also be derived from Rayleigh's principle, using −1 values in the x vector for variables which would be reversed and +1 values for the other variables.

[9] Franklin's inclusion principle states that the first eigenvalue of a matrix is at least as large as the first eigenvalue of the submatrix obtained by deleting the last row and last column of the original matrix. However, reordering the variables in the matrix will not change the eigenvalues, and induction shows that deletion of more than one row (with the corresponding columns) would lead to a submatrix whose λ_1 cannot be greater than the λ_1 of the original matrix. We shall term this the "generalized inclusion principle."

[10] Let $\sum_{r_1}$ represent the sum of the correlations among a subset of m variables. Let $\sum_{r_2}$ represent the sum of the remaining correlations in the matrix. A higher lower bound is found for the submatrix whenever

$$(n - m)\sum r_1 > m\sum r_2.$$

As an example of the calculations involved in equation (7), consider the correlation matrix

1.00	.90	.40	.20
.90	1.00	.30	.25
.40	.30	1.00	−.35
.20	.25	−.35	1.00

To apply the lower bound formula, we must obtain the largest average correlation for each subset of two or more variables. For the entire matrix, the average correlation is .283, which leads to a lower bound estimate of 1.850 using equation (4). The highest average correlation for a subset of three variables is .533 (using the first three variables), which gives a lower bound estimate of $1 + 2(.533) = 2.067$. The largest absolute value of any correlation is .90, so the highest estimate based on two variable subsets is $1 + (2 - 1)(.90) = 1.900$. Since the first eigenvalue must be larger than all of these lower bounds, it is the 2.067 which provides the operative lower bound. In fact, the first eigenvalue is 2.142, very close to the lower bound.

Given how close the lower bound (4) was to the first eigenvalue for the case of all positive correlations, it is appropriate to examine whether lower bound (7) is similarly close to the first eigenvalue for the general case. The correlation matrices generated for regression equations (5) were reexamined. Figure B plots for each three variable correlation matrix the lower bound estimate on the horizontal axis against the first eigenvalue on the vertical axis, and Figure C gives a similar plot for the four variable matrices. The relationships between the eigenvalue and lower bound (7) are clearly near linear. The best fitting linear regression equations are

$$\text{3 variables: } \lambda_1' = .04 + 1.00\text{LB} \ (R^2 = .987;\ N = 1771)$$

$$\text{4 variables: } \lambda_1' = .02 + 1.04\text{LB} \ (R^2 = .990;\ N = 2400), \qquad (8)$$

where LB is the lower bound obtained from equation (7). Thus, the

---

If the diagonal cell entries are not equal to unity (as would be the case for the classical factor model), substitution of a vector of one's in Rayleigh's principle yields a lower bound for λ_1 of $\sum_{jk}\sum r_{jk}/n$. Employing the generalized inclusion principle on submatrices of m variables yields a lower bound:

$$\lambda_1 \geq \max_m \left(\sum_{jk} \sum r_{jk}/m \right).$$

(Once again, it would be appropriate to maximize this also over possible reversals of variables.) This lower bound would also hold if cell values are greater than 1.0, so this is a general result for symmetric matrices. Recent work on bounds for nonsymmetric matrices (for which Rayleigh's principle does not apply) is reviewed in Brauer and Gentry (1976).

FIGURE B

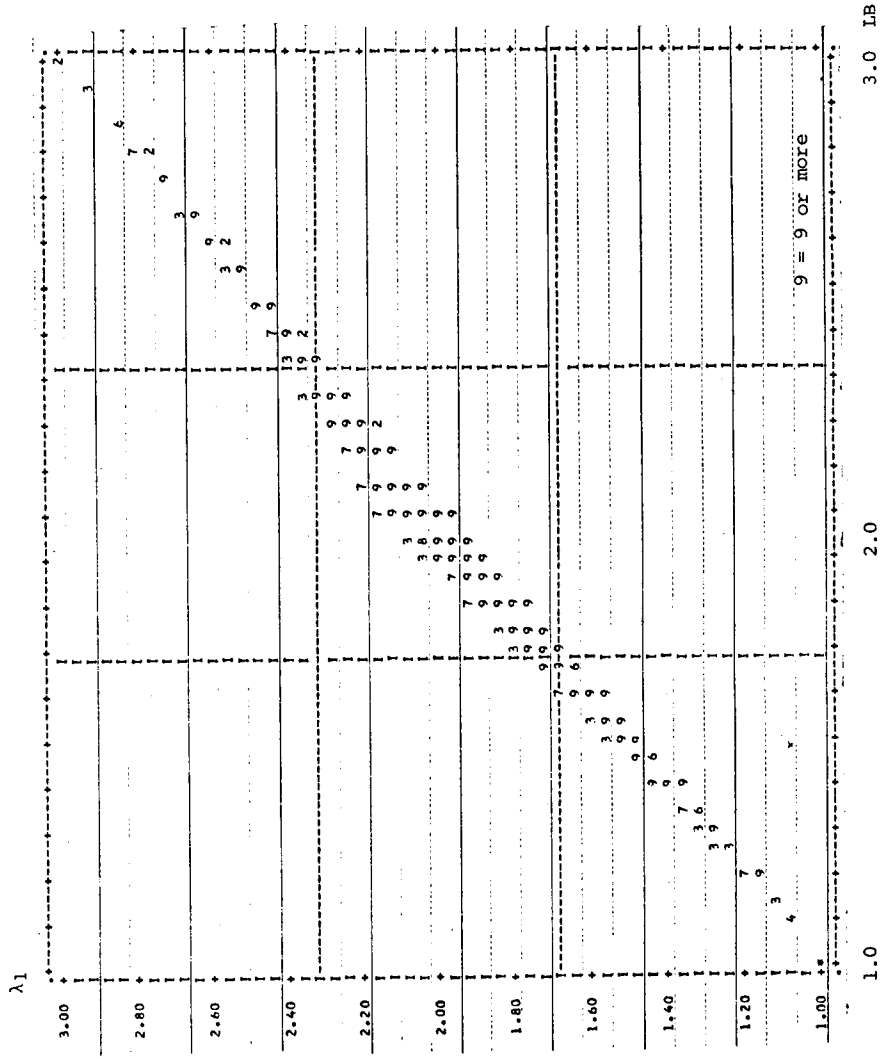
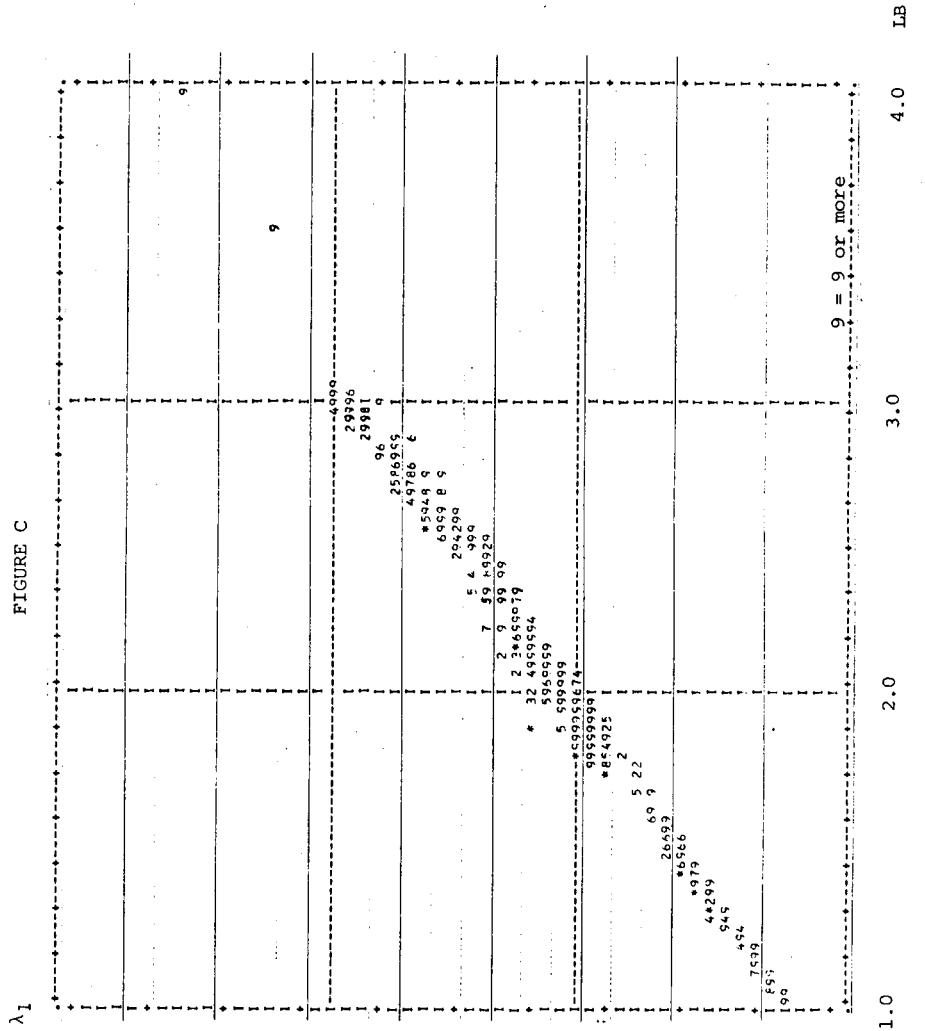

λ1
3.00
2.80
2.60
2.40
2.20
2.00
1.80
1.60
1.40
1.20
1.00
1.0
2.0
3.0 LB
9 = 9 or more

FIGURE C



lower bound gives a very accurate estimate of the first eigenvalue:

$$\lambda_1 \approx 1 + \max_m ((m - 1)\bar{r}_{max}). \quad (9)$$

What properties of the underlying correlations are being monitored by the first eigenvalue in the general case? The size of the primary cluster of variables in the matrix, in terms of its number of variables and its average correlation.[11] It should be noted that this estimate (9) is also appropriate for the all positive correlation case and, if the variables are highly clustered, the revised lower bound can lead to a higher (and hence more accurate) estimate for the first eigenvalue than our original estimate (2).

Conclusions

Computers today permit rapid calculation of eigenvalues, but the analysis here shows how they can easily be estimated by inspection of the correlation matrix. Additionally, this analysis shows how they should be interpreted in terms of the underlying correlations. The first eigenvalue measures the primary cluster in the matrix, its number of variables and average correlation.

[11] Since the second eigenvalue of a correlation matrix is the largest eigenvalue of the matrix obtained by residualizing on the first principal component, that second eigenvalue must similarly monitor the "secondary cluster" of variables, and so on for later eigenvalues.

REFERENCES

Brauer, A. and C. G. Ivey. Bounds for the greatest characteristic root of an irreducible nonnegative matrix II. *Linear Algebra and Its Applications,* 1976, 13, 109–114.

Franklin, J. N. Matrix algebra. New Jersey: Prentice-Hall, 1968.

Mayer, E. P. A measure of the average intercorrelation. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1976, 35, 67–72.

Morrison, D. R. Multivariate statistical methods. New York: McGraw-Hill, 1967.

Verba, S. and Nie, N. H. Participation in America. New York: Harper & Row, 1972.