The Internal Correlation: Its Applications in Statistics and Psychometrics

Article *in* Journal of Educational and Behavioral Statistics · September 1989

DOI: 10.3102/10769986014003211

2 authors:

George Joe
Texas Christian University
151 PUBLICATIONS **7,479** CITATIONS

Jorge L. Mendoza
University of Oklahoma
51 PUBLICATIONS **1,442** CITATIONS

Journal of Educational Statistics
Fall 1989, Vol. 14, No. 3, pp. 211–226

The Internal Correlation: Its Applications in Statistics and Psychometrics

George W. Joe
Texas Christian University
and
Jorge L. Mendoza
Texas A&M University

Key words: *internal correlation coefficient, canonical correlation, ridge regression*

The internal correlation, a measure of dependency in a set of variables, is discussed and generalized. This coefficient is an upper bound to the product moment correlations, multiple correlations, and canonical correlations that can be defined in a set of variables. Applications of the internal correlation coefficient and its generalizations are given for a number of data-analytic situations. Where appropriate, we discuss tests of significance. We illustrate the internal correlation and expand the concept to a series of additional indices: local internal, up-internal, and down-internal correlations. Uses of these indices are illustrated in several areas: multicollinearity, ridge regression, factor analysis, principal components analysis, and test reliability.

The concept of dependence among variables, or linear composites of variables, as represented by the correlation coefficient, is frequently used in educational and psychological research. Two correlation coefficients that have been particularly important in the discussion of the relationship among variables are the multiple correlation coefficient (the correlation between a variable and an optimal linear combination of a set of variables), and the canonical correlation (the correlation between two optimal linear combinations). We will discuss another correlation that may be equally useful to the research community. A relatively new correlation coefficient proposed by Schuenemeyer and Bargmann (1978), which assesses dependence among a set of variables and encompasses the usual correlation coefficient, the multiple correlation, and the canonical correlation, is the internal correlation coefficient ρ(SB). It is surprising that this coefficient has received little attention in the applied literature.

The internal correlation coefficient is the maximum correlation between two linear composites $Y_1 = \underline{a}^T \underline{X}$ and $Y_2 = \underline{b}^T \underline{X}$ of the same set of variables

The authors wish to thank the editors and the referees for their helpful comments.

subject to the constraint that the two weighting vectors are orthogonal. The index combines the dependence represented by the correlation coefficient, the multiple correlation coefficient, and the canonical correlation coefficient into a single coefficient. It is an upper bound to the maximum correlation possible between any linear combinations of two disjoint subsets of variables in a given set of variables, $\underline{X}^T = (X_1, \ldots, X_p)$ (Eaton, 1976; Schuenemeyer & Bargmann, 1978). Consequently, it is a measure of independence/dependence among a set of variables. This correlation coefficient is

$$\rho(\mathrm{SB}) = (\lambda_1 - \lambda_p)/(\lambda_1 + \lambda_p),$$

where λ_1 and λ_p are the largest and smallest eigenvalues of the population correlation matrix, $\underline{R}$, respectively. (The derivation assumes that the eigenvalues are greater than zero.) Correspondingly, the weights for the linear composites are $\underline{a} = \underline{e}_1 - \underline{e}_p$ and $\underline{b} = \underline{e}_1 + \underline{e}_p$, where $\underline{e}_1$ and $\underline{e}_p$ are the population eigenvectors corresponding to λ_1 and λ_p, respectively. Throughout the paper we denote the sample estimate of $\rho(\mathrm{SB})$ by $r(\mathrm{SB})$, and the sample estimates of λ, $\underline{a}$, $\underline{b}$, $\underline{R}$, and $\underline{\Sigma}$ (population covariance matrix) by $\hat{\lambda}$, $\hat{\underline{a}}$, $\hat{\underline{b}}$, $\hat{\underline{R}}$ and $\underline{S}$, respectively.

Venables (1976) derived a coefficient, $\rho(V)$ as a union-intersection test statistic for sphericity, using the covariance matrix, $\underline{\Sigma}$, and assuming that the sample covariance matrix has a Wishart distribution. Venables's index is similar to $\rho(\mathrm{SB})$ but uses the eigenvalues of the covariance matrix instead of the eigenvalues of the correlation matrix $\underline{R}$. In general, those two coefficients differ; that is, $\rho(\mathrm{SB}) \neq \rho(V)$. It is interesting that Bush and Olkin (1961) may have been the first to imply the internal correlation coefficient in their consideration of bounds for a bilinear form for a square symmetric matrix. Also, Khatri (1978) presented several indices, based on functions of internal correlations, representing the total dependence in a set of variables. Because most of the research in education and in psychology deals with the correlation matrix and not with the covariance matrix, we will concentrate on the Schuenemeyer-Bargmann index. This index will be denoted by $\rho(*)$ when there is no need to differentiate between the two indices.

Research on all internal indices, for the most part, has dealt only with theoretical issues, and almost no information is available regarding applications. The main purpose of this paper is to introduce these indices, and to show how they can be used in data-analytical situations. In some instances, these applications represent a different way of obtaining information from a correlation (or a covariance) matrix, which usually has been done through the examination of eigenvalues and condition indices. In other instances, the applications show that additional information, not readily discernible from commonly used indices, may be gained from using the internal correlation. The applications are concerned with identifying analytical problems caused by dependence among variables within a set of variables, determining whether the degree of dependence is of a practical nature, or computing other indices.

To illustrate the maximum internal correlation is an optimal measure of dependence in a set of variables and to demonstrate the usefulness of examining the weights associated with it, consider the following example. Let us use four variables, say *A, B, C,* and *D*. Table 1 gives the correlation matrix, $\underset{\sim}{R}$, for these variables. It may be of interest in some situations to examine the "dependence" in a set of variables, and hence to be able to find an upper bound to the following correlations: $\rho(A.BCD)$, $\rho(B.ACD)$, $\rho(C.ABD)$, $\rho(D.ABC)$, $\rho(AB.CD)$, $\rho(AC.BD)$, and $\rho(AD.BC)$. The period within the parentheses should be read as "with." The reader should note that this set defines the set of all possible (multiple and) canonical correlations, and that at least one of the $2^{p-1} - 1$ elements in this set is larger than all of the elements in the correlation matrix, where p represents the number of variables. As noted, the maximum internal correlation is an upper bound to any (pairwise, multiple, canonical) correlation involving the set of variables *A, B, C,* and *D* or subset of these variables. The maximum internal correlation is defined by

$$\rho(*) = (\lambda_1 - \lambda_p)/(\lambda_1 + \lambda_p), \qquad (1)$$

TABLE 1
Correlation matrix with eigenvalues and vectors

	A	*B*	*C*	*D*
	1.000	0.610	0.669	0.638
		1.000	0.482	0.666
			1.000	0.710
				1.000
		Eigenvalues		
	(2.891	0.523	0.368	0.217)
		Eigenvectors		
	0.5051	−0.1098	0.8004	−0.3035
	0.4738	0.7721	−0.0330	0.4222
	0.4960	−0.6259	−0.1813	0.5738
	0.5238	0.0002	−0.5704	−0.6327
	0.8086			0.2016
$a =$	0.0516		$b =$	0.8960
	0.0778			1.0698
	1.1565			−0.1089

where λ_1 is the largest eigenvalue of $\underset{\sim}{R}$ (the correlation matrix), and λ_p is the smallest one. For our example with variables A, B, C, and D, the multiple and canonical correlations are

$$\begin{aligned}\rho(A.BCD) &= 0.748,\\ \rho(B.ACD) &= 0.714,\\ \rho(C.ABD) &= 0.768,\\ \rho(D.ABD) &= 0.803,\\ \rho(AB.CD) &= 0.750,\\ \rho(AC.BD) &= 0.752,\\ \rho(AD.BC) &= 0.856.\end{aligned}$$

The maximum internal correlation $\rho(*)$ is

$$\rho(*) = (2.891 - 0.217)/(2.891 + 0.217) = 0.860.$$

The reader should note that $\rho(*)$ is very close to the largest canonical correlation, and that the vectors $\underset{\sim}{a}$ and $\underset{\sim}{b}$ given in Table 1 indicate the optimal variable split. This is an important feature of the internal correlation as we will see later in the detection of multicollinearity. In general, this upper bound is close to the largest canonical correlation attainable in the set, especially when the correlation matrix is used. The eigenvalues of the covariance matrix, $\underset{\sim}{\Sigma}$, also can be used in Equation 1 to obtain the Venable index, $\rho(V)$.

Because the correlation matrix (or covariance matrix) may have more than two roots, additional local internal correlations may be defined. The idea of additional local internal correlations has not really been discussed in the literature, but it has been implied in the work of Khatri (1978). We define the ith local maxima population internal correlation as

$$\rho(*)_i = (\lambda_i - \lambda_{p-i+1})/(\lambda_i + \lambda_{p-i+1}), \text{ where } i = 1, \ldots, [p/2].$$

The corresponding weights for the linear composites are $\underset{\sim}{a}_i = \underset{\sim}{e}_i - \underset{\sim}{e}_{p-i+1}$ and $\underset{\sim}{b}_i = \underset{\sim}{e}_i + \underset{\sim}{e}_{p-i+1}$, where $\underset{\sim}{e}_i$ and $\underset{\sim}{e}_{p-i+1}$ are the population eigenvectors that correspond to λ_i and λ_{i-p+1}, respectively.

In some applications the correlation (or covariance) matrix *is* based on variables that are assumed to be stochastic. Hence, the matrix $\underset{\sim}{R}$ (or $\underset{\sim}{\Sigma}$) is not known, but it can be estimated from a sample. In this case, the distribution of the estimate of the ith internal correlation coefficient is not known. However, Srivastava and Khatri (1979, p. 215) indicate how a confidence interval can be obtained using tables calculated by Krishnaiah and Schurmann (1974) for the first (maximal) internal correlation when it is based on the sample covariance matrix, S. In addition, James and

Venables (1980) present the distribution of the internal correlation and a test of significance that has an approximate F distribution under the special condition of a 2×2 covariance matrix.

Under the assumption that the eigenvalues are distinct and greater than zero, Venables (1976) showed the relationship between a likelihood ratio test for sphericity and the internal correlation, but he did not present a test for the hypothesis that the population internal correlation is zero. However, when the internal correlation is based on the covariance matrix, assuming that the eigenvalues are distinct and greater than zero, it is possible to use the asymptotic procedure given by Steiger and Browne (1984), or Srivastava and Khatri's (1979) confidence interval, to test the hypothesis that the population internal correlation [$\rho(V)$] is zero. Unfortunately, the procedure is not appropriate for the internal correlation of a correlation matrix (Steiger & Browne, 1984).

All of the known statistical tests are for internal correlations computed from the sample covariance matrix. No distribution theory or statistical tests exist for the internal correlation based on a sample correlation matrix. Nevertheless, we believe that perhaps Efron's (1979) bootstrap methods could be useful in obtaining a confidence interval for the internal correlation based on a sample correlation matrix. However, it is not known whether the bootstrap methods apply to functions of a (multivariate) correlation matrix. The paper most closely related to this issue appears to be that by Beran and Srivastava (1985), in which they studied the applicability of the bootstrap methods to eigenvalues, eigenvectors, and other functions of a covariance matrix. To date, no study has examined the applicability of the bootstrap methods to functions of a correlation matrix. In view of the properties of data collected in education and psychology, a confidence interval for $\rho(\mathrm{SB})$ based on the bootstrap method would be most desirable because of the bootstrap's nonparametric attributes. An ad hoc procedure, which would not be a test for the internal coefficient but which may be close enough, would be one based on the canonical correlation defined by the variable split indicated by the pair of vectors, $\underline{a}_1$ and $\underline{b}_1$, corresponding to the maximal internal correlation.

In the remainder of the paper we demonstrate how the internal correlation coefficient can be used to assist in identifying multicollinearity in a data set, to give some insight into ridge regression, to aid in the selection of factors in a factor analysis, to aid in the selection of components in a principal components analysis, and to calculate reliability in a test under the assumptions of the classical test model.

The Internal Correlation in Multiple Regression

It is not uncommon when dealing with a relatively large number of independent variables (predictors) in a regression analysis to find that some of the variables may be highly correlated. This high degree of linear relationship among the independent variables causes a condition known as multicollinearity. A very high linear relationship among two or more independent variables might result in a singular covariance/correlation matrix (which results in a matrix that cannot be inverted), or a high degree of multicollinearity might create a "near singular" matrix that is subject to round-off errors when it is inverted. Multicollinearity not only affects the analysis in terms of accuracy, but in terms of precision of future predictions. Multicollinearity affects the accuracy of the analysis, because most computer algorithms cannot provide accurate estimates of the regression parameters $\underline{\hat{B}}^T = (\hat{B}_1, \ldots, \hat{B}_p)$ when $\underline{\Sigma}$ is singular or near singular. The precision of future prediction is affected, because the squared standard error of estimate for each $\hat{B}_i$ is

$$V(\hat{B}_i) = \sigma_e^2 \underline{\Sigma}_{ii}^{-1},$$

where $(\underline{\Sigma}_{ii})^{-1}$ represents the ith diagonal element of $\underline{\Sigma}^{-1}$. That is, the standard error of estimate is a function of

$$H_i = [(1 - R_i^2)N]^{-1},$$

where N is the sample size, and R_i^2 is the squared multiple correlation between x_i and all other independent variables. An increase in R_i^2 increases H_i and consequently the standard error of estimate for $\hat{B}_i$. Specifically, the standard error is inflated by the quantity

$$VIF_i = (1 - R_i^2)^{-1},$$

which is called the variance inflation factor.

Many measures of multicollinearity have been proposed in the literature (see Dillon & Goldstein, 1984, chapter 7). However, one measure that has received considerable attention, has a sound theoretical basis (Belsley, Kuh, & Welsch, 1980, p. 102), and is reported by one major computer package (SAS, proc REG), is the ratio K (condition index), where

$$K = (\lambda_1/\lambda_p)^{1/2},$$

and where λ_1 and λ_p are the largest and smallest eigenvalues, respectively, of the correlation matrix, $\underline{R}$ (or the covariance matrix, $\underline{\Sigma}$). This (condition) index is commonly used as an indicator of multicollinearity in the data, with large values indicating multicollinearity. When the sample estimate of this index is very large, the estimate $\underline{\hat{B}}_i$ may contain a fair amount of numerical error, as well as have a very large standard error. A disadvantage with this approach is that the condition index is not bounded from above, and "very large" is difficult to define and has not been defined in the literature.

However, it is simple to show that the condition index $(\lambda_1/\lambda_p)^{1/2}$ may be rewritten as a monotone function of $\rho(^*)$:

$$K^2 = (\lambda_1/\lambda_p) = [1 + \rho(^*)]/[1 - \rho(^*)].$$

Similarly, K^2 can be written as

$$\rho(^*) = (K^2 - 1)/(K^2 + 1).$$

It is clear then that the internal correlation coefficient can be used as an additional measure of multicollinearity. For this particular application of the internal correlation, the information on multicollinearity is essentially the same as given by the condition index, but the internal correlation coefficient has the advantage of being bounded from above by 1 and from below by 0. Also, because it is a correlation coefficient and because of their experience with interpreting correlations, many researchers may find that the internal correlation has more heuristic value. Another advantage of the internal correlation is that by examining the weight vectors, $\underline{a}$ and $\underline{b}$, one can gain insight into the variables that are causing the multicollinearity problem. Examining the weights can be helpful in identifying the nature of the multicollinearity, especially in situations where the multicollinearity is due to a complex interrelationship among the variables.

Consider a 10-variable correlation matrix that has the eigenvalues given in Table 2. The condition index for this matrix is 118.11. This is undoubtedly a large index, but unless the reader has had considerable experience with multicollinearity or numerical analysis, the index does not connote much information. On the other hand, the internal correlation coefficient $\rho(^*)$ is 0.999, clearly indicating very high multicollinearity. From

TABLE 2
Condition indices and internal correlations

No.	Eigenvalues	Up- condition indices	Up- internal correlations	Down- condition indices	Down- internal correlations
1	9.7657	1.00	0	118.11	0.999
2	0.0648	12.27	0.986	9.62	0.978
3	0.0426	15.14	0.991	7.80	0.967
4	0.0397	15.68	0.992	7.53	0.965
5	0.0322	17.40	0.993	6.78	0.957
6	0.0232	20.52	0.995	5.76	0.941
7	0.0146	25.86	0.997	4.57	0.908
8	0.0127	27.76	0.997	4.26	0.895
9	0.0037	51.55	0.999	2.30	0.681
10	0.0007	118.11	0.999	1.00	0

an interpretive standpoint, this index connotes more information about dependence among the variables than the condition index, K.

Additional condition indices can be obtained by forming "up-condition ratios," $K(u)_i$, which are currently presented in the SAS collinearity diagnostics, and "down-condition ratios," $K(d)_i$, where

$$K(u)_i = (\lambda_1/\lambda_i)^{1/2},$$

and

$$K(d)_i = (\lambda_i/\lambda_p)^{1/2} \qquad \text{for } i = 2, \ldots, p.$$

Similarly, we can define corresponding up-internal correlations by

$$\rho(u)_i = (\lambda_1 - \lambda_i)/(\lambda_1 + \lambda_i),$$

and the corresponding down-internal correlations by

$$\rho(d)_i = (\lambda_i - \lambda_p)/(\lambda_i + \lambda_p).$$

The sets of weights corresponding to these two correlations are $(\underline{e}_1 - \underline{e}_i)$ and $(\underline{e}_1 + \underline{e}_i)$ for $\rho(u_i)$ and $(\underline{e}_i - \underline{e}_p)$ and $(\underline{e}_i + \underline{e}_p)$ for $\rho(d_i)$. Table 2 gives the condition indices and the corresponding up and down internal correlation coefficients for the 10-variable correlation matrix. The indices and correlations could be used in identifying dependencies in the data. The magnitudes of the internal correlations in Table 2 indicate that there are many near dependences among the columns of the data matrix. As a matter of fact, the up-internal correlations indicate that the data are unidimensional. This information would be harder to discern from the condition indices, but not from the eigenvalues. Of course, this use of the indices duplicates the information one would gain by looking at both the eigenvalues and the condition indices.

It is difficult to specify which type of internal correlation (up, down) is the better diagnostic tool. Appropriateness varies from situation to situation. In general, however, when the first eigenvalue dominates, the down-correlations vary more than the up-correlations. On the other hand, the up-correlations are sensitive to differences between the largest eigenvalue and the ith eigenvalue ($i = 2, \ldots, p$). This suggests that the up-correlations would tend to be more effective in identifying unidimensionality, whereas the down-correlations would be more appropriate for identifying multidimensional spaces. Both types of correlations, of course, are equally effective in identifying multicollinearity. They could differ, however, in diagnosing the causes.

Ridge Regression and the Internal Correlation

Ridge regression was designed to overcome the problems encountered when dealing with ill-conditioned data (multicollinearity). Basically, ridge regression tries to stabilize the parameter estimates by trading off unbiasedness in estimation for a reduction in the variance of the estimate. Hundreds of articles have dealt with one or another aspect of ridge regression, and it would be impossible for us to review them here. The reader is referred to Price (1977), Rozeboom (1979), and Freund and Minton (1979) for additional information on ridge regression.

Ridge regression estimates are provided by the formula

$$\underline{\hat{B}}^* = (\underline{R} + k\underline{I})^{-1}\underline{X}^T\underline{y},$$

where $k\underline{I}$ represents a scalar multiplication of the identity matrix $\underline{I}$ by a small positive value k, and R is the correlation matrix of the independent variables. In the (ridge) regression model the correlation matrix R is assumed to be fixed and known. Many solutions have been suggested for k. For our purpose, it suffices to say that k generally will be a value between 0 and 1 (see Freund & Minton, 1979, chapter 5).

One can see how ridge regression works by computing the maximum internal correlation of

$$\underline{R}^* = (R + k\underline{I}).$$

The eigenvalues of $\underline{R}^*$ are the eigenvalues of $\underline{R}$ plus the constant k; that is,

$$\lambda_i^* = \lambda_i + k.$$

Then the internal correlation coefficient for $\underline{R}^*$ is given by

$$\rho(^*) = (\lambda_1 - \lambda_p)/(\lambda_1 + \lambda_p + 2k).$$

To illustrate the point let us go back to the eigenvalues given in Table 2. The internal correlation, $\rho(^*)$, for the correlation matrix corresponding to these eigenvalues is 0.999. Now suppose that one would have performed a ridge regression and obtained a value of $k = 0.8$. The internal correlation of $\underline{R}^*$ with $k = 0.8$ is 0.859. This is a reduction of 0.14 from the maximum internal correlation obtained earlier. Ridge regression stablilizes the estimates by reducing the multicollinearity in $\underline{R}^*$. This is not the sole reason why ridge works better than ordinary least squares in some situations, but it gives an alternative explanation of the procedure. The change in the internal correlation of the augmented matrix, $\underline{R}^*$, provides a measure (scale) of how much the multicollinearity in the matrix, $\underline{R}$, has been reduced, and hence provides a means of comparing the stability of regression solutions. Simply knowing what value k was in the ridge regression solution does not give the researcher an idea of how much multicollinearity was reduced in the solution.

We can gain insight into the relationship between p and k with the internal correlation. The effect of p (the number of variables) on ridge regression can be seen by assuming multicollinearity and computing the maximum internal correlation. Assume that we have a severe case of multicollinearity; that is, the largest eigenvalue approaches a value of p [trace$(R) = \sum_i \lambda_i = p$], and the smallest approaches 0. In this case, the maximum internal correlation coefficient approaches

$$\rho(*) = p/(p + 2k).$$

We can see from the internal correlation above that as p increases, the value of k needed for the ridge solution also increases.

The internal correlation could also help in determining the value of k to use in ridge regression. The value of $\rho(*)$ could be included in the ridge trace. The ridge trace plots the ridge regression estimates for various values of k. The value of k that leads to stable regression coefficients along with a small value of $\rho(*)$ would be selected. Alternatively, one might use as the initial estimate of k the value

$$k = [(\lambda_1 - \lambda_p) - (\lambda_1 + \lambda_p)\bar{\rho}(*)]/2\bar{\rho}(*),$$

where $\bar{\rho}(*)$ is selected as an acceptable value of $\rho(*)$ that does not signify multicollinearity. The values of $\bar{\rho}(*)$ are then changed by small increments to obtain a nonnegative estimate for k.

The Internal Correlation Coefficient and Factor Analysis and Principal Component Analysis

Because the internal correlation, $\rho(*)_1$, is a measure of total dependence among a set of variables, it follows that it and the local maxima internal correlations, $\rho(*)_i = (\lambda_i - \lambda_{p-i+1})/(\lambda_i + \lambda_{p-i+1})$, might be useful in identifying the number of factors and components to retain in exploratory factor analyses or in principal component analyses, respectively. Here we restrict ourselves to general comments on using the coefficient for its heuristic value in this area.

In maximum likelihood factor analysis, the maximum internal correlation of the partial covariance matrix can be computed at each stage of the factoring, and it can be checked to determine if additional factors are needed. Under the usual constraints, the partial covariance matrix in maximum likelihood factor analysis approaches a diagonal matrix; that is, $\underline{S} - \underline{FF}^T \doteqdot \underline{D}$. Because the maximum internal correlation is an upper bound to product moment correlations (and partial correlations for residual matrices in which factors have been extracted), the magnitudes of the successive internal correlations might be helpful in determining whether to keep a factor for rotation. This follows because the internal correlation would be an upper bound to the product of the two largest factor loadings. The latter follows from the fact that $\sum_k f_{ik} f_{jk} < r_{ij}$, where r_{ij} is the product moment correlation between the i and j variables in the correlation matrix. There-