Eigenvalue Shrinkage in Principal Components Based Factor Analysis

Philip Bobko
Virginia Polytechnic Institute and State University

F. Mark Schemmer
Advanced Research Resources Organization, Bethesda MD

The concept of shrinkage, as (1) a statistical phenomenon of estimator bias, and (2) a reduction in explained variance resulting from cross-validation, is explored for statistics based on sample eigenvalues. Analytic solutions and previous research imply that the magnitude of eigenvalue shrinkage is a function of the type of shrinkage, sample size, the number of variables in the correlation matrix, the ordinal root position, the population eigenstructure, and the choice of principal components analysis or principal factors analysis. Hypotheses relating these specific independent variables to the magnitude of shrinkage were tested by means of a monte carlo simulation. In particular, the independent variable of population eigenstructure is shown to have an important effect on shrinkage. Finally, regression equations are derived that describe the linear relation of population and cross-validated eigenvalues to the original eigenvalues, sample size, ordinal position, and the number of variables factored. These equations are a valuable tool that allows researchers to accurately predict eigenvalue shrinkage based on available sample information.

Sample-based factor analysis has become an important tool in many areas of psychology ranging from industrial/organizational theory to models of personality. The technique has also been criticized on a variety of dimensions (cf. Armstrong, 1967). Perhaps the most common criticisms are based on the psychometric freedom a researcher has in using factor analysis to generate a model. There is wide latitude in the choice of particular factor models (e.g., unities or communalities in the correlation matrix) and of decision rules (e.g., the number of factors) as well as latitude in the verbal interpretation of the results of a factor analysis. A second basis for concern stems from the largely unanswered question of the degree to which sample-based factor statistics estimate their respective population parameters.

After conducting sample-based principal components (or principal factors) analyses, it has become fairly routine practice to report the value of $\sum_{i=1}^{k}\hat{\lambda}_i$, where the $\hat{\lambda}_i$ are the sample eigenvalues and k is the number of components (or factors) extracted. In general, k is substantially less than the number of original variables, p (cf. Everett, 1983, or Zwick & Velicer, 1982, for reviews of stopping rules). Note that if a correlation matrix (with unities in the diagonal) is factored, then the total variance (for all variables) equals the number of variables, p. Thus, a common statistic in the psychological literature is $\hat{\lambda}_i/p$, which is interpreted as the proportion of total variance accounted for by the ith principal component. Similarly, $\sum_{i=1}^{k}\hat{\lambda}_i/p$ can be used as an index of the proportion of variance accounted for by the first k principal

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 8, No. 4, Fall 1984, pp. 439–451

0146-6216/84/040439-13$1.90

components. In social science research, the magnitude of $\sum_{i=1}^{k}\hat{\lambda}_i$ is often used to demonstrate the practical significance of the analysis; that is, to show that the k factors span a sufficiently large portion of the p-dimensional variable space. Assuming that the correlation matrix is nonsingular, all p principal components are required to explain all of the variance. However, in the majority of instances, the first few principal components will account for most of the variance.

Given that the above statistics are routinely reported (e.g., for invoking the practical significance of the results), concern is with their accuracy as point estimates. That is, if the k extracted components are used in future samples, to what extent will they continue to explain the same proportion of total variance? Or, to what extent does the sample-based value of $\sum_{i=1}^{k}\hat{\lambda}_i$ reflect the population value $\sum_{i=1}^{k}\lambda_i$? These issues, generally labeled "validity shrinkage" in multiple regression literature (e.g., Darlington, 1968), are crucial for the accurate interpretation of sample factor results. The purpose of this paper is to determine the severity of eigenvalue shrinkage and to develop equations for predicting the magnitude of such shrinkage.

Shrinkage

Lee and Comrey (1979) have indicated that principal components form the basis for more than two-thirds of all factor analyses reported in the literature. It is crucial to remember that each principal component is derived from a maximization perspective. For example, the first principal component is the linear combination of variables which maximizes the variance of that combination in the original sample (subject to the constraint that the squared weights sum to 1.0). Also, it is crucial to remember that such sample eigenvalues are based on all properties of the sample—including sample idiosyncracies. Thus, it would be expected that application of the first few original eigenvectors to a new sample would result in composite variances that are generally less than the original eigenvalues. This expected reduction in variance is labeled "shrinkage."

Note that since the total variance in a correlation matrix is constant (i.e., always equal to p), such overestimation of population eigenvalues will occur only in the early eigenvalues (i.e., the first few that are extracted). Practically speaking, however, this is the area of interest since principal components analyses generally focus on the first few emergent dimensions.

The above anticipates two classes of definitions for shrinkage. First, how much do sample eigenvalues overestimate population eigenvalues? Second, how much will eigenvalues, developed in an original sample, shrink when the original weights (eigenvectors) are applied to a new sample? Both definitions are developed in the Appendix. The first definition is analogous to the multiple regression shrinkage formula developed by Wherry (1931); the second definition is analogous to formulas developed by Lord (1950) and Nicholson (1960). Given results in multiple regression (cf. Darlington, 1968), it is hypothesized that the second (cross-validation) definition of shrinkage is associated with greater eigenvalue loss.

Parallel Analysis

Parallel analysis (Horn, 1965) is a factor technique that is related to eigenvalue shrinkage. In the context of factor extraction (i.e., determining k), Horn suggested that factors be extracted only if their eigenvalues were greater than eigenvalues obtained from a "random" correlation matrix (i.e., based on a sample drawn from a population identity matrix). The underlying assumption is that, because of the freedom to capitalize on chance correlations in the sample matrix, the first few eigenvalues are inflated estimates of the corresponding population eigenvalues.

Rather than generate a random correlation matrix each time a parallel analysis is to be conducted, Montanelli and Humphreys (1976) have empirically derived equations for the purpose of estimating the

expected value of random eigenvalues (from identity populations). Based on a degrees of freedom rationale, they fit a separate equation for eigenvalues of different ordinal positions ($i = 1, \ldots, p$) to predict eigenvalues (ℓ_i) generated from random samples. The general form of their equations is

$$\log(\ell_i) = a_i + b_i \log(N-1) + c_i \log[p(p-1)/2 - p(i-1)] \quad . \tag{1}$$

They generated equations for 21 levels of p (ranging from 6 to 90) and 4 levels of N (ranging from 25 to 1,533). When using these results, factors are to be retained for further analysis if their respective sample eigenvalues are greater than those predicted by Equation 1. This technique has been implemented in a variety of research contexts, including the analysis of binary items (Green, 1983) and the theory of performance ratings (Harvey, 1982; Hulin, 1982).

Shrinkage Formulas

In multiple regression literature, equations for predicting shrinkage typically use sample information, such as the sample multiple correlation, sample size, and the number of independent variables (see Cattin, 1980, for a review of these formulas). These equations can be used as substitutes for half-sampling or other cross-validation schemes (Murphy, 1983).

Regarding estimates of eigenvalue shrinkage, Equation 1 provides a first estimate of the degree to which the magnitude of a sample eigenvalue, ℓ_i, is due to the sample-based maximization criterion of principal components. That is, parallel analysis provides an initial answer to the question, "How much will eigenvalues shrink in cross-validation?" However, as noted above, previous efforts were motivated by factor *selection,* not shrinkage. Thus, several crucial elements are still lacking in Equation 1.

First, the degree of shrinkage may be a function of the magnitude of ℓ_i and not just of the ordinal position. This is analogous to the fact that shrinkage in multiple correlations is dependent on the sample value of R^2. Thus, Equation 1 should be amended to include ℓ_i as a predictor of λ_i.

Second, the equations of Montanelli and Humphreys (1976) are based on samples from identity populations (mutually uncorrelated variables). It is demonstrated below that shrinkage of eigenvalues depends on the underlying population covariance structure. Thus, the present analysis varied this structural aspect, added ℓ_i as a predictor, and generated equations useful for estimating eigenvalue shrinkage in applications of factor analyses.

Determinants of Eigenvalue Shrinkage

An analytic solution for eigenvalue shrinkage for the special case of $p = 2$ is given in the Appendix. It is obvious that this limited case has few direct applications, given that the general goal of principal components analysis is a systematic reduction of a relatively large variable space. Nevertheless, the derivation is of heuristic value. Specifically, shrinkage in factor analysis is hypothesized to be a function of (1) the sample size, N, (2) the number of variables factored, p, (3) the population eigenstructure, (4) the two definitions of shrinkage (see Appendix), (5) the ordinal position of the principal component, i, and (6) use of 1.0 or communalities in the main diagonal of the sample correlation matrix.

Sample and Predictor Size

From the results in the Appendix, it is hypothesized that eigenvalue shrinkage is greater when p is large and N is small. This is consistent with Montanelli and Humphreys' (1976) results and analogous to multiple regression results (cf. Cattin, 1980).

Eigenstructure

Shrinkage should also be a function of the population eigenstructure. That is, if the first population eigenvalue is relatively large, then sample estimates should have proportionately less shrinkage. Following a procedure by Cliff (1970), correlation matrices in the current study were varied according to their pattern of eigenvalues. The three patterns were identity matrices, slow-descent eigenstructures, and steep-descent eigenstructures. For identity population matrices all intercorrelations are zero, and hence, all eigenvalues equal one. For steep-descent eigenstructure populations, the eigenvalues displayed large changes in magnitude from one ordinal position to the next. The slow-descent eigenstructure populations represented an averaging of the identity and steep-descent populations. Note that the identity matrix was included in the study for reasons of completeness and for generating slow-descent matrices. In practice, application of Bartlett's (1954) test for sphericity would usually preclude a factor analysis of sample correlation matrices drawn from identity populations.

Ordinal Position

Assuming that shrinkage is a result of the freedom of sample estimates to capitalize on chance anomalies in the sample, it is hypothesized that the first eigenvalue is associated with greatest shrinkage. Subsequent eigenvalues should exhibit successively smaller shrinkage. This hypothesis is also consistent with the empirical findings of Montanelli and Humphreys (1976).

Communality Estimation

As noted earlier, the unities in the main diagonal of a sample correlation matrix are often replaced by communalities (the principal factors approach). It is hypothesized that shrinkage is increased by this principal factors approach because an additional set of parameters (communalities) is introduced into the process, with no change in sample size.

Method

A monte carlo simulation was used to examine the above hypotheses. The general method was (1) to construct population matrices with properties corresponding to the levels of the independent variables, (2) to generate sample correlation matrices from these populations and obtain their eigenvalue decomposition, and (3) to cross-validate the sample results to gain data corresponding to the two types of shrinkage.

Experimental Design

The independent variables were arranged to yield a split plot factorial design with between replications factors of: (1) the type of eigenstructure of the population correlation matrix, (2) the number of variables factored, p, (3) the p/N ratio that, jointly with p, sets the sample size, N, and (4) the specific population matrix that was nested within the eigenstructure factor. The within replications, or repeated measures, factors were (1) the type of shrinkage, (2) the use of communality estimates versus the use of unities in the diagonal of the factored correlation matrices (i.e., principal factors analysis vs. principal components analysis), and (3) the ordinal position of the eigenvalues.

Three types of population eigenstructures were used in the study. They were: (1) identity correlation (and thus equal eigenvalue) matrices, (2) eigenvalue matrices in which the eigenvalues showed relatively

large sequential reductions in magnitude (steep-descent structure), and (3) eigenvalue matrices in which the eigenvalues had magnitudes halfway between the identity and the steep-descent eigenstructures (slow-descent structure). The rules used to generate these eigenvalues are given below.

The three levels of the number of variables, p, were 10, 20, and 30. These levels were chosen to be representative of many factor analytic studies while also being small enough to allow rapid eigenvalue decomposition of the sample correlation matrices, thus allowing a large number of samples to be drawn.

The four levels of the p/N ratio used in the study were: ½, ¼, ⅛, and 1/20. Given the level of p, these ratios were obtained by setting the sample size, N, to the appropriate level. The ratios were representative of the majority of factor analytic studies appearing in the literature (cf. Bolton & Hinman, 1973; Crawford & Koopman, 1973).

For each combination of the level of p and the population matrix type other than the identity, five population matrices were generated and used in obtaining the data. This was done to insure that cell means were not dependent on one particular population matrix. Ten sample correlation matrices were chosen from each distinct population, resulting in 50 replications per cell.

Two types of shrinkage were evaluated. Analogous to regression procedures, Wherry-type shrinkage was defined as the difference between a sample eigenvalue and its corresponding population eigenvalue. Lord/Nicholson-type shrinkage was defined as the difference between the sample eigenvalue and the "pseudo-eigenvalue" obtained by applying the sample eigenvector to the population matrix (i.e., computing the variance of the sample eigenvector using population intercorrelations). This second definition of shrinkage corresponds to the concern with cross-validation of eigenvalues.[1]

Within replications, eigenvalues of the correlation matrices and of communality-reduced matrices were obtained and cross-validated. Communalities were estimated with squared multiple correlations.

An additional set of data with p equal to 40 was generated for use in the curve-fitting data analysis. To keep computer time and costs within reason when $p = 40$, only one population correlation matrix was generated per cell (with 10 replications per cell). The four levels of the p/N ratio discussed above were used.

Construction of the Population Correlation Matrices

The population matrices were constructed following a method described by Dempster, Schatzoff, and Wermurth (1977). To use the method, a proposed or target eigenvalue matrix is defined. For the type of population with steep-descent eigenvalues, the equation

$$\lambda_i = ke^{-i/2}, \quad k = p(e^{1/2} - 1)/(1 - e^{-p/2}) \quad , \tag{2}$$

was used to generate the set $\{\lambda_i, i = 1, \ldots, p\}$. In this equation, k is a normalizing constant setting the trace of the eigenvalue matrix equal to p.

The target eigenvalues for the slow-descent eigenvalues were obtained by taking the mean of the ith eigenvalue defined by Equation 2 and the value 1.0. For the case of $p = 40$, Equation 2 was altered such that the exponent was $-i/3$, and k again normalized the trace equal to p. This was done so that all target eigenvalues were nonzero within computational rounding error. Exemplary target eigenvalues for the steep- and slow-descent eigenstructures, when $p = 10$, are presented in Table 1.

---

[1]Actually, application of the sample eigenvector to the population correlation matrix is not mathematically equivalent to the expected value obtained when applying sample vectors to new samples. That is, there are two distinct operationalizations of Lord/Nicholson-type shrinkage. Indeed, Rozeboom (1978) has indicated that, in multiple regression contexts, these two operationalizations are based on slightly different assumptions. However, this study's computer simulations demonstrated that these two measures of shrinkage were always within 10^{-2} of each other. Hence, only one measure of Lord/Nicholson shrinkage (sample weights applied to the population) is reported here.

Table 1
Exemplary Target Eigenvalues (p = 10)

Ordinal Position(i)	Matrix Type		
	Identity	Steep-Descent	Slow-Descent
1	1.000	3.961	2.481
2	1.000	2.403	1.701
3	1.000	1.457	1.229
4	1.000	.884	.942
5	1.000	.536	.768
6	1.000	.325	.663
7	1.000	.197	.599
8	1.000	.120	.560
9	1.000	.073	.536
10	1.000	.044	.522

The second step of the Dempster et al. procedure is the construction of a $p \times p$ matrix whose elements are random numbers. This matrix is orthonormalized by columns with the Gram-Schmidt procedure (cf. Green & Carroll, 1976). Call the orthonormalized matrix $\mathbb{G}$ and the diagonal matrix of target eigenvalues $\mathbb{A}$. Then the matrix $\mathbb{\Sigma} = \mathbb{GAG}'$ is a covariance matrix with eigenvalues corresponding to the diagonal elements of $\mathbb{A}$. The population correlation matrix is obtained by standardizing $\mathbb{\Sigma}$. The resulting correlation matrix will not, in general, have the exact eigenstructure reflected in $\mathbb{A}$, but is usually sufficiently close (Dempster et al., 1977). In this manner, five population correlation matrices were generated for each level of p crossed with the steep- and slow-descent eigenstructure types. Following a procedure given by Montanelli (1975), 10 sample matrices were drawn from each of these populations.

Generation of Shrinkage Data

For Wherry-type shrinkage, sample eigenvalues were calculated and compared to population eigenvalues (i.e., obtained from *population* matrices). For Lord/Nicholson-type shrinkage, *sample* eigenvectors were applied to the population. The resultant cross-validated eigenvalues were then compared to the sample eigenvalues. These calculations were performed twice: once with principal components analysis and once with principal factors analysis.

Results

Cell Means

Mean shrinkage data corresponding to the first two ordinal positions (i.e., first two sample eigenvalues) are presented in Tables 2, 3, and 4. These means are collapsed across the five population correlation matrices nested within each type of eigenstructure. They are also collapsed over the replications factor. Thus, each mean is based on 50 shrinkage estimates. Table 2 presents shrinkage results when the population variables are uncorrelated (identity matrix). Tables 3 and 4 present results from slow-descent and steep-descent eigenstructure populations, respectively.

For example, suppose a principal components analysis (PCA) is conducted with $p = 20$ and $N = 160$, and a sample is drawn from an identity population matrix. From Table 2, it can be seen that the first sample eigenvalue will overestimate the population eigenvalue by .652 (Wherry shrinkage, PCA,

Table 2
Mean Shrinkage Values for Principal Components Analysis (PCA) and Principal Factors Analysis (PF) for Samples Drawn From Identity Population Matrices

p and p/N ratio	Wherry Shrinkage				Lord/Nicholson Shrinkage			
	Ordinal Position 1		Ordinal Position 2		Ordinal Position 1		Ordinal Position 2	
	PCA	PF	PCA	PF	PCA	PF	PCA	PF
p = 10								
1/2	1.122	1.536	.718	1.104	1.051	1.120	.714	.741
1/4	.848	1.097	.557	.781	.859	.915	.549	.553
1/8	.585	.718	.393	.510	.575	.605	.411	.422
1/20	.341	.394	.238	.285	.327	.336	.241	.246
p = 20								
1/2	1.298	1.704	1.019	1.413	1.287	1.323	1.023	1.034
1/4	.935	1.176	.746	.977	.952	.989	.739	.777
1/8	.652	.778	.528	.648	.627	.641	.516	.528
1/20	.417	.472	.338	.390	.409	.418	.338	.347
p = 30								
1/2	1.398	1.796	1.170	1.560	1.400	1.443	1.147	1.180
1/4	.983	1.213	.836	1.061	1.029	1.050	.855	.878
1/8	.709	.838	.595	.720	.706	.726	.612	.630
1/20	.428	.480	.367	.417	.451	.457	.374	.376

Table 3
Mean Shrinkage Values for Principal Components Analysis (PCA) and Principal Factors Analysis (PF) for Samples Drawn From Slow-Descent Population Matrices

p and p/N ratio	Wherry Shrinkage				Lord/Nicholson Shrinkage			
	Ordinal Position 1		Ordinal Position 2		Ordinal Position 1		Ordinal Position 2	
	PCA	PF	PCA	PF	PCA	PF	PCA	PF
p = 10								
1/2	.381	.664	.277	.566	.771	.833	.493	.542
1/4	.214	.365	.087	.243	.528	.567	.226	.225
1/8	.081	.167	.051	.119	.322	.360	.139	.131
1/20	.053	.087	-.012	.025	.130	.145	.051	.066
p = 20								
1/2	-.021	.198	.140	.356	.550	.579	.512	.535
1/4	.016	.138	.028	.144	.398	.416	.257	.262
1/8	.066	.131	.013	.078	.312	.320	.124	.128
1/20	-.007	.036	.018	.049	.075	.079	.037	.041
p = 30								
1/2	-.040	.169	.051	.247	.774	.796	.572	.586
1/4	-.194	-.081	.049	.154	.316	.334	.271	.271
1/8	-.070	-.008	.041	.100	.157	.170	.146	.146
1/20	-.037	-.013	.030	.009	.096	.101	.070	.068

$p/N = 20/160 = 1/8$). That is, relative to a population eigenvalue of 1.0, the first sample eigenvalue will be about 1.652. A quick perusal of Tables 2, 3, and 4 indicates that this shrinkage drops dramatically as p/N ratios decrease and/or eigenstructures are nonidentity.

Analysis of Variance

The shrinkage data in Tables 2, 3, and 4 were analyzed with a split-plot analysis of variance. All main effects, other than those for the number of variables (p) and for population matrices nested within eigenstructure type, were statistically significant ($p < .01$). Many of the interactions were also statistically significant.[2] Given the number of replications, and consequent large degrees of freedom, each F-test was subject to high statistical power. Thus, omega-squared estimates were computed to allow practical evaluations of the relative contributions of independent variables. These estimates are given in Table 5. In interpreting Table 5, note that T (the type of population structure) includes identity, slow-descent, and steep-descent matrices. Also, A (the analytic method) was either principal components or principal factors analysis and S (the shrinkage type) implied either a Wherry definition or a Lord/Nicholson definition of shrinkage.

Of all nonerror sources of variance, the type of population eigenstructure accounted for the largest proportion of variance. In fact, this factor accounted for more than half of the systematic variance and 18.63% of the total variance in eigenvalue shrinkage. The effect of the p/N ratio accounted for an additional 7.92% of the total variance. The interaction of population eigenstructure type by the p/N ratio accounted for 1.99% of the total variance, such that high p/N ratios elicited particularly large shrinkage when the population correlations were zero (identity matrix). Thus, the two independent variables of the p/N ratio and type of population eigenstructure accounted for more than 75% of the systematic variance in this data set when the two main effects and their interaction were included.

Curve Fitting (Shrinkage Formulas)

Curve fitting included simulation data from all ordinal positions. Regression analyses were conducted with either population eigenvalues (Wherry shrinkage) or cross-validated eigenvalues (Lord/Nicholson shrinkage) as the dependent variables. Several modifications to Equation 1 were dictated by a theoretical analysis.

First, as initial estimates of population parameters, the *sample* eigenvalues (ℓ_i) were included in the regression. This is information available to a researcher and, as demonstrated above, certainly affects shrinkage. Second, the last term in Equation 1 is calculable only when the ordinal position, i, is less than $p/2$ (otherwise, a logarithm of a negative number is implied). In order to identify all ordinal positions, the linear (not log) factor, $(p/2) - i$, was used. Remaining terms in the equation were generated by interacting this ordinal position factor with Montanelli and Humphreys' (1976) factors of variable size, p, and sample size, N.

Thus, shrinkage equations had the general form:

$$\hat{\lambda}_i = b_0 + b_1(\ell_i) + b_2(p/2 - i) + b_3(N)(p/2 - i) + b_4(p)(p/2 - i) \quad . \tag{3}$$

The empirically derived values of b_i are given in Table 6. For example, the best (least squares) prediction of the ith population eigenvalue, when principal components are used, is

---

[2]A complete analysis of variance is available from the authors.

Table 4
Mean Shrinkage Values for Principal Components Analysis (PCA) and Principal Factors Analysis (PF) for Samples Drawn From Steep-Descent Population Matrices

p and p/N ratio	Wherry Shrinkage				Lord/Nicholson Shrinkage			
	Ordinal Position 1		Ordinal Position 2		Ordinal Position 1		Ordinal Position 2	
	PCA	PF	PCA	PF	PCA	PF	PCA	PF
p = 10								
1/2	.170	.217	.099	.156	.578	.591	.315	.323
1/4	.023	.041	.071	.093	.378	.386	.020	.022
1/8	.109	.117	-.024	-.014	.215	.213	.057	.045
1/20	-.031	-.027	.047	.051	-.032	-.022	.115	.107
p = 20								
1/2	.154	.156	.034	.035	.894	.895	.104	.104
1/4	.103	.104	.037	.037	.575	.575	.024	.024
1/8	.001	.001	.056	.056	.070	.070	.110	.109
1/20	-.040	-.040	.006	.006	.013	.013	.006	.006
p = 30								
1/2	.105	.105	.141	.141	.504	.504	.287	.288
1/4	-.150	-.150	.125	.125	.062	.061	.312	.314
1/8	.026	.026	.073	.073	.042	.042	.254	.253
1/20	-.010	-.010	-.016	-.016	-.020	-.020	.036	.036

Table 5
Omega Squared Estimates for Eigenvalue Shrinkage of the First Two Principal Components

Source	ω^2	Source	ω^2
P (number of variables)	.00	PS	.00
Q (p to N ratio)	7.92	QS	.17
T (population type)	18.63	TS	.64
A (analytic method)	.34	AS	.12
S (shrinkage type)	.42	PI	.24
I (ordinal position)	.63	QI	.25
PQ	.00	TI	.16
PT	.86	AI	.00
QT	1.99	SI	.00
PA	.00	Higher Interactions[a]	1.97
QA	.14		
TA	.18	Residual[b]	64.26

[a]No higher order interactions had an individual contribution of more than .60.

[b]The residual ω^2 is the sum of all sources which include the replications factor.

Table 6
Regression Weights for Predicting Eigenvalues from Equation 3

Type of Analysis and Type of Shrinkage	b_o	b_1	b_2	b_3	b_4	Multiple R^2
Principal Components						
Wherry	-.04065	1.03350	-.05222	.00003	.00076	.973
Lord/Nicholson	-.12240	1.03788	-.05275	.00002	.00078	.969
Principal Factors						
Wherry	-.01543	1.00505	-.06019	.00004	.00087	.957
Lord/Nicholson	-.01070	1.00084	-.05958	.00004	.00086	.959

$$\hat{\lambda}_i = -.04065 + 1.03350(\ell_i) - .05222(p/2-i) + .00003(N)(p/2-i) + .00076(p)(p/2-i) \quad . \quad (4)$$

This equation fits the simulation data extremely well ($R^2 = .973$).

From Equation 3 an estimate of the magnitude of the shrinkage would be calculated as $(\hat{\lambda}_i - \ell_i)$. This value could be computed for each desired component (or factor), or for a particular set of dimensions (say, the first 3 principal components).

Discussion

Independent Variables

The analysis of variance indicated that the type of eigenstructure and the *p/N* ratio had a substantial impact on the magnitude of eigenvalue shrinkage. Perhaps the most important implication of these results is that the population eigenstructure has a far greater effect on shrinkage than do the more commonly examined variables such as sample size and the number of variables factored. The results also confirmed the hypothesis that the magnitude of eigenvalue shrinkage may be substantially less than is implied by previous parallel analysis equations—particularly when the variables to be factored are highly intercorrelated.

The effect of *p/N* ratios on eigenvalue shrinkage is consistent with previous factor analytic research (cf. Montanelli & Humphreys, 1976) as well as with the multiple regression literature. In addition, the effect of population structure on eigenvalue shrinkage is consistent with the findings that changes in the magnitude and pattern of population factor loadings will affect the stability of sample results (cf. Cliff & Pennell, 1967; Guadagnoli & Velicer, 1984).

In the context of variance statistics of the form $\sum_{i=1}^{k} \hat{\lambda}_i$, the above analysis assumed that the number of factors extracted, *k*, was a fixed known quantity. It would be of interest to investigate the interaction of factor selection rules with variance shrinkage. In general, following stepwise regression logic, greater shrinkage would be expected when the value of *k* is allowed to vary. Of course, the *type* of selection rule would specifically influence the value of *k* (cf. Hakstian, Rogers, & Cattell, 1982; Zwick & Velicer, 1982) and consequent shrinkage.

Several caveats to these findings are also in order. First, these results were based on the analysis of correlation, not covariance, matrices—because the majority of factor analyses have been conducted on correlation matrices. Therefore, these results (e.g., the generally small magnitudes of shrinkage, large overlap between the two types of Lord/Nicholson shrinkage, etc.) may be valid only for analyses of correlation matrices.

Second, the present concern has been with the estimation of eigenvalues and *not* with the interpretation

of factors (i.e., not the estimation of factor loadings). The effects of factor patterns and sample size (N) on factor loadings is well documented (e.g., Cliff & Hamberger, 1967; Cliff & Pennell, 1967), and excellent methods exist for estimating the stability of factor interpretation (e.g., Kleinknecht, Thorndike, McGlynn, & Harkavy, 1984). The fact that eigenvalues do not shrink may not necessarily guarantee that factor loadings will remain stable across samples.

Applications

Joint use of Equation 3 and Table 6 enables researchers to predict eigenvalue shrinkage. Because ℓ_i is included as an independent variable, results in Table 6 generalize across a variety of population eigenstructures. Furthermore, in contrast to Tables 2, 3, and 4, Equation 3 is applicable to all ordinal positions.

As an application, consider a factor analysis reported by Murphy, Martin, and Garcia (1982). They extracted two factors from a pool of 10 items ($p = 10$), with a sample size of 45. The analysis was conducted with communalities in the main diagonal (principal factors analysis) and resulted in $\ell_1 + \ell_2 = 5.34 + 1.51 = 6.85$. The total common variance for all 10 variables was 9.133. Thus, in the original sample, the two factors accounted for $(6.85/9.133) \times 100 = 75\%$ of the common variance. Using the weights in the last row of Table 6, application of Equation 3 yields $\hat{\lambda}_1 = 5.14$ and $\hat{\lambda}_2 = 1.35$. Thus, it is predicted that the cross-validated two factor solution would account for $(6.49/9.133) \times 100 = 71\%$ of the common variance—a shrinkage of 4%.

It should be noted that application of Equation 3 to other reported factor analyses revealed similar results. That is, even under relatively large p/N ratios (say, ½), total shrinkage was small (a few percentage points). This is in contrast to multiple regression analyses, when p/N ratios of 1/10 or greater elicit concern about shrinkage in squared multiple correlations (cf. Thorndike, 1978). Thus, there is room for optimism regarding shrinkage in principal components based factor analysis. Researchers are urged to quantify that optimism, or perhaps identify aberrant cases, through the joint use of Equation 3 and Table 6.

References

Armstrong, J. S. (1967). Derivation of theory by means of factor analysis, or Tom Swift and his electric factor analysis machine. *American Statistician, 21,* 17–21.

Bartlett, M. S. (1954). A note on the multiplying factors for various chi-squared approximations. *Journal of the Royal Statistical Society, Series B, 16,* 296–298.

Bolton, B., & Hinman, S. (1973). *Factor analytic studies: 1941–1970.* Fayetteville AK: Arkansas Rehabilitation Center, University of Arkansas, Arkansas Rehabilitation Service.

Cattin, P. (1980). Estimation of the predictive power of a regression model. *Journal of Applied Psychology, 65,* 407–414.

Cliff, N. (1970). The relation between sample and population characteristic vectors. *Psychometrika, 35,* 163–178.

Cliff, N., & Hamberger, C. (1967). The study of sampling errors in factor analysis by means of artificial experiments. *Psychological Bulletin, 68,* 430–445.

Cliff, N., & Pennell, R. (1967). The influence of communality, factor strength, and loading size on the sample characteristics of factor loadings. *Psychometrika, 32,* 309–326.

Crawford, C. B., & Koopman, P. (1973). A note on Horn's test for the number of factors in factor analysis. *Multivariate Behavioral Research, 8,* 117–125.

Darlington, R. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin, 69,* 161–182.

Dempster, A. P., Schatzoff, M., & Wermuth, N. (1977). A simulation of alternatives to ordinary least squares. *Journal of the American Statistical Association, 72,* 77–91.

Everett, J. E. (1983). Factor comparability as a means of determining the number of factors and their rotation. *Multivariate Behavioral Research, 18,* 197–218.

Green, P., & Carroll, J. D. (1976). *Mathematical tools for applied multivariate analysis.* New York: Academic Press.

Green, S. B. (1983). Identifiability of spurious factors using linear factor analysis with binary items. *Applied Psychological Measurement, 7,* 139–147.

Guadagnoli, E., & Velicer, W. (1984). *The relationship of sample size to the stability of component patterns: A simulation study*. Unpublished manuscript, University of Rhode Island, Department of Psychology, Kingston.

Hakstian, A. R., Rogers, W. T., & Cattell, R. B. (1982). The behavior of number-of-factors rules with simulated data. *Multivariate Behavioral Research, 17,* 193–219.

Harvey, R. (1982). The future of partial correlation as a means to reduce halo in performance ratings. *Journal of Applied Psychology, 67,* 171–176.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179–185.

Hulin, C. (1982). Some reflections on general performance dimensions and halo rating error. *Journal of Applied Psychology, 67,* 165–170.

Kleinknecht, R. A., Thorndike, R. M., McGlynn, F. D., & Harkavy, J. (1984). Factor analysis of the dental fear survey with cross-validation. *Journal of the American Dental Association, 108,* 59–61.

Lee, H. B., & Comrey, A. L. (1979). Distortions in a commonly used factor analytic procedure. *Multivariate Behavioral Research, 14,* 301–321.

Lord, F. M. (1950). *Efficiency of prediction when a regression equation from one sample is used in a new sample* (Research Bulletin RB50–40). Princeton NJ: Educational Testing Service.

Montanelli, R. G. (1975). A computer program to generate sample correlation and covariance matrices. *Educational and Psychological Measurement, 35,* 195–197.

Montanelli, R. G., & Humphreys, L. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonals: A monte carlo study. *Psychometrika, 41,* 314–348.

Mood, A. M., & Graybill, F. A. (1963). *Introduction to the theory of statistics*. New York: McGraw-Hill.

Murphy, K. (1983). Fooling yourself with cross-validation: Single sample designs. *Personnel Psychology, 36,* 111–118.

Murphy, K., Martin, C., & Garcia, M. (1982). Do behavioral observation scales measure observation? *Journal of Applied Psychology, 67,* 562–567.

Nicholson, G. (1960). Prediction in future samples. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp. 322–330). Stanford: Stanford University Press.

Rozeboom, W. (1978). Estimation of cross-validated multiple correlation: A clarification. *Psychological Bulletin, 85,* 1348–1351.

Thorndike, R. M. (1978). *Correlational procedures for research*. New York: Wiley.

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the multiple correlation coefficient. *Annals of Mathematical Statistics, 2,* 440–457.

Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research, 17,* 253–269.

Acknowledgment

The authors thank Lloyd Humphreys and an anonymous reviewer for their valuable comments on an earlier draft of this article.

Author's Address

Send requests for reprints or further information to Philip Bobko, Department of Psychology, Virginia Polytechnic Institute and State University, Blackburg VA 24061, U.S.A.

Appendix

An analytic solution to eigenvalue shrinkage is derived below for the restricted case when $p = 2$. The derivation has heuristic value.

Let ℓ_1 be the first eigenvalue of a sample correlation matrix drawn from a population correlation matrix with eigenvalue λ_1. Similarly, let r be the sample correlation coefficient between the two measured variables and let ρ be the respective population correlation. Also, let (a_1, a_2) represent sample weights for the first eigenvector. Then, ℓ_1 is the sample variance of $a_1 z_1 + a_2 z_2$ and,

$$E\{\ell_1\} = E\{a_1^2 + a_2^2 + 2a_1a_2r\} \quad . \tag{4}$$

Remember that the a_i are constrained such that $a_1^2 + a_2^2 = 1$. In addition, when $p = 2$, it is easily demonstrated that $a_1a_2 = \frac{1}{2}$ if r is positive and $a_1a_2 = -\frac{1}{2}$ if r is negative. Thus, letting $f(r)$ be the density function of r,

$$E\{\ell_1\} = 1.0 + \int_0^1 r \cdot f(r)\, dr - \int_{-1}^0 r \cdot f(r)\, dr \quad . \tag{5}$$

Let Wherry-type shrinkage be defined as the difference between the sample eigenvalue and the corresponding population eigenvalue (see Darlington, 1968, for the regression analogy). Again, if $p = 2$, it is easily demonstrated that $\lambda_1 = 1 + |\rho|$. Then,

$$\begin{aligned}\text{Shrinkage} &= E\{\ell_1\} - \lambda_1 \\ &= [\textstyle\int_0^1 r \cdot f(r)\, dr - \int_{-1}^0 r \cdot f(r)\, dr + 1.0] - [1.0 + |\rho|] \\ &= \textstyle\int_0^1 r \cdot f(r)\, dr - \int_{-1}^0 r \cdot f(r)\, dr - |\rho| \quad . \end{aligned} \tag{6}$$

Similarly, let Lord/Nicholson-type shrinkage be defined as the difference between the sample eigenvalue and the "pseudo-eigenvalue" obtained by cross-validating the sample eigenvector onto the population correlation matrix. Denoting the pseudo-eigenvalue as λ_1^*, note that $\lambda_1^* = 1.0 + 2a_1a_2\rho$ where a_1 and a_2 are the eigenvector entries obtained from the sample. Then,

$$\begin{aligned}\text{Shrinkage} &= E\{\ell_1\} - E\{\lambda_1^*\} \\ &= E\{\ell_1\} - 1.0 - \rho E\{2a_1a_2\} \\ &= \textstyle\int_0^1 r \cdot f(r)\, dr - \int_{-1}^0 r \cdot f(r)\, dr - \rho\, [\int_0^1 f(r)\, dr - \int_{-1}^0 f(r)\, dr] \quad . \end{aligned} \tag{7}$$

In order to facilitate comparisons, note that the first two integral terms in Equations 6 and 7 are identical. Then, it can be seen that the subtractive term for Wherry-type shrinkage is greater than the subtractive term for Lord/Nicholson-type shrinkage. This is because the bracketed terms in Equation 7 must sum to 1.0. Therefore, their difference (which multiplies ρ in Equation 7) must be less than 1.0. Hence, Wherry-type shrinkage is smaller.

In the extreme case when $p = 2$ and $\rho = 0$, further simplification is possible. That is, the last term of Equations 6 and 7 becomes zero and the two types of shrinkage are identical. Also, when $\rho = 0$, the density function of r is (Mood & Graybill, 1963, p. 357)

$$f(r) = \frac{\frac{n-3}{2}!\,(1 - r^2)^{(n-4)/2}}{\sqrt{\pi}\,\frac{n-4}{2}!} \quad , \tag{8}$$

where n is the sample size. Upon integrating, it follows that eigenvalue shrinkage is

$$\textstyle\int_0^1 r \cdot f(r)\, dr - \int_{-1}^0 r \cdot f(r)\, dr = \dfrac{\frac{n-3}{2}!}{\frac{n-2}{2}!\,\sqrt{\pi}} \tag{9}$$

The resultant value, and hence shrinkage in ℓ_1, is reduced as the sample size is increased.