PSYCHOMETRIKA—VOL. 60, NO. 3, 355–369
SEPTEMBER 1995

ASSESSING THE STABILITY OF PRINCIPAL COMPONENTS USING REGRESSION

ATANU R. SINHA

JOHN E. ANDERSON GRADUATE SCHOOL OF MANAGEMENT
UNIVERSITY OF CALIFORNIA, LOS ANGELES

BRUCE S. BUCHANAN

LEONARD N. STERN SCHOOL OF BUSINESS
NEW YORK UNIVERSITY, NEW YORK

This paper presents an analysis, based on simulation, of the stability of principal components. Stability is measured by the expectation of the absolute inner product of the sample principal component with the corresponding population component. A multiple regression model to predict stability is devised, calibrated, and tested using simulated Normal data. Results show that the model can provide useful predictions of individual principal component stability when working with correlation matrices. Further, the predictive validity of the model is tested against data simulated from three non-Normal distributions. The model predicted very well even when the data departed from normality, thus giving robustness to the proposed measure. Used in conjunction with other existing rules this measure will help the user in determining interpretability of principal components.

Key words: sampling stability, PCA, simulation, regression analysis.

1. Introduction

In principal components analysis (PCA), one must often decide how many components, m, are needed to provide an adequate description of the latent structure in the p manifest variables ($m < p$). Common approaches to this problem, such as looking for the "elbow" in the scree plot, or retaining all components with corresponding eigenvalues greater than 1, are ad hoc rules of thumb, whose primary justification, despite some attempts to put them on a more formal basis, is that they are intuitively appealing. Yet, there are other more formal approaches like Velicer's Partial Correlation Procedure where that many components are retained for which the average squared partial correlation is minimized. (For a review of these and other procedures see Jackson 1991[1]). Moreover, given that a principal component is kept, the next task often is to provide some form of interpretation for it; that is, to examine the values of its coefficients and to describe its relationship with the manifest variables. Indeed, one often finds in practice that a component is likely to be kept to the extent that its coefficients yield a plausible interpretation.

The processes of component retention and interpretation are complicated by the fact that the estimated coefficients have themselves some sampling variability. Thus, components obtained from PCA are likely to differ from one randomly drawn sample to another and may, in fact, possess very little stability. If one takes stability to be a

The authors would like to thank the four anonymous reviewers and the two editors for their valuable comments. Atanu R. Sinha gratefully acknowledges the research support received from the Marketing Studies Center, AGSM, UCLA. Send requests for reprints to Atanu R. Sinha, B418 Gold Hall, 110 Westwood Plaza, Los Angeles, CA 90095.

[1] We thank the fourth anonymous reviewer for this important reference.

0033-3123/95/0900-93149$00.75/0

prerequisite to useful interpretation, then it behooves the analyst to know whether or not the component he or she is attempting to interpret is stable. The focus in this paper is to provide the user with a simple measure by which to judge whether or not a component can be interpreted with confidence. Such a measure is to be used in conjunction with other existing rules.

Sampling theory does exist for some special cases of PCA, such as when the covariances of manifest variables form an identity matrix (Srivastava & Khatri, 1979, p. 86), and more generally for the largest and smallest eigenvalues (Srivastava & Khatri, 1979, p. 205). And a number of statistical tests for PCA have been described (e.g., Anderson, 1951, 1963; Girschik, 1936, 1939). But on the whole analytical results pertaining to the stability of principal components are limited by their requisite distributional assumptions or asymptotic nature.

Further, the analytical results are mostly developed for PCA from covariance matrices. According to Anderson (1963, p. 136) the asymptotic distribution theory of components extracted from correlation matrices is much more complicated than for covariance matrices; general results cannot be given in a simple form. What has been developed for some special cases is *extremely* complicated (Jackson, 1991, p. 99). Also, in situations where the original variables are in different units or the variances of the variables differ widely it is preferable to use the correlation matrix. As a matter of fact a large number of applications of PCA have employed the correlation matrix (Morrison, 1976, p. 268 and Jackson, p. 65).

The paucity of useful analytical results relating to PCA led to the development of a number of studies using simulation. Our study follows that tradition. The purpose of this study is to develop an approach for assessing the stability of principal components when working with correlation matrices. Such an approach, based on simulation and free of restrictive distributional assumptions can be applied to reasonably small samples. Once the stability of a component is known, the analyst can make a judgment as to whether or not it is stable enough to warrant an interpretation. Given this objective there is need for a model which the analyst can easily use. Regression modeling provides an answer to this need. We regressed on the variables which are most likely to impact the stability measure as defined by us.

The structure of this paper is as follows. In section 2 we review some analytical results on the stability of principal components, and draw some intuitive conclusions from them. The literature in this area is vast. To conserve space throughout this paper we reference only those articles that are relevant to our work. We refer the reader to the book by Jackson (1991) for a comprehensive review. In section 3 we describe the measure of component stability that will be used in our analysis. In section 4 we describe our simulation procedure and relate it to other studies which have used simulation or bootstrap methodology. In section 5 we develop a predictive model of component stability, then calibrate and test it using the results of the simulation procedures. In section 6 we demonstrate the usefulness of the model with the help of three examples. We conclude with some remarks concerning limitations and applications of the model.

2. Review of Analytical Results

We review a few relevant analytical results that help in drawing some conclusions regarding the important variables that go into the regression model. As stated above, for correlation matrices it is difficult to obtain results in a simple form. We decided, instead, to learn by studying results for covariance matrices. We did this solely for the purpose of identifying the important variables in order to build our measure, being

aware that there is no one-to-one correspondence between the components obtained from a correlation matrix and those obtained from a covariance matrix (Jackson, 1991).

Consider, as above, p manifest variables. Let the population mean vector be $\boldsymbol{\mu}$ and the covariance matrix be $\boldsymbol{\Sigma}$. Denote the eigenvalues of $\boldsymbol{\Sigma}$ by $\lambda_1 > \cdots > \lambda_p$ (we assume components are ordered by decreasing eigenvalue) and the corresponding eigenvectors by $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p$. Suppose we have n independent p-tuplet observations from the population just described, denoted $\mathbf{x}_1, \ldots, \mathbf{x}_n$. These observations will have sample mean vector $\mathbf{m}$ and sample covariance matrix $\mathbf{S}$. The p sample eigenvalues will then be $l_1 > \cdots > l_p$, and their corresponding normalized eigenvectors $\mathbf{a}_1, \ldots, \mathbf{a}_p$, so that $\mathbf{S} = l_1\mathbf{a}_1\mathbf{a}_1' + \cdots + l_p\mathbf{a}_p\mathbf{a}_p'$.

If these observations are drawn from a multivariate Normal population, and the p eigenvalues are distinct, then according to Morrison (1976) the asymptotic sampling properties of the coefficients of the i-th eigenvector are given by the following covariance matrix

$$\operatorname{cov}(\mathbf{a}_i - \boldsymbol{\alpha}_i) = \frac{\lambda_i}{n} \sum_{\substack{h=1 \\ h \neq i}}^{p} \frac{\lambda_h}{(\lambda_h - \lambda_i)^2} \boldsymbol{\alpha}_h \boldsymbol{\alpha}_h'. \tag{1}$$

This equation shows that the sampling variances of the coefficients of the sample eigenvector, and thus, the stability of the sample principal components, depend qualitatively on two factors. First, the sampling variances are inversely proportional to the sample size. Second, the coefficients of the i-th vector have smaller sampling variances to the extent that its corresponding eigenvalue is *distinct* from other eigenvalues. That is, to the extent that all other eigenvalues are either greater or less than λ_i, the sampling variances of the coefficients of $\mathbf{a}_i$ will be less. This is shown by the denominator of (1). As λ_h approaches λ_i, the denominator goes to zero and the cov $(\mathbf{a}_i - \boldsymbol{\alpha}_i)$ goes to infinity. These two factors—the sample size and the distinctness—are used in our model. The importance of distinctness in determining the stability of individual components is also borne out in the theoretical influence function developed for eigenvectors (see Pack, Jolliffe, & Morgan, 1988, Equation 2, p. 42).

Geometrically, the above result is obvious when we consider that any two successive principal components form the basis of a plane in the p-dimensional space (see Morrison, 1976, p. 275 for a geometrical interpretation with a figure). The projection of the sample points onto this plane will be more and more ellipsoidal (i.e., more elongated) to the extent that the two eigenvalues are more distinct from each other; and more and more circular to the extent that they are equal. If the projection is more ellipsoidal, then it is easy to determine the major and minor axes of the projection, and hence the eigenvectors corresponding to the larger and smaller eigenvalues. But as the projection becomes more circular, the major and minor axes become more difficult to determine and are, in the limit, indeterminate.

Krzanowski (1984) investigated the issue of component stability through the effect that ε-perturbation in eigenvalues have on the stability of principal components. The author analytically derived the maximum change in the coefficients of the principal components for a given ε change in the variance of the component. He found that "it is not the absolute size of the variance of any component which determines whether that component is stable or not, but rather its separation in terms of variance from the next component. Relatively isolated (early) components with large variance should therefore be fairly stable, but later components which all have similar variances will not be stable" (p. 562). Analytically he shows that the effect on the component coefficients $\boldsymbol{\alpha}_i$, of an ε-reduction in λ_i, is an inverse function of $\lambda_i - \lambda_{i+1}$.

Pack et al (1988) have studied the problem of influential observations in PCA. They demonstrated, through a case study, that the theoretical influence functions provided good estimates of the actual changes observed when individual observations are deleted from a principal components analysis. Green (1977) recognized that interpretation of parameters of multivariate models requires knowing how much the fit of the model to the data is impaired by changes in the parameters. His study focused on how sensitive principal components are to rounding and zeroing of coefficients.

3. A Measure of Principal Component Stability

Our approach starts from a correlation matrix. In devising a scalar measure to summarize the sampling properties of eigenvector coefficients, we did the following:

Let Ψ_i be the unit vector in p-dimensional space described by the population coefficients of α_i. Then let $\mathbf{Y}_i$ be a unit vector of the i-th principal component, described by the coefficients $\mathbf{a}_i$, from a sample of size n randomly drawn from the population. (Note that the population and sample eigenvectors α_i and $\mathbf{a}_i$ both have unit norm.) The stability of the sample i-th principal component is determined by the extent to which, on average, the directionality of $\mathbf{Y}_i$ corresponds to that of Ψ_i over repeated independent samples. A simple way to capture this correspondence is to project $\mathbf{Y}_i$ onto Ψ_i. This is done by taking the inner product of the two unit vectors $\mathbf{Y}_i$ and Ψ_i. Since we are dealing with unit vectors this inner product represents the cosine of the angle between the two vectors $\mathbf{Y}_i$ and Ψ_i. The choice of signs in PCA is arbitrary. Hence we take the expectation of the absolute values of the inner product to arrive at an average measure of stability (STB). The average is taken over independent samples of size n. This quantity can be written in terms of the eigenvector coefficients as

$$\text{STB} = E[|\mathbf{a}_i'\alpha_i|]. \tag{2}$$

This is our measure of principal component stability; namely, the expectation of the absolute value of the inner product between the sample vector and the population vector, where the expectation is taken over independent samples of size n.

4. Simulation

First we describe the principle underlying the simulation. Then we describe how the simulated data was created.

Consider a set of p manifest variables. From a particular "population" of N p-tuplet observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$, we draw samples of size n from the distribution $P[\mathbf{x} = \mathbf{x}_i] = 1/N$, for $i = 1, 2, \ldots, n$. In other words, we draw a sample of size n, with replacement, from the "population". We do this a total of B times to obtain B samples. In the simulation p and n are varied over a range of values. The choice of N as the "population" size and B as the number of samples is based on judgment.

We perform a PCA on the "population" correlation matrix to get a set of eigenvalues $(\lambda_1, \ldots, \lambda_p)$ and corresponding eigenvectors $(\alpha_1, \ldots, \alpha_p)$. We then perform PCA on correlation matrices from the randomly drawn samples (denoted by s), obtaining eigenvalues $(l_1^s, \ldots, l_p^s)$ and eigenvectors $(\mathbf{a}_1^s, \ldots, \mathbf{a}_p^s)$ from each $s \in B$. For, say, the i-th component, we calculate the set of B absolute inner products $|\mathbf{a}_i^{s'}\alpha_i|$ for all $s \in B$. The estimated stability of the i-th component is then the mean of these absolute inner products across all B samples

$$\text{STB}_i = \frac{1}{B}\sum_{s=1}^{B} |\mathbf{a}_i^{s'}\alpha_i| \tag{3}$$

There is a rich tradition of using simulation to investigate different aspects of principal components analysis. For the sake of brevity we shall mention only a few studies. Zwick and Velicer (1986) used simulation to compare the performance of the five methods used for determining the number of components to retain. These methods include Horn's parallel analysis, Velicer's minimum average partial correlation, Cattell's scree test, Bartlett's chi-square test, and Kaiser's eigenvalue greater than 1.0 rule. The authors systematically varied several factors including sample size, number of variables, number of components, and component saturation (defined as the magnitude of the correlation between the manifest variables and the components). It was found that under a variety of situations the parallel analysis and the minimum average partial correlation gave the best results.

Guadagnoli and Velicer (1988) used simulation to study the relation between sample size and sampling stability of the components. Stability was defined relative to the population pattern. Besides sample size they systematically varied number of variables, number of components, and component saturation. The authors concluded that sample size as a function of number of variables was not an important factor in determining stability. Component saturation and absolute sample sizes (defined as sample size *per se*, as distinct from sample size relative to the number of variables) were the most important factors, with number of variables per component ($= p/m$) being important as well. Their study provided credence to our use of sample size and number of variables in the regression model.

Some researchers have applied the bootstrap method to principal components analysis. Diaconis and Efron (1983) used the bootstrap to assess the variability of individual eigenvector coefficients, and Daudin, Duby and Trecourt (1988) studied the stability of principal component subspaces using a stability measure based on the bootstrap. But as far as we know this is the first study to assess the directional stability of individual eigenvectors in PCA.

In the context of factor analysis Lambert, Wildt and Durant (1991) used bootstrap to construct approximate confidence intervals for the factor loadings. Their purpose was broadly similar to ours—assessing sampling variability of loadings estimated by exploratory factor analysis. But there are some key differences. One, we are using our approach for principal components analysis. Secondly, we are proposing a different approach to assessing stability—a regression model. We believe our measure is very simple to implement and can be computed with a calculator.

Creating Simulated Data

To perform the modeling described in the next section, we first had to create a large number of simulated data sets to serve as "populations". These varied by their eigen structures and number of variables (p). The procedure used to create each of these "populations" is in the Appendix.

For the purpose of calibration of the measure we used Normal "populations". We realized that in order for the measure to have validity it was necessary to test the measure beyond settings of the Normal distribution. There are a variety of situations where the data may not conform to a Normal distribution. Thus, we tested for robustness against departures from Normality using three other distributions—uniform distribution, gamma distribution with shape parameter 1, and gamma distribution with shape parameter 2. These three between them cover a wide variety of distributions—on the one hand the uniform distribution, and on the other hand a right-tailed distribution with mode at zero, that is, gamma with shape = 1.

"Populations" were created for 11 different values of p (from 7 to 27 in increments of 2). For each of these 11 different values of p, a "population" was created with a

different eigenstructure (the p eigenvalues in each "population" were obtained by taking p draws from uniform random number generator and constraining their sum to be p, and the p eigenvectors were obtained as described in the Appendix).

Next, a sample size n was set. For a given n, 50 samples were created from each "population". For each principal component, i, from each sample, s ($s = 1, \ldots, 50$), an inner product was calculated with the corresponding component from the "population"; then its absolute value enumerated. These 50 absolute inner products were then averaged to obtain an estimated stability measure, as indicated in (3). This average stability measure is for a given n and p. As mentioned earlier we used 11 different eigenstructures corresponding to 11 values of p. We used 18 different values of n (from 30 to 200 in increments of 10). Altogether $11 \times 18 = 198$ combinations of p and n were used.

5. Modeling Component Stability

The discussion in section 2 suggests that a principal component will be more stable to the extent that (i) the sample size increases, and (ii) its corresponding eigenvalue is distinct from those of the preceding and succeeding components. Actually, we define two forms of distinctness. The predistinction (PREDST) of component i is measured with respect to component $i - 1$; it is given by $(l_{i-1} - l_i)/l_{i-1}$. Post-distinction (PSTDST) is measured with respect to the succeeding component, and is given by $(l_i - l_{i+1})/l_i$. The values of both distinction measures are bounded on the range (0, 1). By definition, the first component cannot have a predistinction and the last (or p-th) component cannot have a postdistinction. Thus for components 2 through $p - 1$ we used both PREDST and PSTDST in the regression model. However in PCA the first component is important and cannot be ignored. Hence, for the first component we developed a separate regression model using PSTDST[2]. The last component is rarely of practical relevance in PCA and was not considered a candidate for modeling.

In addition to these three factors, it is reasonable to expect that, other things being equal, a principal component will be less stable to the extent that the number of variables, p, increases because the number of eigenvector coefficients that must be estimated increases with p^2, while the size of the data matrix only increases with p. This intuition was borne out by the results we obtained as shown in Table 1. Table 1 shows how average stability increases with sample size n and decreases with number of components p. Thus for a given sample size, say, 30, the average stability decreases from 0.77 for $p = 7$ to 0.27 for $p = 27$. Similarly, for a given value of p, say, 19, the average stability increases from 0.34 for $n = 30$ to 0.67 for $n = 200$.

Estimation

The model was estimated on multivariate Normal simulated data. To quantify the relationship between these variables, viz., PREDST, PSTDST, n, and p, and the stability of principal components STB, we performed regression analysis. The stability measure was bounded (0, 1). So, to obtain a more tractable dependent variable in the regression, we employed a logit transformation. Because we expected the effects of sample size n and, of number of variables p to be proportionate rather than absolute, we took logarithms of these two variables.

We used two separate models—one, for estimating the stability of Components 2 through $p - 1$, and another, for estimating the stability of Component 1. This was necessary because the variable PREDST is defined for Components 2 through $p - 1$, and not defined for Component 1.

[2] We thank the first anonymous reviewer for this suggestion.

Table 1

Change in Stability (STB) with Sample Size n, and Number of Manifest Variables p

Sample Size n	Number of Manifest Variables p				
	7	13	19	23	27
30	0.77	0.48	0.34	0.28	0.27
60	0.90	0.58	0.45	0.39	0.39
100	0.93	0.62	0.54	0.46	0.48
150	0.96	0.72	0.63	0.53	0.50
200	0.98	0.65	0.67	0.64	0.60

Estimation of Model for Components 2 through p − 1

Across 198 combinations of p and n, a total of 2,970 stability measures were generated. The 2,970 simulated data points were then randomly divided such that the model was calibrated on 1,989 data points. The remaining 981 points formed the holdout sample. The regression model we obtained using data from the calibration sample is:

$$\text{Logit(STB)}^3 = -2.485 + 3.616\ \text{PREDST} + 2.732\ \text{PSTDST} + 0.784\ \log_e n - 0.735\ \log_e p, \quad (4)$$

where all variables were defined above. On the calibration sample, the model yielded an adjusted R^2 of 0.881. All coefficients were significantly different from 0 at the 0.001 level of significance, and all signs were as expected. It should be pointed out that one could try to fit other regression models to the simulated data. But the model we used with the logit transformation of the dependent variable and the logarithmic transformation of the independent variables n and p, has considerable intuitive appeal as explained above.

Estimation of Model for Component 1

Recall that for component one PREDST is not defined. Thus, as stated already, for modeling stability of component 1 we made use of only three variables—PSTDST, n and p. Across 198 combinations of p and n, a total of 198 stability measures were generated. Then the 198 simulated data points were randomly divided such that the model was calibrated on 127 data points. The remaining 71 data points formed the holdout sample. The regression model we obtained using data from the calibration sample is:

$$\text{Logit(STB)} = -2.709 + 6.579\ \text{PSTDST} + 0.996\ \log_e n - 0.647\ \log_e p, \quad (5)$$

where all variables were defined above. On the calibration sample, the model yielded an adjusted R^2 of 0.940. All coefficients were significantly different from 0 at the 0.001 level of significance, and all signs were as expected. Further analysis failed to detect the

[3] Logit(STB) = $V \Rightarrow$ STB = $\exp(V)/(1 + \exp(V))$.

Table 2

Correlation and Absolute Differences between STB and PSTB for Components 2 through p-1, for some Typical Distributions

Distributions	Correlation between STB and PSTB	Percentage of absolute differences within +/- 0.05	Percentage of absolute differences within +/- 0.10
Normal	0.931	49.5	80.2
Gamma(Shape=1)	0.899	32.9	67.1
Gamma(Shape=2)	0.912	44.4	73.3
Uniform	0.908	46.7	78.2

presence of outliers or deviations from the usual regression assumptions. The models fitted the data very well. Thus, we felt these models were appropriate for the purpose of modeling component stability.

Prediction

The models were tested for predictive validity using the data from the holdout samples. Predicted stability, PSTB were computed using the above models and compared with the measure of stability, STB using (3). The STBs are called observed stability.

Predictive Validity of Model for Components 2 through p − 1

PSTB was computed using (4) for each of the 981 data points. The result of the comparison of PSTB with STB is depicted in Table 2.

Using Normal Distribution. The correlation between PSTB and STB is 0.931. Additionally, the table of the absolute differences between PSTB and STB (row 2, Table 2) shows 49.5% of all predicted stability PSTB within +/−0.05, and 80.2% within +/−0.10 of the observed stability STB.

Using non-Normal distributions. The model in (4) was estimated using multivariate Normal simulated data. We tested for the model's robustness to non-Normality using three different cases—two right skewed multivariate distributions, namely, gamma distribution with shape parameter 1, and gamma distribution with shape parameter 2, and uniform distribution (Table 2). The rationale behind choice of these distributions is that between these and the Normal distribution we are able to cover a lot of distributions that one is faced with in dealing with real data. For each of these three non-Normal distributions we created "populations" using the procedure as described in section 4 for different eigenvalues and eigenvectors structure. Then values of STB were calculated using (3). These were our observed stability for each component. We computed the predicted stability PSTB for each component using (4). The predicted stability was compared with observed stability using two measures—correlation and absolute differences. Table 2 shows that based on the correlation measure (Column 2)

Table 3

Correlation and Absolute Differences between STB and PSTB for Component 1, for some Typical Distributions

Distributions	Correlation between STB and PSTB	Percentage of absolute differences within +/- 0.05	Percentage of absolute differences within +/- 0.10
Normal	0.928	91.5	95.8
Gamma(Shape=1)	0.891	74.7	94.9
Gamma(Shape=2)	0.93	87.9	100.0
Uniform	0.929	84.8	99.0

there is hardly, if any, degradation in going to non-Normal data. The correlation measures are 0.899 for the gamma distribution with shape parameter 1, 0.912 for the gamma distribution with shape parameter 2 and, 0.908 for the uniform distribution. These compared very well with that of 0.931 for the Normal distribution. Now, based on the table of absolute differences (columns 3 and 4 of Table 2) there is some degradation from Normal distribution (80.2% of absolute differences lie within +/−0.10) to the gamma distribution with shape parameter 1 (67.1% of absolute differences lie within +/−0.10). For the gamma with shape = 2 and the uniform distribution, these values were 73.3% and 78.2% respectively. Hence moving away from the Normal distribution, the degradation for the two latter distributions were not significant. Certainly there are other distributions for which we have not tested for robustness. But the distributions for which we tested represent a reasonable variety and thus give us confidence in our results.

Predictive Validity of Model for Component 1

PSTB was computed using (5) for each of the 71 data points. The result of the comparison of PSTB with STB is depicted in Table 3.

Using Normal Distribution. The correlation between PSTB and STB is 0.928. Additionally, the table of the absolute differences between PSTB and STB (row 2, Table 3) shows 91.5% of all predicted stability PSTB within +/−0.05, and 95.8% within +/−0.10 of the observed stability STB.

Using non-Normal distributions. The model in (5) was estimated using multi-Normal simulated data. We tested for the model's robustness to deviations from Normality using the three different distributions as in the previous case. The rationale behind choice of these distributions has been described above. We calculated observed stability STB using (3) and the predicted stability PSTB for Component 1 using (5). Table 3 shows that based on the correlation measure (Column 2) there is hardly, if any, degradation in going to non-Normal data. The correlation measures are 0.891 for the gamma distribution with shape parameter 1, 0.93 for the gamma distribution with shape parameter 2 and, 0.929 for the uniform distribution. These compare very well with that

Table 4

Results for Example 1

Component i	Eigenvalue l_i	Variance Explained (%)	PSTB
1	2.12	23.56	0.66
2	2.06	22.89	0.67
3	1.24	13.78	0.78
4	1.10	12.22	0.73
5	0.66	7.33	0.73
6	0.65	7.22	0.49
7	0.56	6.22	0.71
8	0.38	4.22	0.85
9	0.23	2.56	N/A

of 0.928 for the Normal distribution. Now, based on the table of absolute differences (Columns 3 and 4 of Table 3) there is some degradation from Normal distribution (95.8% of absolute differences lie within +/−0.10) to the gamma distribution with shape parameter 1 (94.9% of absolute differences lie within +/−0.10). For the gamma with shape = 2 and the uniform distribution, these values are 100.0% and 99.0% respectively. (Note that moving away from Normality produced even better results; it is due to sampling error. There is no significant difference in the numbers.) Hence moving away from the Normal distribution, there is no degradation for the two latter distributions.

6. Examples

To demonstrate the use of this model we consider three examples—one simulated and two from real data.[4] In the following examples we computed PSTBs for Component 1 using (5), and for components 2 through $p - 1$ using (4). This can be done using a hand held calculator. (Recall that the last component is hardly useful to know about and has been dropped from consideration.)

Example 1

Consider the example summarized in Table 4. Here we simulated a data matrix of 9 manifest variables, with a sample size of 100. The correlation matrix for the data has eigenvalues as summarized in the second column. The questions an analyst might ask include: how many components to retain? and, how strong an interpretation to put on any retained component? The third column contains the percentage of total variance explained by the components. Examination of the second and third columns show that the "elbow," in so far as there is one, occurs between the third and fourth component (eigenvalues of 1.24 and 1.10, respectively). Further, the first four components account for 72.45% (23.56% + 22.89% + 13.78% + 12.22%) of the total variance. Thus, given

[4] We would like to thank Professor Margaret Campbell of UCLA for giving us access to these data.

that the eigenvalue of the fourth component is greater than 1, we might be inclined to retain the first four components.

Are all the four components interpretable? Results in the fourth column (Table 4), where we have calculated PSTB for Components 1 through 8, shows that the third component with PSTB 0.78 is definitely more stable than the first and the second component with PSTBs 0.66 and 0.67, respectively. The fourth component, with PSTB 0.73, is slightly more stable than the first and the second component. Thus, given that four components are to be retained, the third component can be interpreted with reasonable confidence; however, the first or the second component does not beckon intrepretation individually. Though these components explain roughly 24% and 23% of the total variance, respectively, they do not do so in a very stable way, hence the interpretation ascribed to them may vary randomly from sample to sample. The fourth component suffers from instability as well but not as much as the first two components. Thus the objective of data reduction is attained; however, overall, interpretation of components is suspect.

Note that the measure being proposed, namely, PSTB is used as supplemental to rules such as eigen value greater than 1. Hence, we won't go beyond retaining four components though some later components have very high stability (the 7th and the 8th components have PSTB of 0.71 and 0.85). Moreover, in this situation, going up to the 7th or 8th component would defeat the very purpose of data reduction since there are only 9 manifest variables.

Example 2

Here we have data obtained from 140 respondents on 22 psychological variables. The data were obtained as part of a dissertation research on advertising effects, in an experimental setting. Table 5 provides a summary of the findings. Both the "eigenvalue greater than one" and the "elbow" rule applied to this analysis would suggest the first four components be retained. They all have stability of at least 0.78. If we now look at the fifth component its eigenvalue 0.97 is only marginally less than 1. Also, the amount of variance explained by the fourth and the fifth component is very similar. But the stability measure PSTB is very low for the fifth component, namely, 0.40, and does not suggest inclusion of the fifth component in the analysis. Hence here the stability measure provides added confidence in retention and interpretation of the coefficients of the four components.

Example 3

These data were obtained from a different 140 respondents on a different set of 22 psychological variables as another part of the dissertation research mentioned above. The summary information is given in Table 6. The "elbow rule" suggests retaining the first four components, whereas the "eigenvalue greater than one rule" suggests retaining five components. Here the two rules do not lead to an unambiguous answer. The stability measure PSTB may be of help in such a situation. The stability measures in Table 6 suggest that the fourth and the fifth components have very low sampling stability, namely, with coefficients 0.48 and 0.54. Any interpretation of the corresponding coefficients is suspect. Hence, if interpretation is the objective it is not advisable to use more than 3 principal components. If the objective is data reduction then one would be hard pressed to justify using four components as per the scree rule. The eigenvalues of the fourth and the fifth component are so near to each other and component five has a higher stability than component four. It maybe advisable to use a five components solution for purposes of data reduction.

Table 5

Results for Example 2

Component i	Eigenvalue l_i	Variance Explained (%)	PSTB
1	7.27	33.04	0.98
2	2.83	12.85	0.87
3	2.25	10.23	0.78
4	1.09	4.94	0.78
5	0.97	4.4	0.40
6	0.94	4.26	0.45
7	0.74	3.38	0.49
8	0.72	3.27	0.38
9	0.64	2.93	0.43
10	0.59	2.68	0.46
11	0.50	2.26	0.44
12	0.48	2.20	0.36
13	0.45	2.06	0.40
14	0.41	1.84	0.41
15	0.38	1.76	0.40
16	0.35	1.60	0.49
17	0.28	1.27	0.48
18	0.27	1.21	0.34
19	0.26	1.18	0.37
20	0.24	1.08	0.49
21	0.19	0.86	0.61
22	0.15	0.70	N/A

7. Discussion

We have shown that a simple regression model can provide a reasonably accurate assessment of the stability of individual principal components. Though approximate, such a model avoids the need for relying on asymptotic results and on strict distributional assumptions. Of course, one could also perform a simulation, but such analyses are computer intensive and require access to the original data matrix. Our model can be applied after the fact on a calculator using only the sample eigenvalues, sample size, and number of variables. Thus used in conjunction with existing rules it may provide a useful benchmark for the analyst who does not have the resources or inclination to perform a simulation, or who does not have access to the original data matrix.

Table 6

Results for Example 3

Component i	Eigenvalue l_i	Variance Explained (%)	PSTB
1	9.22	41.89	0.99
2	2.71	12.31	0.95
3	1.43	6.52	0.78
4	1.20	5.47	0.48
5	1.11	5.06	0.54
6	0.79	3.60	0.57
7	0.75	3.43	0.43
8	0.64	2.92	0.47
9	0.59	2.67	0.41
10	0.54	2.44	0.44
11	0.47	2.12	0.53
12	0.38	1.74	0.51
13	0.35	1.60	0.45
14	0.30	1.35	0.45
15	0.28	1.29	0.37
16	0.27	1.22	0.34
17	0.26	1.18	0.50
18	0.19	0.86	0.66
19	0.15	0.69	0.56
20	0.13	0.57	0.45
21	0.12	0.57	0.46
22	0.10	0.46	N/A

This model is appropriate for PCA when working with correlation matrices. There are a variety of situations when one wants to work with covariance matrices. This model would have to be revised for use with PCA from covariance matrices.

Appendix

Creating the "populations"—a step by step exposition of the procedure used:

i. The number of variables, p, was set, and the eigenvalues $(\lambda_1, \ldots, \lambda_p)$ and the eigenvectors $(\alpha_1, \ldots, \alpha_p)$ of the "population" were specified. We wanted to have a wide variety of eigenstructures represented in our simulated data. That would provide strong internal validity to the development of our measure and

guard against biases that could arise if only some typical eigenstructures were used. For a given choice of p, the (p) eigenvalues ($\lambda_1, \ldots, \lambda_p$) were obtained by drawing randomly from a uniform distribution $U[0, p]$ and constraining their sum to be p.

ii. In order to obtain the eigenvectors use was made of several correlation matrices. For a given correlation matrix $\mathbf{R} = ((r_{i,j}))$ a principally partitioned submatrix of order p was drawn out randomly. Here is how it worked. Let the size of the correlation matrix be $q \times q$, where $q > p$. Draw a random integer between 1 and $q - p + 1$, call it u. Now extract the principally partitioned submatrix:

$$\begin{Vmatrix} r_{u,u} & \cdot & \cdot & r_{u,u+p-1} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ r_{u+p-1,u} & \cdot & \cdot & r_{u+p-1,u+p-1} \end{Vmatrix}.$$

A decomposition of this p-dimension correlation submatrix gives the eigenvectors ($\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p$).

iii. The eigenvalues ($\lambda_1, \ldots, \lambda_p$) from (*i*) and the eigenvectors ($\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p$) from (ii) were used to create the "population" covariance matrix $\boldsymbol{\Sigma}$ using matrix multiplication; that is, $\boldsymbol{\Sigma} = \lambda_1 \boldsymbol{\alpha}_1 \boldsymbol{\alpha}_1' + \cdots + \lambda_p \boldsymbol{\alpha}_p \boldsymbol{\alpha}_p'$.

iv. A Cholesky decomposition was performed on $\boldsymbol{\Sigma}$, yielding a factor matrix $\mathbf{C}$ such that $\boldsymbol{\Sigma} = \mathbf{CC}'$.

v. A univariate random Normal generator was used to create p independent standard Normal samples of size N. These were arranged into p-tuplets, as such constituting an N-by-p independent Normal sample, denoted $\mathbf{X}^{\mathrm{I}}$. N was chosen to be 2000.

vi. The independent Normal sample was converted to a "population" with the desired eigenstructure by post multiplication with the Cholesky factor of the covariance matrix for the desired population; that is, $\mathbf{X}^P = \mathbf{X}^{\mathrm{I}}\mathbf{C}$.

References

Anderson, T. W. (1951). The asymptotic distribution of certain characteristic roots and vectors. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (pp. 103–130). Berkeley and Los Angeles: University of California Press.

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics, 34*, 122–148.

Daudin, J. J., Duby, C., & Trecourt, P. (1988). Stability of principal component analysis studied by the bootstrap method. *Statistics, 19*, 241–258.

Diaconis, P., & Efron, B. (1983). Computer intensive methods in statistics. *Scientific American, 248*(5), 116–130.

Girschik, M. A. (1936). Principal components. *Journal of the American Statistical Association, 31*, 519–528.

Girschik, M. A. (1939). On the sampling theory of roots of determinantal equations. *Annals of Mathematical Statistics, 10*, 203–224.

Green, B. F. (1977). Parameter sensitivity in multivariate methods. *Journal of Multivariate Behavioral Research, 12*, 263–287.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*(2), 265–275.

Jackson, J. E. (1991). *A user's guide to principal components*. New York: Wiley.

Krzanowski, W. J. (1984). Sensitivity of principal components. *Journal of the Royal Statistical Society, Series B, 46*(3), 558–563.

Lambert, Z. V., Wildt, A. R. & Durand, R. M. (1991). Approximating confidence intervals for factor loadings. *Journal of Multivariate Behavioral Research, 26*(3), 421–434.

Morrison, D. F. (1976). *Multivariate Statistical Methods* (2nd ed.). New York: McGraw-Hill.

Pack, P., Jolliffe, I. T., & Morgan, B. J. T. (1988). Influential observations in principal components analysis: A case study. *Journal of Applied Statistics, 15*(1), 39–52.

Srivastava, M. S., & Khatri, C. G. (1979). *An Introduction to Multivariate Statistics*. New York: North Holland.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*(3), 432–442.

Manuscript received 1/20/92
Final version received 6/28/94