Jackson, A User's Guide to Principal Components
Wiley, 1991

the subject of simple structure and rotation and, in that context, Hawkins and Fatti (1984) suggested performing a Varimax rotation on the deleted pc's, $y_{k+1}, \ldots, y_p$, and testing the maximum absolute value of these. The possible advantage of this procedure would lie in its diagnostic properties.

More of an adjunct than an alternative might be the use of an Andrews plot (Andrews, 1972) of $\mathbf{x} - \hat{\mathbf{x}}$ for a set of data. Curves that tend to cluster should exhibit the same sorts of residual behavior and might enhance the interpretation of this phenomenon. Limits for these curves, if desired, could be based on the procedure due to Kulkarni and Paranjape (1984) except that they would be based on $\mathbf{S} - \mathbf{V}_k\mathbf{V}_k'$ rather than $\mathbf{S}$ itself. The Andrews technique will described in Section 18.6.

Before the distribution of Q had been derived, Gnanadesikan and Kettenring (1972) suggested making a gamma probability plot of Q, and this might still have merit in data analysis. This does require an estimate of the shape parameter of this distribution, which Gnanadesikan and Kettenring suggested obtaining on the basis of the unretained pc's. They also suggested:

1. Probability plots of the unretained pc's
2. Scatter plots of combinations of the unretained pc's
3. Plots of retained vs. unretained pc's

Related work on gamma plots was done by Wilk et al. (1962) and Gnanadesikan (1964). Other plotting techniques that may be useful by Andrews (1979) in connection with robustness and by Daniel & (1971), which, while designed for regression, have applications in PCA.

2.8 WHEN TO STOP?

2.8.1 Introduction

One of the greatest uses of PCA is its potential ability to adequately represent a p-variable data set in $k < p$ dimensions. The question becomes: "What is k?" Obviously, the larger k is, the better the fit of the PCA model; the smaller k is, the more simple the model will be. Somewhere, there is an optimal value of k; what is it?

To determine k, there must be a criterion for "optimality." This section will describe a large number of criteria that are in use. These criteria range all the way from significance tests to graphical procedures. Some of these criteria have serious weaknesses, but will be listed anyhow since they are commonly used.

2.8.2 When to Start?

Hotelling (1933) spoke of the "sand and cobblestone" theories of the mind with regard to test batteries. *Cobblestone* referred to the situation where a few pc's

fairly well characterized the data. *Sand*, on the other hand, meant that there were many low correlations and the resultant major characteristic roots were relatively small with some possibly indistinguishable. Sections 2.8.3 through 2.8.11 will be concerned with a number of stopping procedures. However, it would be convenient to have a procedure that would tell us at the beginning whether or not we should even begin to embark on the PCA process. There are a number of guides available.

From a statistical point of view, the best way to determine whether more than a sandpile exists is to perform the significance test on the equality of the characteristic roots. A form of this was given in (2.6.1) and will be expanded in Section 2.8.3. This requires the computation of all of the roots and while this is generally not a problem these days it might be desirable to have some quick measures available that do not involve the roots.

In Section 4.2.3, some bounds are given on the maximum root. These bounds are simple functions of the elements of the covariance or correlation matrix. A "one-number" descriptor was proposed by Gleason and Staelin (1975) for use with correlation matrices:

$$\varphi = \sqrt{\frac{\|\mathbf{R}\|^2 - p}{p(p-1)}} \tag{2.8.1}$$

where $\|\mathbf{R}\|^2 = \sum\sum r_{ij}^2 = \sum l_i^2$ with all summations running from 1 to p. They called this a measure of *redundancy* but we shall not use this term in order to avoid confusion with another use of the word in Section 12.6. If there is no correlation at all among the variables, $\|\mathbf{R}\|^2 = p$ and $\varphi = 0$. If all of the variables are perfectly correlated, $l_1 = p$ and $\varphi = 1$. Although this coefficient has the same range as a multiple correlation coefficient, it is not unusual to obtain values of less than .5 for situations that appear to have fairly strong structure. The distribution of this quantity is unknown and it would appear that experience will be the best guide as how to interpret it. However, this quantity could be useful in comparing two or more data sets.

If one is working with covariance matrices, (2.8.1) becomes

$$\varphi = \sqrt{\frac{\|\mathbf{S}\|^2 - \sum_{i=1}^{p}(s_i^2)^2}{\sum_{i=1}^{p}\sum_{j\neq i}^{p}(s_i s_j)^2}} \tag{2.8.2}$$

For the Ballistic Missile example, which involved a covariance matrix,

$$\varphi = \sqrt{\frac{115\,887 - 47\,895}{136\,238}} = .706$$

Another single number descriptor is the *index* of a matrix, $\sqrt{l_1/l_p}$ which, for this example, would be $\sqrt{335.34/16.41} = 4.52$. Unlike the Gleason–Staelin

statistic, this index does not have an upper bound and will generally increase with increasing p. The index of a singular covariance matrix is undefined.

A quantity that may be useful in connection with stopping rules is an extension of equation (1.5.4), the correlation of an original variable with a principal component. The multiple correlation of an original variable with all of the retained pc's is

$$R_{x_j(y_1 \cdots y_k)} = \sqrt{\sum_{i=1}^{k} \left(\frac{l_i u_{ji}^2}{s_j^2} \right)} = \sqrt{\sum_{i=1}^{k} \left(\frac{v_{ji}}{s_j} \right)^2} \tag{2.8.3}$$

2.8.3 Significance Tests

Before the advent of Bartlett's test, discussed in Section 2.6, the statistical stopping rules were, by necessity, ad hoc procedures such as the use of reliability coefficients (Hotelling, 1933) or tests for the rank of the covariance matrix using the generalized variance (Hoel, 1937). The various forms of Bartlett's test are large-sample procedures assuming that the sample size is large enough for the χ^2-approximation to hold. Section 4.4 will contain a number of small-sample procedures in detail. What is important here is to discuss the philosophy of significance tests in this context.

For significance tests in general, one should make the distinction between *statistical* and *physical* significance. Consider a simple t-test for a sample mean. If one obtained a sample of one million observations, nearly any departure of the sample mean from the hypothetical mean, no matter how small, would be judged significant. In more practical terms, if the precision with which the sample mean is estimated is small, a mean difference may be detected that is not considered important to the experimenter and in fact, the notion of "importance" or physical significance is the basis for sample size determinations. On the other hand, a true difference may exist between the true and hypothetical means but this difference may be obscured by a large variance in the sample. This analogy carries over to PCA. Bartlett's test is for the null hypothesis that the last $(p - k)$ characteristic roots are equal. It may be that this procedure will continue selecting pc's to be included in the model that explain very little of the total variance and, in addition, may not be readily interpretable. On the contrary, the precision of the estimates may be such that the PCA model may be oversimplified, that is, some pc's that one might expect to show up in the model are prevented from doing so by excessive variability. While either can happen, most practitioners would agree that the former is more prevalent, that is, Bartlett's test ends up retaining too many pc's.

Suppose one performs a significance test of this type on, say, a $p = 20$-variable problem and finds that only the last four roots are not significantly different so that 16 pc's should be retained. On reflection, this individual elects to delete 6 more and end up with 10 pc's in the model. Is this being hypocritical? Not necessarily. If the extra six discards were interpreted to be uncontrollable

inherent variability, there might be a case for dropping them. Some of the procedures to be described below will address some of these issues but they should be considered as adjuncts to, not substitutes for, significance tests. This would be particularly true in the case of the situation where high variability suppresses some pc's; do NOT include pc's in the model that do not belong there statistically. The answer to this dilemma would be to investigate why this situation may have occurred. A possibility, not to be overlooked, is the existence of one or more outliers; this is an excellent place to use the residual analysis techniques of the previous section or some of the sensitivity techniques given in Section 16.4.

2.8.4 Proportion of Trace Explained

Over the years, a very popular stopping rule has been one related to the *proportion* of the trace of the covariance matrix that is explained by the pc's in the model. Its popularity lies in the fact that it is easy to understand and administrate. Some computer packages use this rule; the user will specify some particular proportion (the default is commonly .95) and when that much of the trace is explained, the process is terminated.

This procedure is NOT recommended. There is nothing sacred about any fixed proportion. Suppose, for $p = 20$, that the last 15 roots are not significantly different so that only five pc's would be retained, and further suppose that these five pc's explain only 50% of the trace. Should one keep adding pc's related to these other roots until the magic figure is reached? Definitely not. Conversely, the example to be presented in Section 2.9 has over 95% explained by the first two pc's and yet one more should be included in the model.

Having said that, it must be admitted that there are occasions when PCA is used as an exploratory tool when very little is known about the population from which the sample is obtained. In these instances, the proportion of the trace explained may be of some use in developing provisional PCA models and we shall not be above employing this procedure occasionally.

2.8.5 Individual Residual Variances

What would seem to be a better stopping rule would be one based on the *amount* of the explained and unexplained variability. In this procedure, one determines in advance the amount of residual variability that one is willing to tolerate. Characteristic roots and vectors are obtained until the residual has been reduced to that quantity. (Again, this should be carried out *after* the significance test so that one does not include pc's that the test would not permit, even if the desired residual variance were unobtainable.)

In some situations, prior information with regard to inherent variability may be available that might serve as the desired target residual. This method has been employed for some of the examples in this book—in particular, the audiometric examples in Chapters 5 and 9. While it is best to specify the desired

residual variability for each variable separately, this may not be possible or practical. In this situation, an average residual could be specified that could then be expressed in form of the residual trace. For a chemical example, see Box et al. (1973).

2.8.6 The SCREE Test

This is a graphical technique. One plots all of the characteristic roots of the covariance matrix, the values of the roots themselves being the ordinate and the root number the abscissa. A typical SCREE plot is shown in Figure 2.2. [The name *SCREE* is due to Cattell (1966), scree being defined as the rubble at the bottom of a cliff, i.e. the retained roots are the cliff and the deleted ones are the rubble.] Note that the last few roots are much smaller than the rest and are nearly in a straight line. There is a break of sorts between the first three roots and these remaining five roots. (This break is sometimes called an *elbow*.) Cattell and Jaspers (1967) suggested using all of the pc's up to and including the first one in this latter group. In this example, the model would include four pc's.

This technique has become quite popular although there can be some problems with it. First, it is only a graphical substitute for a significance test. Second, the plot may not have a break in it; it could appear for instance as a fairly smooth curve. Alternatively, it may have more than one break. In this case, it is customary to use the first break in determining the retained pc's. However, Wernimont suggested that for some analytical chemistry problems in which he found this, the pc's before the first break represented the components describing the system under study and the second set, between the breaks, was

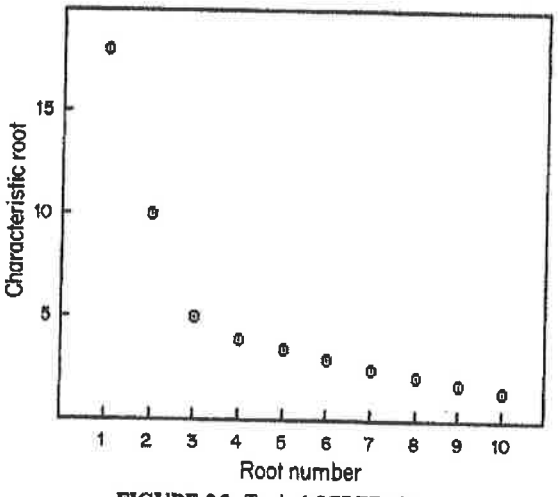

FIGURE 2.2. Typical SCREE plot.

made up of components representing instrumentation and/or other identifiable testing and measurement variability. An idealized representation of such a situation is shown in Figure 2.3. If these intermediate pc's are to be retained, their roots should be distinct. Cattell and Vogelmann (1977) also discussed procedures for multiple SCREEs.

Box et al. (1973) described a similar situation in which the third group of roots should have all been equal to zero because they represented linear constraints. However, due to rounding error, they were positive quantities, although much smaller than the roots representing inherent variability. [The opposite situation, $l_p = 0$ when $\lambda_p > 0$ is not likely. Yin et al. (1983) and Silverstein (1984) showed that l_p is bounded away from zero under fairly general conditions.]

A third situation that may occur is that the first few roots are so widely separated that it is difficult to plot them all without losing the detail about the rubble necessary to determine the break. This problem may be diminished by plotting the logs of the roots instead. (This is called an LEV, or log-eigenvalue, plot.)

In an attempt to take some of the guesswork out of these procedures, Horn (1965) suggested generating a random data set having the same number of variables and observations as the set being analyzed. These variables should be normally distributed but uncorrelated. A SCREE plot of these roots will generally approach a straight line over the entire range. The intersection of this line and the SCREE plot for the original data should indicate the point separating the retained and deleted pc's, the reasoning being that any roots for the real data that are above the line obtained for the random data represent roots that are larger than they would be by chance alone. (Just *larger*, not *significantly*

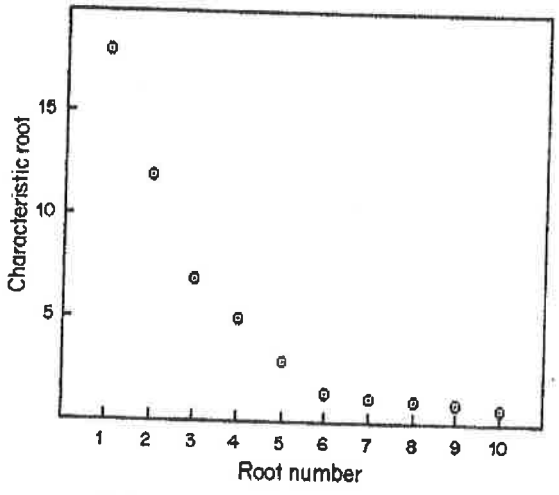

FIGURE 2.3. "Wernimont" SCREE plot.

larger.) Farmer (1971) carried out a similar study for LEV plots using both random and structured data. These procedures are sometimes called *parallel analysis.* Expressions for the expected values of these roots, using regression from these simulations, were obtained by Allen and Hubbard (1986), Lautenschlager et al. (1989), and Longman et al. (1989). Longman et al. also included the 95th percentile. Lautenschlager (1989) felt that interpolation might be superior to regression in many instances. He also included extensive tables. Lambert et al. (1990) used bootstrap techniques to study the sampling variability of the characteristic roots and their effect on the stopping rule.

2.8.7 The Broken Stick

A rather quick-and-dirty version of Horn's technique is the *broken stick* model (Jolliffe, 1986). This is based on the fact that if a line segment of unit length is randomly divided into p segments, the expected length of the kth-longest segment is

$$g_k = \frac{1}{p} \sum_{i=k}^{p} \left(\frac{1}{i} \right) \tag{2.8.4}$$

As long as the proportion explained by each l_k is larger than the corresponding g_k, retain the corresponding pc.

For example, the proportions for the case of four independent variables would be .521, .271, .146, and .062. The proportion explained by each pc for the Ballistics Missile example are .782, .112, .068, and .038. Only the first pc accounts for more than would be expected by chance alone.

2.8.8 The Average Root

An even quicker technique is to retain only those pc's whose roots exceed $\mathrm{Tr}(\mathbf{S})/p$, which is the size of the average root. The rationale here is that any pc that is deleted will have a root smaller than the average. In the next chapter, one of the topics to be discussed will be PCA starting with a correlation rather than covariance matrix. In this case the average root is equal to 1 and for this reason, the average root rule is widely used in the fields of psychology and education, among others, where correlation matrices are generally employed. This method, sometimes called the *Guttman–Kaiser* criterion, has been criticized as being too inflexible. In fairness to Guttman, he derived this criterion as a lower bound for the number of common factors in factor analysis (Section 17.2.2) along with two other criteria that were better; PCA was never mentioned (Guttman, 1954).

For the Ballistics Missile example, the average root is $429.11/4 = 107.28$, which is larger than all of the roots save the first and would imply that only one pc should be retained. Jolliffe (1972) contended that this cutoff was too high because it did not allow for sampling variability and, based on some

simulation studies, recommended using 70% of the average root. For the Ballistics Missile example, this would amount to 75.09, which would leave the conclusion unchanged.

2.8.9 Velicer's Partial Correlation Procedure

This procedure is based on the partial correlations among the original variables with one or more pc's removed (Velicer, 1976b). Let

$$\mathbf{S}_k = \mathbf{S} - \sum_{i=1}^{k} \mathbf{v}_i \mathbf{v}_i' \qquad k = 0, 1, \ldots, p-1 \tag{2.8.5}$$

and

$$\mathbf{R}_k = \mathbf{D}_s^{-1/2} \mathbf{S}_k \mathbf{D}_s^{-1/2} \tag{2.8.6}$$

where $\mathbf{D}_s$ is a diagonal matrix made up of the diagonal elements of $\mathbf{S}_k$. $\mathbf{R}_0$ is the original correlation matrix, $\mathbf{R}_1$ is the matrix of correlations among the residuals after one pc has been removed and so on.

Let

$$f_k = \sum\sum_{i \neq j} (r_{ij}^k)^2 / [p(p-1)] \tag{2.8.7}$$

where r_{ij}^k are the correlations in the matrix $\mathbf{R}_k$. f_k is the sum of squares of the partial correlations at stage k and, as a function of k, will normally have a minimum in the range $0 < k < p-1$ and the value of k for which this occurs will indicate the number of pc's to retain.

The logic behind Velicer's test is that as long as f_k is declining, the partial covariances are declining faster than the residual variances. This means that Velicer's procedure will terminate when, on the average, additional pc's would represent more variance than covariance. This represents a departure from the other stopping rules, which are concerned only with the characteristic roots themselves, and is somewhat of a compromise between these other procedures and the method of factor analysis that will be discussed in Chapter 17. It also precludes the inclusion of pc's that are primarily a function of one variable (Jolliffe, 1986).

To illustrate this method, consider the Ballistics Missile example with one pc retained:

$$\mathbf{S}_k = \mathbf{S} - \mathbf{v}_1 \mathbf{v}_1' = \begin{bmatrix} 29.25 & -6.77 & -5.01 & -16.25 \\ -6.77 & 18.81 & -7.01 & -11.28 \\ -5.01 & -7.01 & 13.92 & .48 \\ -16.25 & -11.28 & .48 & 31.79 \end{bmatrix}$$

whence

$$\mathbf{R}_1 = \begin{bmatrix} 1 & -.29 & -.25 & -.53 \\ -.29 & 1 & -.43 & -.46 \\ -.25 & -.43 & 1 & .02 \\ -.53 & -.46 & .02 & 1 \end{bmatrix}$$

and

$$f_1 = [(-.29)^2 + \cdots + (.02)^2]/[(4)(3)] = .14$$

The results are: $f_0 = .50, f_1 = .14, f_2 = .38$, and $f_3 = 1.00$. The minimum occurs at $k = 1$ so one pc would be retained. This is a typical result in that fewer pc's are retained than would be under Barlett's test. In Section 2.8.2 it was suggested that Bartlett's test might retain more pc's than was realistic, particularly for large n, and it was because of this that Velicer's procedure was devised.

Reddon (1985) carried out some simulations using data generated from populations having unit variances and zero covariances. In this case, f_0 should be less than f_1. The smaller the sample size relative to the number of variables, the larger the probability that this will not happen (Type I error). His example suggests a sample size of $n = 3p$ will be required to produce $\alpha = .05$ and of $n = 5p$ to reduce this below .01. This, of course, is only part of the story and Reddon also used this technique with some real data and found the results to be consistent with those already reported in the literature.

2.8.10 Cross-validation

Yet another suggestion for determining the optimum number of pc's is the cross-validation approach advocated by Wold (1976, 1978) and Eastment and Krzanowski (1982). This approach is recommended when the initial intention of a study is to construct a PCA model with which future sets of data will be evaluated. Cross-validation will be discussed in Section 16.3 but, briefly, the technique consists of randomly dividing the sample into g groups of n/g observations each. A PCA, using only the first pc, is performed on the entire sample save the first group. Predictions of these deleted observations are then obtained using (2.7.2) and a value of Q is obtained for each of these observations. The first group is then returned to the sample, the second group is deleted, and the procedure is repeated. This continues until all g groups have had their turn. The grand average of all the Q-values, divided by p, is called the *PRESS-statistic*. The entire procedure is repeated using a two-component model, a three-component model, and so on, from which additional PRESS-statistics are formed. The stopping rule is based on some comparison schemes of PRESS-statistics. As a stopping rule, this method requires the original data for its implementation. The others require only the covariance matrix.

2.8.11 Some Other Ad Hoc Procedures

There have been a number of other stopping rules proposed from time to time. Most of these are intuitive, many of them coming from the chemometricians, and their distributional properties are generally unknown. Among them are:

1. *Indicator function* (Malinowski, 1977).

$$\text{IND} = \sqrt{\frac{\sum_{j=k+1}^{p} l_j}{n(p-k)^3}} \tag{2.8.8}$$

   Obtain IND as a function of k and terminate when IND reaches a minimum. Droge and Van't Kloster (1987) and Droge et al. (1987) compared IND with some early cross-validation techniques. For comments on this comparison and a discussion of cross-validation, see Wold and Sjöström (1987).

2. *Imbedded error* (Malinowski, 1977).

$$\text{IE} = \sqrt{\frac{\sum_{j=k+1}^{p} l_j}{np(p-k)/k}} \tag{2.8.9}$$

   Obtain IE as a function of k and terminate when IE is a minimum.

3. The ratio l_i/l_{i+1} (Hirsh et al., 1987).
4. Number or percentage of residuals outside of univariate limits (Howery and Soroka, 1987).

Another practice is to stop at an arbitrary k because of some prior information or hypothesis. Newman and Sheth (1985) did this in a study of primary voting behavior. They had $n = 655$ observations on $p = 88$ variables. They stopped at $k = 7$ because that is what the model called for although there were more than seven roots (from a correlation matrix) that were greater than unity. The first seven only explained 32%.

With this last example, it is perhaps appropriate to quote Morrison (1976). "It has been the author's experience that if that proportion [he had previously recommended at least 75%] of the total variability cannot be explained in the first four or five components, it is usually fruitless to persist in extracting vectors even if the later characteristic roots are sufficiently distinct . . . ." If one knows in advance that the inherent variability is 60% or 70% and that there are a small number of real pc's, it is still permissible to obtain pc's up to that point but in most situations this is not known and, as Morrison suggests, obtaining a large number of pc's will invariably produce pc's that will be difficult to interpret at best and may well represent inherent variability.

2.8.12 Conclusion

In this section, a number of procedures have been presented for determining the optimum number of retained pc's. Some of these techniques have been included merely because they enjoy widespread use despite the fact that they have severe shortcomings. The Ballistic Missile example has been used as an example in this section solely for illustrative purposes because there were only four variables. Because the first pc, explaining 78% so dominates the others, some of these techniques retained only that one (Broken Stick, Average Root, and Velicer). Anderson's version of Bartlett's Test retained two as would have the SCREE plot except that one cannot really do much with only four variables. If one had used 95% of the trace of S (quite often used as a default on some computer programs), three pc's would have been retained. In the next section, a 14-variable example will be introduced that will produce a much more realistic comparison of these various methods.

In Section 2.4, it was shown that zero roots were indicative of linear constraints among the variables. For this reason, one should obtain *all* of the roots even though there is no intention of using all of the pc's in the model.

A number of studies have been carried out to compare these various procedures. Many of these were simulation studies. Among them was one by Krzanowski (1983) in which he found, for the four procedures he compared, that Bartlett's test retained the most pc's, followed by the 75% trace rule, the average root, and finally cross-validation, although the last two were similar. Krzanowski showed that cross-validation looked for gaps between the roots, thus providing an analytical version of the SCREE plot. Wold (1978) concluded that the cross-validation technique retained fewer pc's because it included a predictive procedure using the Q-statistic. Bauer (1981) concluded that the average root criterion was usually within ± 1 of the true dimensionality and never off by more than 3. The SCREE plot usually had one too many. Zwick and Velicer (1986) compared simulation results with the underlying models and concluded that, of the five procedures they compared, parallel analysis was most accurate followed by Velicer's partial correlation technique and the SCREE plot. The average root generally retained too many pc's. Bartlett's test, they felt, was erratic. The Zwick–Velicer paper refers to a number of other comparisons of these methods, many with conflicting results. This is probably the result of the different ranges of variations employed in the various experiments.

2.9 A PHOTOGRAPHIC FILM EXAMPLE

2.9.1 Introduction

The two examples used so far, the Chemical analyses and the Ballistics Missile tests, were included primarily for purpose of illustrating the mechanics of PCA since they involved two and four variables, respectively. Although a number of