PRINCIPAL COMPONENTS

Many research projects gather a large number of variables. For example, administering the MMPI (Minnesota Multiphasic Personality Inventory) gives one at least 566 variables because the MMPI has 566 items. For most research purposes, it is unwieldy and impractical to analyze and then subject the reader to an analysis of 566 variables. Thus, we would like some objective means of reducing the variables to a manageable number. One of the primary purposes of principal components analysis or PCA is to reduce the number of variables.

The logic of PCA is as follows. We would like to reduce the number of variables but preserve as much of the variability in the data as possible. To do this, let us create a new variable from the data that is a linear combination of the original variables. Let Z denote this new variable and let X_1, X_2, ... X_p denote the original variables. By making Z a linear combination of the original variables, we imply the mathematical formula

$$Z = a_1X_1 + a_2X_2 + ... a_pX_p \qquad (1.0)$$

where a_1, a_2, ... a_p are weights assigned to the original p variables. Note that Equation (1.0) is similar to the formula for a multiple regression equation--Z is analogous to the "dependent" variable, the a's are the b weights, and the X's are the predictor variables. In PCA, however, there is no intercept and there are no residuals.

To preserve as much variability as possible, we want to make the variance of Z as large as possible. Thus, the goal of PCA is to select values of the a's so that the variance of Z is as large as possible. But we have to be sensible about this. The variance of Z will begin approaching infinity as the individual a's go to either plus or minus infinity. To avoid this, PCA selects the a's subject to the constraint that $\sum_i a_i^2 = 1.0$, or the sum of the squared a's equals unity.

Technically, this is called "normalizing" a vector, the vector in this case being the vector of a's. Once the weights are calculated, then the variable Z is called the first principal component or first PC.

To see how the weights are identified, let **C** denote the covariance or correlation matrix for the *X*s. The variance of *Z* then equals $\mathbf{a}^T\mathbf{C}\mathbf{a}$ where **a** is the vector of *a* weights. We want to maximize the variance of *Z* subject to the constraint that $\mathbf{a}^T\mathbf{a} = 1$. Thus, we can construct the augmented Lagrangian function

$$F(\mathbf{a}) = \mathbf{a}^T\mathbf{C}\mathbf{a} + \lambda(\mathbf{a}^T\mathbf{a} - 1) \tag{1.1}$$

where λ is the Langrangian multiplier. We want to maximize $F(\mathbf{a})$ with respect to vector **a**. We may do this by taking the first derivative of $F(\mathbf{a})$ with respect to **a** and setting it to a vector of 0s. Doing this gives

$$\frac{\partial F(a)}{\partial a} = 2Ca - \lambda a = 0 \tag{1.2}$$

or

$$(\mathbf{C} - \lambda\mathbf{I})\mathbf{a} = \mathbf{0}\,. \tag{1.3}$$

One solution to (1.3) is to set all the *a*s equal to zero. This, however, is called a *trivial* solution. We require a nontrivial solution or, in other words, substantive values for the elements of **a**. To solve for this, suppose for the moment that we have some real value for λ such that the matrix $(\mathbf{C} - \lambda\mathbf{I})$ is known and has an inverse. Premultiplying both sides of (1.3) by the inverse of $(\mathbf{C} - \lambda\mathbf{I})$ gives

$$\mathbf{a} = (\mathbf{C} - \lambda\mathbf{I})^{-1}\mathbf{0}$$

or **a** must equal **0**. Hence, if matrix $(\mathbf{C} - \lambda\mathbf{I})$ has an inverse, the only solution to (1.3) is the trivial solution.

We must conclude then that in order to have a nontrivial solution, the matrix $(\mathbf{C} - \lambda\mathbf{I})$ must NOT have an inverse. That is, $(\mathbf{C} - \lambda\mathbf{I})$ must be singular and its determinant must equal 0, or

$$|\mathbf{C} - \lambda\mathbf{I}| = 0 \tag{1.4}$$

In our overview of matrix algebra, we saw that (1.4) is called the *characteristic equation* for matrix **C** and that, because **C** is symmetric, there will be *p* real values of λ. These *p* λs are the

eigenvalues of **C**. And for any given eigenvalue, the solution of (1.3) for **a** reveals that **a** is an *eigenvector* of **C**.

Now let us return to (1.3) and rewrite it as

$$\mathbf{Ca} = \lambda\mathbf{a} \tag{1.5}$$

Premultiply both sides by $\mathbf{a}^T$ giving

$$\mathbf{a}^T\mathbf{Ca} = \mathbf{a}^T\lambda\mathbf{a} = \lambda\mathbf{a}^T\mathbf{a}. \tag{1.6}$$

But $\mathbf{a}^T\mathbf{Ca}$ is the variance of variable *Z* and we have constrained $\mathbf{a}^T\mathbf{a}$ to equal unity. Consequently, (1.6) implies that

$$\lambda = \text{Var}(Z).$$

In other words, the variance of a principal component is actually an eigenvalue of the matrix **C**. Hence, to maximize Var(*Z*), all we have to do is select the largest eigenvalue of **C** and take its associated eigenvector as the weights in vector **a**.

The second principal component may be defined as that linear combination of the *X*s that has the second largest variance, or

$$Z_2 = a_{12}X_1 + a_{22}X_2 + ... + a_{p2}X_p$$

Here, we subscript the *Z* and double subscript the *a*s so that a_{ij} is the weight assigned to the *i*th variable for the *j*th principal component. Since Var(Z_2) is an eigenvalue of **C**, we simply select the second largest eigenvalue and its associated eigenvector as the solution to $\mathbf{a}_2$. We may continue with this logic, each time selecting the next highest eigenvalue and its associated eigenvector as the next principal component.

A consequence of using the eigenvalues as the variance for the principal components and the eigenvectors as the weights is that the principal components will be uncorrelated. Recall that the matrix of eigenvectors is an *orthogonal* or *orthonormal* matrix. That is, if **A** is the matrix of eigenvectors, then $\mathbf{A}^T\mathbf{A} = \mathbf{I}$. We may write the equation for the principal componets in matrix form as

$$\mathbf{z} = \mathbf{Ax}$$

where **z** is a *p* by 1 vector of principal components and **x** is a *p* by 1 vector of variables. Premultipling both sides by $\mathbf{A}^T$, giving

$$\mathbf{A}^T\mathbf{z} = \mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{I}\mathbf{x} = \mathbf{x}, \quad (1.7)$$

reveals that the variables may be written as a linear combination of the principal components. Now take the covariance matrix of **x** in (1.7),

$$\mathbf{A}^T\mathbf{C}_{zz}\mathbf{A} = \mathbf{C}$$

where $\mathbf{C}_{zz}$ is the covariance matrix among the principal components. But in the overview on matrix algebra, we have seen that $\mathbf{C} = \mathbf{A}^T\Lambda\mathbf{A}$ where Λ is the diagonal matrix of eigenvalues. Hence, the covariance matrix among the principal components is a diagonal matrix, making all the components uncorrelated with one another.

There is another definition of the principal components. We have seen that we may write the variables as a function of the weights and the principal components. That is true as long as we have as many components as there are variables. If there are fewer components, then there will be some error in writing the variables. Thus, for the first component, we may write

$$X_1 = a_{11}Z_1 + U_1$$
$$X_2 = a_{21}Z_1 + U_2$$
$$.$$
$$X_p = a_{p1}Z_1 + U_p$$

where U_i denotes a residual for the *i*th variable. In general, let **X** denote the *N* by *p* matrix of observed scores for *N* individuals on *p* variables, let $\mathbf{A}_j^T$ denote the *j* by *p* matrix of weights for the first *j* principal components (i.e., the columns of $\mathbf{A}_j$ contain the first *j* eigenvectors), let **Z** denote the *N* by *j* matrix of principal component scores for the individuals and let **U** denote the *N* by *p* matrix of residuals. Then we may write

$$\mathbf{X} = \mathbf{Z}\mathbf{A}_{jT} + \mathbf{U}$$

Mathematically, it can be shown that the first *j* principal components are those uncorrelated variables that are the best linear predictors of the observed variables in the sense that they minimize the trace($\mathbf{U}^T\mathbf{U}$). That is, if one could construct any linear predictor of the set of *p*

observed variables and regress the *p* variables on the predictor, then the best (in terms of least squares error and maximum variance) predictor would be the first principal component. The next best predictor would be the second principal component, etc.

There is yet another interpretation of PCA. Because eigenvectors and eigenvalues "redimension" a matrix, PCA redimensions the original variables by finding the dimension on which there is the largest individual differences (the first PC), a perpendicular dimension on which there is the next largest amount of variability (the second PC), etc.

So what's the big deal? We began with 566 MMPI items and we end up with 566 PCs! How does that solve the problem? It actually solves the problem if we realize that each successive PC is less variable that its predecessor. At some point we are going to be deriving PCs that account for a trivial amount of variability. The advantage of PCA comes when we ignore those trivial components. By using the most important PCs instead of the original variables, we can reduce the size of the problem. The big difficulty with PC analysis (and with factor analysis) lies in deciding where to draw the line between an important PC and a trivial PC. There are no clear-cut answers here.

Among the potential solutions to this problem, two have gained wide popularity. The first has been termed the Kaiser (1960) criterion. The Kaiser criterion accepts only those components with an eigenvalue greater than 1.0. The second criterion is the *scree test* proposed by Cattell (1966). In this test one plots the eigenvalues and then attempts to draw a straight line connecting the last *j* eigenvalues. All those components that have eigenvalues above the straight line are accepted.

An Example

Table 1.1 presents the correlation matrix for six tests of cognitive ability for 250 individuals: Vocabulary, Reading Comprehension, Sentence Completion, Mathematics, Geometry, and Analytical Reasoning. Instead of using all six tests, it might be desirable to have only a few measures of cognitive ability on everyone. Principal components may be used to

reduce the number of variates.

Table 1.1. Correlation matrix for six tests of cognitive ability.

	Vocab	Read	Sent	Math	Geom	Anlyt
Vocab.	1.000	.803	.813	.708	.633	.673
Read.	.803	1.000	.725	.660	.526	.636
Sentences	.813	.725	1.000	.618	.575	.618
Math.	.708	.660	.618	1.000	.774	.817
Geom.	.633	.526	.575	.774	1.000	.715
Anlyt.	.673	.636	.618	.817	.715	1.000

The eigenvalues of the correlation matrix are 4.43, .68, .32, .24, .17, and .15. Because the sum of the eigenvalues is 6, the first principal component accounts for 4.43/6 or 73.8% of the variance. The second component accounts for .68/6 or 11.3%, and the remaining components accounts for 5% or less. According to the eigenvalue criteria, only one component would be selected. The scree criteria, on the other hand, suggests that 2 components be computed and that is what we shall interprete.

The eigenvectors appear in Table 1.2. The first eigenvector suggests a broad component that is weighted almost equally by every cognitive test. This variate might be called "general cognitive ability." The second eigenvector is bipolar, one pole being characterized by Mathematics, Geometry, and Analytical Reasonsing. The opposite pole is marked by negative weights for Vocabulary, Reading Comprehension, and Sentence Completion. Scores on this second component would reflect quantitative ability versus verbal ability. The remaining eigenvectors are shown in Table 1.2 for completeness. In an ordinary application, they would not

be presented or interpreted.

Table 1.2. Eigenvectors of the correlation matrix for six tests of cognitive ability.

	Eigenvector:					
Test	1	2	3	4	5	6
Vocabulary	.427	-.338	.123	.123	-.411	-.710
Reading	.401	-.434	-.471	.479	.387	.225
Sentences	.400	-.434	.488	-.491	.063	.410
Math	.421	.360	-.237	.071	-.663	.438
Geometry	.388	.510	.545	.415	.345	-.044
Analytical	.411	.346	-.415	.580	.344	-.289